On the 3D hollow organ cartography using 2D endoscopic images

Tan-Binh Phan

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

On the 3D hollow organ cartography using 2D endoscopic images

THÈSE

présentée et soutenue publiquement le 12 Novembre 2020

pour l'obtention du

Doctorat de l'Université de Lorraine

Mention: Automatique, Traitement du Signal et des Images, Génie Informatique

par

Tan Binh PHAN

Composition du jury

Rapporteurs :	Mireille GARREAU	PU, Université de Rennes 1, LTSI Rennes.
	Franck MARZANI	PU, Université de Bourgogne, laboratoire ImVia.
Examinateurs :	Ernest HIRSCH	PU, Université de Strasbourg, Laboratoire Icube.
	Marie-Odile BERGER	DR INRIA, LORIA, équipe-projet Magrit.
	Dominique LAMARQUE	PUPH, Hôpital Ambroise Paré, Boulogne-Billancourt.
Invités :	Dinh-Hoan TRINH	Docteur, Laboratoire ImVia, équipe Vibot.
Directeur de thèse :	Christian DAUL	PU, Université de Lorraine, CRAN.
Co-Directeur de thèse :	Didier WOLF	PU, Université de Lorraine, CRAN.

CRAN

Centre de Recherche en Automatique de Nancy
UMR 7039 CNRS – Université de Lorraine

2, avenue de la forêt de Haye 54516 Vandœuvre-lès-Nancy
Tél.+33 (0)3 72 74 52 90 Fax +33 (0)3 83 59 56 44

Acknowledgment

First of all, I would like to express my deepest gratitude to my supervisor Prof. Christian Daul. I am very grateful to him for his support and very nice advice about my scientific study. He also helped me when I was disoriented and corrected a lot of mistakes when I wrote this thesis. Without his help, this thesis would be impossible.

I would also like to express my gratitude towards my co-supervisor Prof. Didier Wolf, who always attended my seminars, discussed together to come up with new ideas.

I am particularly grateful to Dr. Dinh Hoan Trinh, who guided me, gave me many fruitful discussions and useful ideas.

I deeply appreciate Prof. Mireille Garreau and Prof. Franck Marzani for their time and patience for reading my thesis and giving necessary corrections. I want to give my greatest thankfulness to Prof. Ernest Hirsch, Prof. Marie-Odile Berger, and Prof. Dominique Lamarque for their acceptance to be in the PhD scientific board, and fruitful discussions, and corrections. In particular, I would also like to thank Prof. Dominique Lamarque for sharing the patient data with me.

I would like to give my special regards to University of Lorraine which provided the finance for my thesis through a doctoral contract. I want to express my acknowledgment to Christine and my colleagues in the CRAN laboratory for their help and support.

I also can't forget a great time being with the Vietnamese student union in Nancy. We have organized some weekend meetings, made Asian traditional foods, and played football together. All of these things make my life in France being wonderful.

Finally, I wish to express my gratitude to my family and my girlfriend, who always love, care, and encourage me. They are always with me and motivated me to try to finish my work.

I would like to dedicate this thesis to my parents and my three elder brothers.

Table of Contents

Résumé étendu

Mots-clés: structure from motion, flot optique dense, mosaïquage 3D, endoscopie, dermatologie.

Contexte médical et objectifs scientifiques

L'endoscopie est une technique incontournable pour la détection des lésions cancéreuse ou inflammatoires des organes creux comme la vessie, l'estomac, le côlon ou l'oesophage. L'endoscopie est la seule procédure médicale permettant d'observer les couleurs naturelles et les textures des parois intérieures des organes creux. Dans les examens cliniques classiques, Les systèmes endoscopiques sont équipés d'une source de lumière et d'une caméra qui transmet des images à travers un tube rigide ou flexible, les images étant visualisées en temps réel sur un moniteur de télévision couleur. Le but d'un tel examen est d'explorer visuellement le paroi épithéliale (l'épithélium est le tissu qui tapisse la paroi interne de tous les organes).

Les endoscopes fournissent aux cliniciens des images à haute résolution ayant un faible champ de vue (FoV). Comme ces images à FoV limité ne permettent de visualiser que partiellement les régions d'intérêt, elles ne facilitent pas le diagnostic des lésions et le suivi des patients. En outre, les séquences vidéo endoscopiques sont difficiles à interpréter après l'endoscopie. Ce fait constitue un obstacle à l'archivage des données et à la traçabilité des examens. Une méthode pour résoudre les problèmes médicaux mentionnés ci-dessus consiste à étendre le champ de vision en calculant des images panoramiques (mosaïques 3D) à partir de la séquence d'images acquise lors d'une endoscopie. De tels FoVs 3D étendus améliorent le diagnostic puisque les lésions complètes (et les repères anatomiques potentiels) sont visibles sur une seule image. De plus, la comparaison des images de FoV étendues calculées pour deux ou plusieurs examens endoscopiques rend possible le suivi du patient par les cliniciens (endoscopistes ou chirurgiens). L'évaluation de l'évolution d'une lésion est très difficile et fastidieuse en comparant deux séquences vidéo acquises pour un même patient à un intervalle de quelques semaines ou mois.

Les algorithmes de mosaïquage 2D (ou 3D) placent les pixels (ou points 3D) vus dans des vues partielles de la scène "non reliées géométriquement" dans un système de coordonnées commun dans lequel une représentation de la scène complète est construite (c'est-à-dire que l'algorithme de mosaïquage détermine la relation géométrique entre les pixels/points 3D des différentes vues partielles). Les mosaïques 2D construites avec les algorithmes proposés par le passé par divers laboratoires

représentent une réelle avancée en termes de diagnostic, de suivi des patients et d'archivage des données. Cependant, les mosaïques 2D présentent des inconvénients intrinsèques en raison de leur représentation bidimensionnelle de la scène. En effet, en raison de la distorsion géométrique qui se produit lors du placement de données 3D dans les mosaïques 2D, des parties d'organes importantes ou complètes ne peuvent pas être entièrement représentées par une image 2D panoramique unique. En outre, seule la première image d'une mosaïque a la résolution d'origine, la résolution des autres images placées dans la mosaïque 2D dépend de la trajectoire de l'endoscope et des points de vue dans l'organe creux. Pour éviter ces problèmes, cette thèse vise à développer un algorithme de mosaïquage 3D qui peut être utilisé pour reconstruire de grandes surfaces intérieures d'organes creux.

Seules quelques techniques de reconstruction 3D ont été proposées pour les organes creux et appliquées avec succès aux données cliniques. Les algorithmes visant à reconstruire des parties de scène en 3D à l'aide de données endoscopiques peuvent être classés en deux groupes d'approches, à savoir le groupe des méthodes de vision active et le groupe des algorithmes de vision passive. Alors que les méthodes basées sur la vision active (par exemple, les techniques de temps de vol ou de lumière structurée) nécessitent la projection d'une lumière contrôlée dans l'environnement, les méthodes de vision passive (par exemple, "Shape from X, Simultaneous Localization and Mapping", structure à partir du mouvement, etc. Dans cette thèse, le choix a été fait d'étudier une approche de vision passive pour la construction de mosaïques 3D d'organes creux afin d'éviter les modifications matérielles des systèmes endoscopiques (les systèmes de vision active sont techniquement difficiles à mettre en œuvre sur les systèmes endoscopiques et impliquent des adaptations matérielles importantes et coûteuses). Après une analyse des avantages et des inconvénients de diverses techniques de vision passive, une technique de structure à partir du mouvement (SfM) a été sélectionnée comme approche générale pour la reconstruction des surfaces intérieures des organes creux à l'aide de séquences vidéo endoscopiques.

La SfM est le processus d'estimation de la structure 3D d'une scène (c'est-à-dire que la forme des objets ou des surfaces est représentée par des nuages de points) à partir d'un ensemble d'images 2D. Les méthodes SfM classiques se composent essentiellement de deux étapes principales :

1. Détermination d'un ensemble de correspondances ponctuelles entre les images. Dans les techniques SfM classiques, les correspondances de points sont données par des pistes de points homologues, les points homologues étant un ensemble de points 2D issus de la projection d'un même point 3D dans des images acquises à partir de points de vue différents. Les points homologues peuvent être suivis le long de la séquence d'images en détectant et en faisant correspondre des points caractéristiques. L'avantage de ces méthodes est que les points caractéristiques sont localisés avec une précision inférieure au pixel, tandis que leurs descripteurs caractéristiques peuvent être invariants en fonction de l'échelle, de la rotation et des changements d'intensité.

2. Détermination de la structure de la scène. Les traces de points caractéristiques sont ensuite utilisées pour récupérer simultanément la pose de la caméra (posi-

tion et orientation de la caméra) pour chaque point de vue (image) et la forme des objets représentés par les nuages de points 3D.

Au cours des dernières décennies, les approches SfM ont conduit à de nombreux et impressionnants résultats de reconstruction de scènes dans des applications très différentes. Cependant, la plupart des résultats ont été obtenus pour des scènes comprenant de nombreuses textures et structures contrastées, et pour des conditions d'acquisition dans lesquelles les points de vue des caméras peuvent être contrôlés. Ces données et conditions d'acquisition facilitent la tâche des algorithmes de correspondance de caractéristiques utilisés dans la première étape des techniques SfM classiques, de sorte qu'un grand nombre de longues trajectoires de points peuvent être obtenues. L'obtention de grands ensembles de points homologues, qui sont précisément localisés dans les images et vus de points de vue très différents, joue un rôle important en termes de robustesse et de précision des méthodes SfM. Cependant, dans les scènes endoscopiques, les données ne sont ni avec des textures et des structures contrastées, ni acquises à partir de points de vue contrôlés. En effet, les scènes d'organes creux sont souvent caractérisées par un manque de textures, de fortes variations d'illumination entre les images, des réflexions spéculaires et des trajectoires de caméra très difficiles à contrôler. De telles conditions d'acquisition et de scène rendent les méthodes SfM classiques basées sur la détection et la correspondance des caractéristiques souvent inopérantes en endoscopie.

Cette thèse propose de nouvelles approches SfM pour retrouver les surfaces intérieures des organes creux. Les solutions sont basées sur l'idée qu'une exploitation appropriée des la correspondance dense de points (mais moins précises) fournies par les algorithmes de flot optique peut compenser la précision sub-pixel des méthodes basées sur les caractéristiques. Les algorithmes de flux optique dense (DOF) ont un fort potentiel pour trouver des points homologues dans des scènes endoscopiques complexes [TD19]. Malgré ce potentiel élevé, les techniques DOF ont été très peu associées aux techniques SfM. Dans ce travail, le DOF est utilisé seul ou en combinaison avec les méthodes d'appariement des caractéristiques. Ainsi, selon le type d'organe creux, cette thèse propose deux méthodes SfM (une pour les scènes presque sans texture, et une autre pour les scènes avec une quantité de texture insuffisante). De plus, et quelle que soit la méthode proposée dans cette thèse, l'utilisation des champs DOF permet d'établir de nombreux et grands groupes de points homologues qui conduisent à la reconstruction d'un nuage de points dense. Par conséquent, le pipeline de reconstruction 3D proposé dans cette thèse ne nécessite pas l'utilisation d'un algorithme stéréo multi-vues (MVS). Une étape MVS est obligatoire pour les pipelines courants de construction de scènes 3D car les méthodes SfM classiques basées sur la détection et la correspondance de caractéristiques ne peuvent fournir que des nuages de points épars.

En supposant que toutes les scènes endoscopiques envisagées dans cette thèse soient presque rigides, et que les caméras ne doivent pas être calibrées dans des conditions médicales, les méthodes SfM proposées doivent d'abord prouver leur précision sur des fantômes non déformables dont la vérité terrain est connue. Outre la précision, la robustesse est le critère le plus important pour une application médicale. Dans le contexte de l'endoscopie, la robustesse signifie qu'une méthode

SfM avec des paramètres d'algorithme constants peut être appliquée avec succès à diverses scènes endoscopiques.

Principales contributions and discussions

La principale contribution présentée dans cette thèse réside dans la proposition de deux nouvelles méthodes SfM qui peuvent traiter des conditions de scène complexes comme celles rencontrées en endoscopie. Les algorithmes DOF et FMDOF ont tous deux été intégrés dans un schéma SfM et ont donné des résultats réalistes en termes de reconstruction des surfaces d'organes creux. Dans ce travail, les scènes médicales étaient presque rigides, tous les tests de construction de surface ont été effectués avec des paramètres d'algorithme constants et aucune étape de calibrage de la caméra n'a été nécessaire pour assurer une récupération de forme robuste et réaliste. En outre, les tests de reconstruction avec des paramètres intrinsèques de caméra connus et inconnus ont conduit à des performances similaires en termes de précision et de robustesse. Ainsi, les algorithmes SfM basés sur le DOF et le FMDOF ont tous deux un intérêt pratique puisque lors des procédures médicales, les calibrages d'endoscopes doivent être évités (les examens médicaux doivent rester inchangés). Le pipeline de reconstruction 3D proposé est également capable de fournir directement un nuage de points 3D dense sans aucun algorithme stéréo multi-vues.

Les résultats obtenus pour les fantômes texturés ont montré que la précision des méthodes SfM proposées peut se rapprocher de celle de deux méthodes de référence basées sur la détection de caractéristiques. Il était important d'évaluer objectivement la précision inhérente des méthodes proposées avant de les appliquer aux scènes médicales réelles qui sont sans vérité de terrain connue.

D'un point de vue scientifique, le flot optique a été rarement utilisé en SfM en raison de l'accumulation d'erreurs lors du suivi de points le long d'images consécutives. Cette thèse propose une stratégie originale pour exploiter les champs de vecteurs basée sur la sélection d'images de référence qui permettent à la fois d'établir des groupes de points homologues nombreux et importants, et de construire des surfaces sans discontinuité. Les champs de vecteurs entre une image de référence et chacune de ses images superposées permettent d'éviter le suivi de points le long de la séquence vidéo. De plus, ce travail a permis une estimation précise des du flot optique en utilisant un nouveau descripteur invariant aux illuminations pour traiter les scènes avec peu de textures/structures, et affectées par de forts changements d'illumination.

Du point de vue médical, les méthodes SfM proposées permettent de reconstruire les surfaces acquises pour diverses scènes endoscopiques et modalités d'imagerie. Les images à FOV étendus facilitent la détection des régions avec des anomalies (par exemple, avec des polypes) ou des inflammations (par exemple, des inflammations autour de la région de l'antre pylorique de l'estomac). En outre, les surfaces d'une même région reconstruites par les algorithmes SfM proposés pour deux ou plusieurs examens permettent de diagnostiquer l'évolution d'une lésion ou d'évaluer la rémission d'un tissu après une intervention chirurgicale par exemple. Pour les

applications médicales testées, les algorithmes de reconstruction 3D décrits ont conduit systématiquement à des formes 3D cohérentes (en accord avec l'anatomie de l'organe), sans discontinuités de textures ou de structures, ainsi qu'à une résolution acceptable quel que soit l'emplacement observé sur la surface. Les images 3D texturées des surfaces des parois internes des organes permettent également l'échange d'informations entre médecins de différentes spécialités.

Récemment, les travaux présentés dans [MWP+19] ont porté sur la visualisation en temps réel des surfaces de morceaux de côlon (petites parties du côlon) en associant une méthode SfM et une méthode SLAM. Les auteurs ont utilisé un réseau de neurones récurrents (RNN) pour prédire les cartes de profondeur et les poses des caméras pour les séquences d'images coloscopiques, la vérité terrain (petites parties texturées de la surface du côlon) pour l'entraînement du réseau étant déterminé avec la méthode COLMAP SfM. Le RNN a été intégré dans une approche SLAM standard pour corriger les cartes de profondeur et les poses de caméra prédites par le réseau RNN. Un maillage de texture global est finalement obtenu après la fusion des cartes de profondeur des cadres coloscopiques. Les résultats obtenus dans [MWP+19] sont impressionnants dans le sens où la méthode permet de réaliser des reconstructions en temps réel et de visualiser des régions potentiellement manquantes (non scannées ou occultées). Cependant, la longueur des parties du côlon reconstruites reste limitée en raison de la vérité terrain utilisée pour l'entraînement du RNN qui prédit les cartes de profondeur et les poses des caméras. En effet, les méthodes SfM basées sur les caractéristiques comme COLMAP reconstruisent des scènes complexes de coloscopie avec une précision modérée (comme dans l'estomac, il y a souvent un manque de textures dans de nombreuses images). À l'avenir, la méthode SfM basée sur du flot optique pourrait être utilisée pour fournir des vérités de terrain réalistes (surfaces des organes) afin d'améliorer l'entraînement du réseau neuronal et d'augmenter la précision de la carte de profondeur et de la prédiction de la pose de la caméra dans l'approche de reconstruction proposée dans [MWP+19]. Ces méthodes basées sur un réseau neuronal pourraient être appropriées à la fois pour le côlon et l'estomac.

Une extension naturelle des algorithmes de reconstruction proposés consisterait à les adapter à des scènes non rigides. La scène non rigide qui sera considérée dans un travail rapproché est située à la jonction du cardia (partie supérieure de l'estomac) et du fond de l'œsophage. À cette jonction, le muscle du sphincter cardiaque agit comme une valve qui s'ouvre et se ferme. Lorsque le gastroscope se trouve dans l'œsophage et pointe vers le cardia, il observe une forme approximativement cylindrique dont le fond s'ouvre et se ferme. En cas de reflux gastrique chronique (reflux acide de l'estomac vers l'œsophage), on observe un changement de couleur des tissus œsophagiens inférieurs (des "traînées" roses remontent sur le tissu blanc sain sans inflammation). Les mouvements dus au muscle du sphincter cardiaque déforment cette zone et il est difficile d'observer cette surface afin de quantifier l'œsophage de Barrett. Dans ce contexte, l'objectif d'un algorithme NRSfM (non-rigid structure from motion) ne serait pas d'étendre le FoV, mais plutôt de construire une vidéo d'une surface dont la forme change avec le temps. Avec une telle vidéo 3D, le gastro-entérologue peut choisir l'état de surface qui lui permet d'observer l'étendue de l'œsophage de Barrett.

General introduction

This thesis was written at the CRAN laboratory (Centre de Recherche en Automatique de Nancy, UMR 7039 CNRS/Université de Lorraine) in the BioSiS (Biologie, Signaux et Systèmes en Cancérologie et Neurosciences) department. This thesis was funded through a doctoral contract of the Université de Lorraine and the participation to conferences was supported by the Agence nationale de la recherche (ANR-15-CE17-0015) in the frame of the EMMIE project (Endoscopie MultiModale pour les lésions Inflammatoires de l'Estomac). One of the goals of this project lies in the early detection and characterization of the mucosal inflammations preferentially occurring in the pyloric antrum region of the stomach and observed in gastroscopic video-sequences. The general principle of the inflammation characterization and documentation in the EMMIE project relies on the detection and classification of inflammations using multi-spectral data and/or narrow-band imaging data and on the superimposition of this information onto 3D panoramic images computed with white light video-sequences acquired with an endoscope. This thesis focusses on the construction of extended and textured 3D surfaces of the pyloric antrum region using only a sequence of endoscopic images with a limited field of view (FoV). However, the proposed 3D cartography (or 3D mosaicing) algorithms should be applicable beyond the standard white light modality of gastroscopy which is under consideration in the EMMIE project. The proposed mosaicing methods should be usable in gastroscopy and cystoscopy, as well as in various image modalities as white-light, narrow-band imaging, and fluorescence. Medical expertise and data were notably supplied by Prof. Dominique Lamarque from the Ambroise Paré Hospital in Boulogne Billancourt (AP-HP Paris), France.

Endoscopy is an unavoidable technique for the detection of cancerous or inflammatory lesions in hollow organs as the bladder, stomach, esophagus, or colon. Endoscopy is the only medical procedure allowing to observe the natural colours and textures of the inner walls of hollow organs. In classical clinical examinations, endoscopic systems are equipped by a light source and a camera which transmits images through a rigid or flexible tube, the images being visualized in real-time on a colour TV-monitor. The purpose of such an examination is to visually explore the epithelial wall (the epithelium is the tissue that lines the internal wall of all hollow organs).

Endoscopes provide clinicians with high-resolution images having a small FoV. Since such limited FoV images only partially visualize regions of interest, they do not facilitate lesion diagnosis and patient follow-up. Moreover, the endoscopic video-sequences are difficult to interpret after the endoscopy. This fact is a barrier for data

archiving and examination traceability. One method to solve the above mentioned medical issues is to extend the FoV by computing panoramic images (3D mosaics) using the image sequence acquired during an endoscopy. Such extended 3D FoVs improve the diagnosis since complete lesions (and potential anatomical landmarks) are seen in one unique image. Moreover, comparing extended FoV images computed for two or more endoscopic examinations makes the patient follow-up possible for clinicians (endoscopists or surgeons). A lesion evolution assessment is very difficult and tedious by comparing two video-sequences acquired for a same patient at a time interval of some weeks or months.

2D (or 3D) mosaicing algorithms place the pixels (or 3D points) seen in "geometrically unconnected" partial views of the scene into a common coordinate system in which a representation of the complete scene is constructed (i.e. the mosaicing algorithm determines the geometrical relationship between the pixels/3D points of the different partial views). The 2D mosaics built with the algorithms proposed in the past by various laboratories represent a real advance in terms of diagnosis, patient follow and data archiving. However, 2D mosaics have intrinsic drawbacks due to their bi-dimensional scene representation. Indeed, due to the geometrical distortion occurring when placing 3D data in 2D mosaics, large or complete organ parts cannot be entirely represented with a unique 2D panoramic image. Moreover, only the first image of a mosaic has the original image resolution, the resolution of the other images placed in the 2D mosaic depends on the endoscope trajectory and viewpoints in the hollow organ. To avoid these problems, this thesis aims to develop a 3D mosaicing algorithm which can be used to reconstruct large inner surfaces of hollow organs.

Only few 3D reconstruction techniques were proposed for hollow organs and successfully applied to clinical data. The algorithms aiming to reconstruct 3D scene parts using endoscopic data can be classified into two groups of approaches, namely the group of active vision methods and the group of passive vision algorithms. While the methods based on active vision (e.g., Time-of-Flight or structured light techniques) require controlled light to be projected into the environment, the passive vision methods (e.g., Shape from X, Simultaneous Localization and Mapping, Structure from Motion, etc.) only require 2D images. In this thesis, the choice was made to study a passive vision approach for the construction of 3D mosaics of hollow organs to avoid hardware changes of endoscopic systems (active vision systems are technically difficult to implement on endoscopic systems and involve significant and expensive hardware adaptations). After an analysis of the advantages and drawbacks of various passive vision techniques, a Structure from Motion (SfM) technique was selected as a general approach for the reconstruction of the inner hollow organ surfaces using endoscopic video-sequences.

SfM is the process of estimating the 3D structure of a scene (i.e. the shape of objects or surfaces is represented by point clouds) from a set of 2D images. In essence, classical SfM methods consist of two main stages:

1. Determination of a set of point correspondences between the images. In classical SfM techniques, the point correspondences are given by homologous point tracks, homologous points being a set of 2D points issuing from the projection

of a same 3D point in images acquired from different viewpoints. Homologous points can be tracked along the image sequence by detecting and matching feature points. The advantage of these methods is that feature points are located with sub-pixel accuracy, while their feature descriptors can be invariant to scale, rotation, and intensity changes.

2. Scene structure determination. The feature point tracks are then used to simultaneously recover the camera pose (camera position and orientation) for each viewpoint (image) and the shape of the objects represented by 3D point clouds.

Over the last decades, SfM approaches led to numerous impressive scene reconstruction results in very different applications. However, most of the results were achieved for scenes including numerous contrasted textures and structures, and for acquisition conditions in which the camera viewpoints can be controlled. These data and acquisition conditions facilitate the task of the feature matching algorithms used in the first stage of classical SfM techniques so that a large set of long point tracks can be obtained. Obtaining large sets of homologous points, which are precisely located in the images and seen from very different viewpoints, plays an important role in terms of robustness and accuracy of SfM methods. However, in endoscopic scenes, the data are neither with contrasted textures and structures, nor acquired from controlled viewpoints. Indeed, hollow organ scenes are often characterized by a lack of textures, strong illumination changes between images, specular reflections and very difficult to control camera trajectories. Such acquisition and scene conditions make the classical SfM methods based on feature detection and matching often inoperative in endoscopy.

This thesis proposes new SfM approaches to recover inner hollow organ surfaces. The solutions are based on the idea that appropriate exploitation of the dense (but less accurate) point correspondences provided by optical flow algorithms can compensate for the subpixel accuracy of feature-based methods. Dense optical flow (DOF) algorithms have a high potential to find homologous points in complex endoscopic scenes [TD19]. Despite this high potential, DOF techniques were very barely associated with SfM techniques. In this work, DOF is either used alone or in combination with the feature matching methods. Thus, according to the hollow organ type, this thesis proposes two SfM methods (one for scenes with almost no textures, and another for scenes with an insufficient texture amount). Moreover, and whatever the method proposed in this thesis, the use of DOF fields enable to establish numerous and large groups of homologous points that lead to the reconstruction of a dense point cloud. Therefore, the 3D reconstruction pipeline proposed in this thesis does not require the use of a multiview stereo algorithm (MVS). A MVS step is mandatory for common 3D scene construction pipelines since classical SfM methods based on features detection and matching can only delivers sparse point clouds.

Assuming that all endoscopic scenes under consideration in this thesis are almost rigid, and that the cameras should not be calibrated in medical conditions, the proposed SfM methods need first to prove their accuracy on non-deformable phantoms with known ground truth. Besides the accuracy, robustness is the most important

criteria for a medical application. In the context of endoscopy, robustness means that a SfM method with constant algorithm parameters can be successfully applied to various endoscopic scenes.

This thesis is organized into the following chapters:

Chapter 1: Medical context and scientific objectives. This chapter first introduces the medical context and explains why the construction of extended 3D surfaces are medically of interest in the endoscopy of hollow organs. Then, a brief overview of the 3D reconstruction approaches (active and passive vision methods) and their adaptation to endoscopy is presented. The strengths and weaknesses of each 3D endoscopic data construction method are assessed and discussed. This study justifies the potential of a SfM approach in terms of feasibility and robustness of the construction of hollow organ surfaces from endoscopic video-sequences. An introduction to SfM methods, to the main SfM pipelines, and to the main issues of SfM when this technique is applied to endoscopic scenes is also given in this chapter. This chapter also justifies why, among the three SfM pipeline categories present in the literature (global, incremental, and hierarchical SfM pipelines), the incremental pipeline was chosen. The chapter end discusses the scientific issues of SfM when it is applied to endoscopic scenes and gives first indications about the solutions to be studied for obtaining an inner hollow organ reconstruction method which is of interest in clinical conditions.

Chapter 2: Classical Structure from Motion. This chapter is dedicated to the literature review on the different classical parts/steps of an incremental SfM approach which provides both a sparse 3D point cloud (first representation of the surface to be recovered) and the camera pose parameters (sensor position and orientation) for each viewpoint. Starting with the homogeneous notation for points and lines on a plane, the Pinhole camera model is used to describe the projection of 3D scene points onto a 2D image plane. Then, this chapter presents a brief introduction of the camera calibration techniques which represent an important issue in the classical SfM pipelines because several SfM steps require the knowledge of the camera parameters. To do so, the two-view SfM algorithm is first presented and used to describe the principle of the multi-view SfM-part. The latter consists of two main stages, namely the determination of point tracks and the incremental reconstruction pipeline. In the two-view SfM section, essential concepts of stereo-vision are also given. These concepts relate to the role and determination of homographies, of the epipolar geometry (fundamental matrix, essential matrix, etc.), of feature matching methods, etc. The Multi-view SfM pipeline description is followed by the presentation of the COLMAP SfM pipeline [SF16].

Chapter 3: Dense Optical Flow based Structure from Motion. As highlighted in Chapter 1, classical SfM techniques fail to construct surfaces of endoscopic data mainly due to the feature matching step which cannot lead in a robust way to homologous point tracks. In this chapter, a novel SfM method is proposed by replacing, in the point correspondence determination step, the classical feature matching algorithms by an original dense optical flow (DOF) based algorithm. The DOF algorithm determines the groups of homologous point based on the observation that 2D homologous points of the same group lie on the overlapping (common) regions of the

images to which these points are belonging. To present in detail the algorithm, this chapter first introduces a robust optical flow method which delivers accurate flow fields between image pairs and which is based on a data-term incorporating a new illumination-invariant descriptor. Then, a strategy for searching in a video-sequence numerous consecutive and non-consecutive images which share common scene parts is described. This strategy is based on the search of reference images having a significant overlap with numerous other images and corresponding to small surface parts covering the whole surface to be reconstructed. The group of homologous points is finally determined by using the DOF fields between a reference image and each image overlapped with it. Two points linked by a vector of the flow field are considered as being homologous if they are preserved by specular reflections and occlusions. It has to be noticed that the second stage of the proposed DOF-based SfM is the same as that of the state-of-the-art COLMAP algorithm [SF16]. Indeed, the multi-view-SfM stage in classical SfM methods exhibits high performances when homologous point groups can be determined in an accurate and robust way. A set of optimal parameters for the optical flow computation and the point group determination is given by conducting numerous experiments on phantom data with known ground truth in terms of shape and dimensions, and covered by various textures. These parameters are kept constant for all experiments and tests performed later in the thesis. The robustness of the proposed DOF algorithm is shown for real endoscopic data through a comparison with the results obtained with classical feature-based SfM methods.

Chapter 4: SfM based on the feature matching method combined to DOF. The principle of the method is to exploit the accuracy of feature points in the image regions including enough contrasted textures, while the use of DOF is limited to the remaining regions without contrasted textures due to complex acquisition conditions or missing tissue structures (e.g., blood vessels under the epithelium). This combination is flexible in terms of textures due to the fact that the FMDOF algorithm works in a similar way as the DOF algorithm described in Chapter 3 when SIFT features cannot be extracted (as in gastroscopic scenes) and, on the contrary, automatically exploits features in image parts where they are available (e.g., as in cystoscopy). Similarly to the DOF-based SfM, the second stage of the FMDOF-based SfM method uses the classical multi-view reconstruction of COLMAP. A set of parameters for the computation of the FMDOF matches is also fixed in this chapter, this setting remaining constant for all the results given in the rest of the thesis. At the end of this chapter, cystoscopic data are used to highlight the difference when treating bladder images with a classical feature-based method and with the FMDOF-based approach.

Chapter 5: Epithelial surface reconstruction results. This chapter gives and discusses the surface reconstruction results achieved by the SfM methods presented in Chapters 3 and 4. Tests on phantom data are first conducted to evaluate the accuracy of the proposed methods. The accuracy of these methods is compared with that of the state-of-the-art COLMAP [SF16] and VisualSFM [Wu13] SfM approaches. Other tests described in this chapter were conducted to assess the impact of the knowledge of the intrinsic camera parameters (e.g., the focal length) on the

SfM results of the DOF- and FMDOF-algorithms. For practical reasons, endoscopes should not be calibrated in clinical situation. For this study, reconstruction tests were conducted on the same image sequences with and without calibrated cameras. The robustness of the proposed solutions was finally highlighted with the data of various medical scenes (gastroscopy, cystoscopy, dermatology) and different image modalities (white light for all medical applications, narrow-band imaging for gastroscopy, and fluorescence in cystoscopy). Besides that, a reconstruction of a small kitchen room has proven the potential of the proposed methods for non-medical scenes. All reconstruction tests were performed with constant algorithm parameters.

Finally, a conclusion summarizes the major contributions of this thesis and gives some perspectives which can improve the potential of the described work.

Chapter 1

Medical context and scientific objectives

Contents

1.1 Medical context

This chapter introduces general aspects relating to endoscopy, such as the benefits of this imaging technique, the different endoscopic examinations and modalities, the working principle of an endoscope, etc. Even if endoscopes present valuable medical advantages, their limited field of view (FoV) and the 2D nature of the acquired images are not optimal information for a lesion diagnosis, lesion evolution follow-up and data archiving. The construction of extended three-dimensional (3D) organ surfaces can significantly improve the exploitation of endoscopic data. This chapter gives an overview of existing 3D reconstruction techniques in endoscopy, proposes solutions and choices to improve the 3D reconstruction of hollow organs

and exposes the scientific challenges to be met for endoscopic scenes with a complex content (images with almost no textures, important changes in terms of acquisition conditions over the time, etc.).

1.1.1 Endoscopy and clinical diagnosis

Radiography, echography, X-ray tomography and magnetic resonance imaging are acquisition techniques which deliver bidimensional (2D) or tridimensional and morphological or functional information of the inside of the human body. However, endoscopy is the only imaging technique of internal organs which provides images with natural colour and texture information. These images, acquired with classical CDD (Charge Coupled Device) cameras, represent valuable data for lesion (e.g., cancers or inflammations) detection and diagnosis based on textures, structures and colours corresponding to information standardly used by the human brain to analyse a scene.

(a) Upper endoscopy

(b) Rigid cystoscope, Karl Storz company.

(c) Flexible cystoscope (Endo-EYE model from the Olympus company).

Figure 1.1: Procedure and instruments in endoscopy. (a) An upper endoscopy is the procedure allowing to diagnose and, sometimes, to treat lesions that affect the upper part of the digestive system which includes the esophagus, the stomach and the beginning of the small intestine (i.e., the duodenum). A flexible endoscope is inserted in the mouth (acting as meatus) and can be displaced in the complete upper digestive tract. In the more and more widespread digital technology, the CCD-matrix is fixed on the distal tip (the endoscope tip which is inside the body), while the camera (not visible here) is connected on the proximal tip (only the eyecup pointed by the "endoscope arrow" is visible). Biopsies can be performed through the operating channel since the passageway from the mouth to the small intestine opening is usually unobstructed. (b) and (c) In urology, the endoscope is inserted through the urinary meatus (external urethral orifice) and is used to scan the bladder wall or the urethra. This procedure is sometimes done with a rigid endoscope (see Fig. 1.1(b)) when surgery is required, but most often flexible endoscopes (fiberscopes, see Fig. 1.1(c))) are used to facilitate the examination and increase the patient's comfort. The illustration are taken from [Ali16, Endb].

Endoscopic examinations are often non-invasive, or sometimes minimally invasive, since large skin incisions are never required for this type of medical procedures. Endoscopes are inserted in the body through either a small incision or a natural meatus (in anatomy, a meatus defines the orifice of an organ). The tissue of internal or external walls of a great variety of organs can be observed with endoscopes, namely that of the lungs, liver, bladder, colon, oesophagus, stomach, joints for instance. The primary function of endoscopes is the acquisition of images or video sequences but, depending on the instrument, it is also possible to perform biopsies using instruments that pass through an operating channel. This channel can also be used to perform minimally invasive surgical procedures, such as gallbladder removal. The principle of the upper endoscopy procedure (or esophagogastroduodenoscopy-EGD) is sketched in Fig. 1.1(a).

Endoscopes, as shown in Figs. 1.1(b)-1.1(c) are very useful tools for inspecting numerous human body parts, as

- the gastrointestinal tract (esophagus, stomach, duodenum, colon, rectum and anus),
- the respiratory tract (the upper tract consits of the nasal cavity, the paranasal sinuses and the pharynx, while the lower tract includes the trachea, bronchi, bronchioles and lungs),
- the ear (outer, middle, and inner ear),
- the urinary tract (by going from the top to the bottom: the kidneys, ureters, bladder and urethra),
- female reproductive tract (cervix, uterus, fallopian tubes),
- joints (knees and elbow) and
- various cavities like the abdominal or the pleural cavity for instance [Enda].

During certain examinations, endoscopy can sometimes be combined with other acquisition techniques, such as ultrasound for instance. An ultrasound probe can be attached to the endoscope to obtain images giving additional and complementary information of the internal wall of the oesophagus or stomach for example. An endoscopic ultrasound procedure can also provide images of organs (such as the pancreas) that are difficult to reach with the endoscope alone.

Finally, for the small intestine, there is a particular means of endoscopic exploration: the capsule video-endoscopy. The capsule embeds a miniaturized wireless CCD camera that acquires a few images per second (typically 5 to 10 images per second) which are either transmitted by radio waves to a sensor placed on the skin, or stored in an on-board memory to be read after the capsule is "returned" by the patient. The main indication for capsule endoscopy is the presence of chronic bleeding in the gastrointestinal tract that could not be detected with conventional endoscopic examinations (gastroscopy and colonoscopy). In this situation capsule endoscopy is the main means of exploration.

Most conventional endoscopes consist of following components.

1. *A rigid or flexible tube.* The tube is the endoscope part that allows to reach the organ of interest inside the body. It connects the distal tip of the endoscope that navigates through the patient's body to the proximal tip which is itself connected to the endoscopic column. In some cases this tube is rigid, mainly when the objective of endoscopy is minimally invasive surgery, for example in the abdominal cavity or in the knee. Most endoscopes have a flexible tube, especially in the case of hollow organ inspection, which requires easy navigation and crushable distal tips.

2. *Light source and transmission.* In most examinations the source emits white light. However, as discussed later, in numerous examinations complementary light sources exits. In most endoscopes, the light is transmitted from the proximal tip to the distal type by optical fibres. A cone of diffuse light illuminates the scene.

3. *Optics and CCD matrix on the distal tip.* The current trend of the endoscope technology is to fix the CCD matrix for sampling the image lines directly on the distal tip. This way to proceed eliminates the honeycomb effect of previous technologies for which the reflected light was transmitted towards the proximal tip with a bundle of optical fibres. The sampling of this light by a CCD matrix led to images superimposed by a more or less pronounced periodic structure due to the optical fibres. Chip on the tip technologies avoid this honeycomb effect and lead usually to well contrasted images when the endoscope speed is moderate enough. The optics are conceived to obtain short focal lengths ensuring a minimum of angular aperture. However, since the acquisition distances are short (to ensure high resolution) the field of view remains limited. Moreover, short focal lengths are related to barrel distortions (see Chapter 2).

4. *Acquired signal transmission.* With the chip on the tip technology, electric signals are conveyed from the distal tip to the proximal tip, and images are digitized after the transmission.

The length of the tube, its diameter, and the name of the endoscope depend on the organ or the system track to be inspected. Thus, arthroscopes, bronchoscopes, nephroscopes are dedicated to organs (joints, bronchial tree and the renal cavities, respectively), while gastroscopes, colonoscopes, cystoscopes and ureteroscopes are used for body parts like the upper digestive track, the lower digestive track, the lower urinary tract and the upper urinary tract, respectively. Laparoscopes are endoscopes used in mini-invasive surgery. In same way, the name of the endoscopic examination depends on the organ or system tract (gastroscopy, cystoscopy, ureterology, etc.).

Similarly to the external body surface covered by skin, the internal walls of hollow organs like the oesophagus, colon, small intestine, stomach, and bladder are lined by epithelial tissue. The aim of endoscopic examinations is to visualize various information types, namely "normal" textures corresponding to healthy epithelium and signals/structures corresponding to lesions. The standard light source colour of most of the endoscopic examinations is white. In this widespread image modality,

natural rendering of epithelial colours and textures allow endoscopists to recognize the organ anatomy, anatomical landmarks helping to understand the endoscope position in the organ and at least a part of the lesions to be detected. White light is the modality in which endoscopists are best able to identify parts of an organ. However, with white light sources, not all types of lesions can always be diagnosed or, when a lesion can be detected, it may not be possible to diagnose diseases early. For this reason it is often possible to switch between two light sources giving complementary information: white light for the natural scene rendering and a first diagnosis, and another light source facilitating the early detection of lesions, but with a less natural visualisation of the organ.

For a gastroscopy, in order to make the epithelial tissues accessible to the camera, the patient must be fasting (meals, drinks and tobacco are prohibited six hours before the examination) and air is injected into the digestive tract to allow for a better visualization. In the case of the stomach, even if the organ is inflated by air, the walls are not totally rigid and the surface deforms itself over time, in particular due to the movement of the endoscope. Two areas are usually monitored in this organ: the pyloric antrum in the lower stomach region (see Fig. 1.2) and the cardia, at the junction with the oesophagus. In these areas, chronic inflammations appear in a privileged way, which can cause cancer. The white light modality (WL, see Fig. 1.2(a)) allows a natural representation of these areas, but inflammations are usually only detected at an advanced stage. In the Narrow Band Imaging (NBI, green-blue light source) modality which is shown in Fig. 1.2(b), the epithelial surfaces are less natural but a stronger texturing of the epithelial surfaces allows an earlier detection of the inflammations. In the pyloric antrum, the NBI modality makes it possible to better detect intestinal metaplasia and to characterize dysplastic areas.

In urology, cystoscopes enable the detection and follow-up of various bladder lesions (polyps, multi-focal cancers, etc.). During the examination, the bladder is

(a) (b)

Figure 1.2: Gastroscopic images of the antrum (a) The white light modality visualizes the textures and colours in a natural way, (b) while the NBI modality facilitates an early detection of an antral gastritis due to the diffusion of intestinal metaplasia. Images taken from [Par].

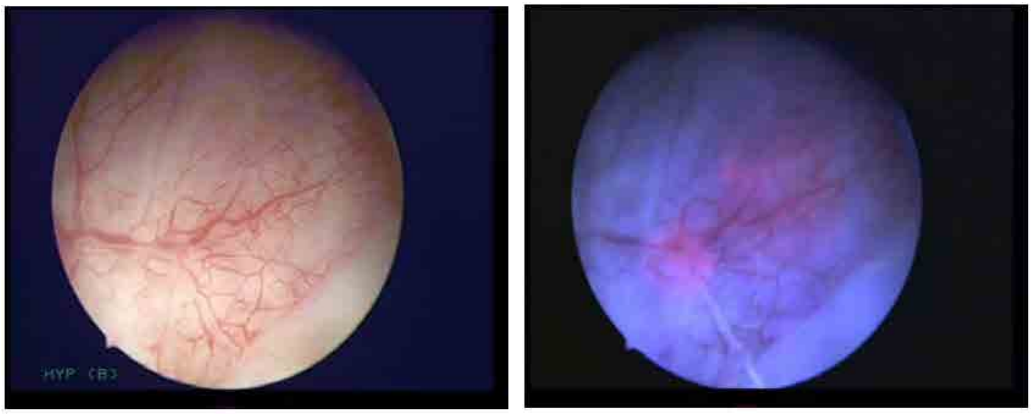
filled with a saline isotonic liquid which, on the one hand, swells the organ and on the other hand, "stiffens" the wall (the shape of the surface remains relatively constant during an image sequence acquisition). The cystoscopic video sequences are classically acquired in the white light modality, see Fig. 1.3(a) or sometimes in the fluorescence light modality (FL, see Fig. 1.3(b)). For the bladder, white light is the standard imaging modality and the exploratory examination of reference. FL induced cystoscopy is a complementary modality recommended by the European Association of Urology for Carcinoma In Situ (CIS) patients because it improves the detection rates and allows for a more complete removal of tumors. Simply put, CIS is a carcinoma that, in its early stages, is located under the first layers of the epithelium. It is therefore not visible in white light [Wei13]. The fluorescence modality allows for a deeper vision of these first layers, and thus an earlier detection becomes possible, to the detriment of an unnatural representation of the textures of the epithelium.

Gastroscopy, cystoscopy and other endoscopic examinations have all a common point: the images are acquired close to the tissue to ensure high image resolution.

1.1.2 The need of extended surface representations

1.1.2.1 Limitations of endoscopic images in terms of interpretability

In comparison to the numerous scenes for which a camera acquires large object parts from a distant viewpoint, endoscopes acquire images close to the epithelial surfaces. The major limitation of such acquisition conditions is that the clinician can only perceive epithelium areas through a limited field of view (FOV). In an endoscopic video-sequence, regions of interest usually spread over tens or hundreds of images. A unique image only visualizes a limited part of the complete region of interest. The consequences are multiple from the medical point of view.



(a) (b)

Figure 1.3: Cystoscopic imaging modalities. (a) Reference white light modality which facilitates the navigation inside the bladder. (b) Fluorescence modality enabling an early detection of carsinoma in situ. Images taken from [Wei13].

1. In images with a limited FoV, lesions (like multifocal cancerous lesions in cystoscopy or inflammations in gastroscopy) can only be partially observed. The partial view of lesions does not facilitate the diagnosis.

2. In organs like the bladder, urologists or surgeons cannot simultaneously observe lesions and anatomical landmarks (like the urethra or the ureters). However, during the diagnosis, endoscopists have to mentally reconstruct and understand the scene. To do so, the endoscopist displaces the instrument with back and forth and/or zigzag movements in order to alternatively see the regions with lesions and landmarks. This procedure is tedious and time-consuming.

3. Another drawback relating to previous point is that a recorded endoscopic video alone (i.e., without an endoscope in the hand) is often not sufficient for a clinician to mentally reconstruct the scene and to easily interpret the images. That is the reason why in examinations like urology and gastroscopy videos are standardly not recorded. The consequence is twice: one the one hand, no media exists after an examination to favour an information exchange between specialists of different medical fields and, on the other hand, there is no examination traceability.

4. Two video-sequences acquired for a same patient at a time interval of some weeks or month are inappropriate for a comparison of their content. Thus, an assessment of a lesion evolution cannot be performed.

5. Last but not least, the wall of a hollow organ such as the bladder or stomach must be scanned without any gaps. With a reduced field of view, it is difficult for an endoscopist to be certain that this goal has been achieved and lesions can be missed.

All these facts highlight the interest of extended FoVs in endoscopy.

1.1.2.2 Image mosaicing : 2D approaches versus 3D methods.

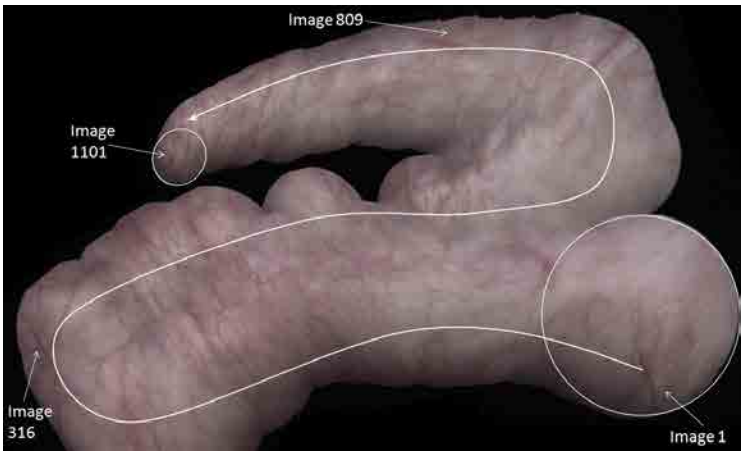
The FoV of a scene can be extended by calculating either 2D or 3D panoramic images. Whatever the chosen approach, the computed mosaics must give a visually consistent rendering without structure, texture or colour discontinuities due to the placement of pixel information from different images into the common mosaic coordinate system. When passing from an image sequence to the mosaic, a loss of resolution must be avoided and the resolution must ideally remain equal to that of the images, regardless of the location in the extended field of view map.

In the last two decades, 2D image mosaicing algorithms were proposed in the fields of urology (bladder wall, see [WDW+12, BSGA09]), gastroscopy (stomach, see [ADGB16, TDBL18]) and other endoscopic modalities [SLH06, CS07]. The different parts of classical 2D mosaicing approaches are the following.

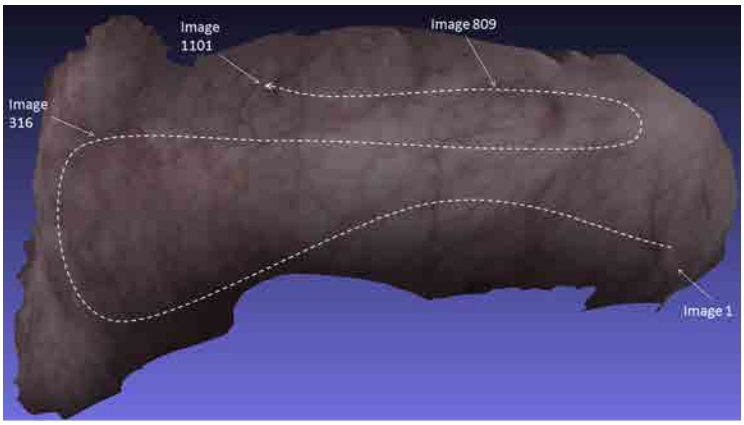
- Finding the geometrical link between the pixels of consecutive image pairs. This geometrical relationship is given either by global transformations linking

all pixels of an image pair (e.g., a homography as in [WDW+12]), or by vector fields, each vector linking a given pixel pair [ADGB16].

- The known geometrical relationship between image pairs is then used for an initial placement of the pixels into the common coordinate system of the 2D mosaic [LDB+08].
- Global bundle adjustment methods are then used to simultaneously adjust either the parameters of the global geometrical transformations between images or directly the individual pixel positions [WDWR12], to minimize the structure and texture discontinuities between images.



(a)



(b)

Figure 1.4: 2D and 3D mosaic examples. (a) 2D mosaic constructed with 1101 images [Ali16]. The mosaic path is materialized by the white curve, the pixels of image $\mathbf{I}_i$ being added to the current mosaic built with information of images $\mathbf{I}_1$ to $\mathbf{I}_{i-1}$. As visible, the first ($\mathbf{I}_1$) and last ($\mathbf{I}_{1101}$) images correspond to ellipses with very different areas in pixels. The resolution of an image in the mosaic plane depends strongly on the viewpoint changes of the endoscope between images, leading to both strong image distortion and significant resolution losses (only image $\mathbf{I}_1$ is without resolution loss). (b) 3D mosaic without gaps contructed with the same images as the 2D mosaic in (a).

- In a last step, colour discontinuities (due to the fact that different viewpoints of the endoscope lead to orientation changes of the illumination source) are compensated by selecting the pixels whose colours minimize the discontinuities or by smoothing the colour transitions by a weighted averaging of the colours of neighbour pixels [WDWR12].

In the specific case of cystoscopy (the only field with numerous publications in terms of 2D endoscopic mapping) lessons can be learned from the results obtained. In the first state-of-the-art publication in this field [LDB+08], the pixels were iteratively placed in the mosaic coordinate system (pixels of a new image were added to the mosaic which iteratively grows after each registration of an images pair). Even with a robust and accurate image registration, the errors that accumulate along the video sequence led very quickly to visual incoherencies (discontinuities), even for a moderate extension of the 2D map area. The results have also shown that even elaborated algorithms [WDW+12] for global map corrections lead to visually consistent maps only for limited field of view extensions.

More globally (i.e., for different endoscopic modalities), although 2D mosaics increase the FoV, such a representation of human organs has two major drawbacks. On the one hand, the 3D organs are projected on a 2D plane defined by the image taken as a reference for the mosaicing. When moving away from this reference image in the mosaic plane, the projection distortions become strong and result both in a loss of image resolution and in an incorrect organ representation at the borders of the mosaics which remain of limited size (see Fig. 1.4(a) which shows an extended bladder FoV mosaic with gaps and strong resolution changes leading to a strong loss of resolution for a large map part). On the other hand, 2D mosaics are not in accordance with the 3D mental organ representation of endoscopists.

Obtaining extended 3D FoV mosaics can be of high interest in endoscopy since the drawbacks of 2D mosaics can be avoided when reconstructing a 3D surface and projecting image textures on this surface from appropriate (known) camera viewpoints.

- If the shape of a reconstructed surface is close to the true (ground truth) surface shape, then the projection of the image textures from the known camera viewpoints lead to a resolution which is relatively constant on the surface (more or less independent from the camera viewpoint and movement) and whose value is at least close to the resolution of the images (super-resolution techniques could even increase the resolution).

- The correction of colour discontinuities can benefit from additional information: local surface orientations can be used to correct colour discontinuities due to the changes of local surface orientations and camera viewpoints.

- In 3D mosaics, texture gaps (surfaces with holes) are due to surface part which were not scanned by an endoscope and are not due to accumulating mosaicing errors as in 2D mosaics (this point can be verified by comparing Fig. 1.4(a) and Fig. 1.4(b)).

To sump up, 3D image mosaicing can lead to a large field of view extension, while ensuring the best possible texture resolution (close to that of the images) on the surfaces. Compared to 2D mosaics, 3D mosaicing improve the visual coherence in terms of texture, structure and colours discontinuities.

1.2 3D reconstruction for endoscopy

The aim of this section is to present the different published 3D mosaicing methods and their application in various endoscopic examinations. An analysis of the advantages and drawbacks of these methods in endoscopy allows for choosing a structure from motion based pipeline as a potential solution for the 3D cartography of hollow organs. The structure from motion step of classical pipelines is also presented.

1.2.1 Principles of 3D reconstruction approaches

The 3D reconstruction is the process of recovering the shape or structure of the surfaces that are seen in images. This research topic, which has occupied researchers for decades, is still topical, particularly because of the numerous application fields of 3D scene construction: computer aided geometric design (CAGD), computer graphics, computer animation, computer vision, medical imaging, computational science, virtual reality, digital media, etc. [JR12, HSDF15, FGG+10]. It is noticeable that, according to the application, the aim of the 3D reconstruction can either be to find the exact dimensions of objects (as in the field of dimensional measurement of manufactured parts), or just to recover the shape of surfaces independently of a scale. 3D reconstruction methods can be grouped into two main categories of approaches referred to as "active" and "passive" vision reconstruction methods.

The active vision methods are all based on devices that project artificial light with perfectly controlled characteristics into the scene to be reconstructed. This artificial light, reflected by the surfaces of the scene and acquired with a camera, is the information used for the 3D reconstruction. Two active vision approaches are often mentioned in the literature, namely those based on time-of-flight (ToF) cameras [LSKK10, PHS+09], and those based on structured light [BHSD+13, SFS+12].

In contrast to active vision methods, passive vision approaches exploit only the content of images acquired from different viewpoints by one or several cameras. Besides the images, these approaches do not require any additional information. While stereoscopy was historically the first passive 3D vision approach, more elaborated and recent methods became popular: Simultaneous Localization And Mapping (SLAM, [Dav03]), monocular Shape from-X (SfX, [CTS95, BHB00, SF16]) that notably include Deformable Shape-from-Motion (DSfM, [BGCC12]), Structure-from-Motion (SfM, [SF16]) and Shape-from-Shading (SfS,[ZTCS99]).

Both active and passive vision based methods led to accurate and robust 3D scene reconstructions in numerous application fields like industrial quality control [GBDH95, Die16], robotics [RSEM09, KCCH11], geology [ZSZ+09, JR12], etc. However, as illustrated by the literature relating to the 3D reconstruction of endoscopic

scenes, the implementation of active or passive vision solutions for complex medical scenes remains a real challenge.

1.2.1.1 Active vision systems in 3D endoscopy

A strong point of active vision systems lies in the fact that no homologous image features (structures or textures) have to be segmented and matched to reconstruct 3D points. Such structure/texture information which can be missing or very difficult to be extracted for endoscopic scenes is replaced by an information provided by the light emitted by a projector.

3D reconstruction attempts in endoscopy with ToF systems. The ToF technique is an active reconstruction method which exploits the time required by a light emitted by an illumination unit to travel to an object and to come back to a detector. The principle of ToF has been used for the development of new range-sensing devices based on standard CMOS (Complementary Metal Oxide Semiconductor) or CCD technology which equips ToF cameras [HMB+13]. Such cameras often consist of the following components: illumination unit, optics, image sensor, driver electronics and distance computation unit [TOF]. The principle of a ToF camera reconstructing 3D points is sketched in Fig. 1.5.

The phase-based ToF technology suffers from some specific problems that cause systematic calibration errors and parameter correlation issues. Due to the physical realization of light modulation with light emitting diodes (LEDs are typically used

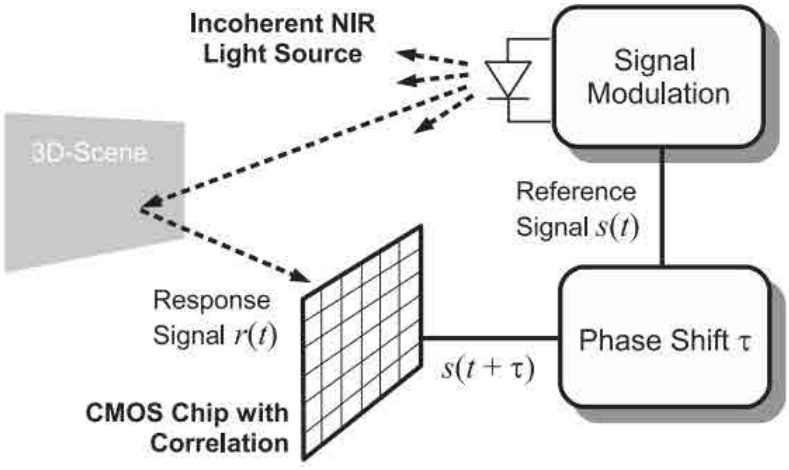


Figure 1.5: Priniple of the Photonic Mixer Device (PMD) for ToF-measurement. Its aim is to estimate full-range distance information in real time by illuminating the scene with modulated infrared light and determining the phase shift between the reference signal and the reflected light [LSKK10]. The illumination units of the camera emits intensity modulated near infrared light (NIR), which is activated by an internal reference signal. The emitted signal is reflected by the surface of objects and detected by appropriate smart pixels of the ToF camera. The illustration is taken from [LSKK10].

by the illumination unit of ToF cameras), the ideal sine-waveform light emittance is approximated by a band-limited rectangular waveform. This causes nonlinear depth distortions called wiggling errors. Further challenges to be addressed include so-called flying pixels (i.e. pixels that observe the regions with discontinuities in depth) [FAT11, HMB+13]. A lot of research currently try to minimize the sources of errors and their effect. For instance, the flying pixel errors can be attenuated using camera simulation techniques exploiting the intensity-based calibration model proposed by Lindner et *al.* [LSKK10].

The first ToF-based endoscope was proposed by Penne et *al.* [PHS+09]. A commercial ToF camera (PMD[vision]3k-S, PMD Technologies, Siegen, Germany) with a sensor resolution of 48×64 was combined with a rigid standard endoscope optics and operated at up to 25 fps. The authors assessed the ToF accuracy on phantoms (an in vitro excised, pig stomach). The measurement errors are estimated for each pixel by determining the standard deviation of the distances obtained when observing a static scene. An average precision of 0.89 mm and a median precision of 0.71 mm was computed from 100 acquired depth maps.

The major advantage of ToF-cameras lies in the fact that they provide both depth and intensity data for each pixel of the images which are acquired at a high frame rate using compact systems [HMB+13].

However, the ToF technology presents also strong drawbacks when it is used in endoscopy. It requires significant, complex and expensive hardware modifications when ToF cameras must be associated with standard endoscopic systems. The inhomogeneous scene illumination caused by the endoscopic optics impedes accurate range measurements using the infrared light [HM18]. This issue causes severe systematic distance error measurements and noise in endoscopy. Moreover, although ToF cameras can provide a high 3D point density, the acquired field of view remains small, and the 2D image resolution of classical endoscopes remains by far larger. This implies that classical 2D images must be acquired and exploited jointly with the (inaccurate) 3D point clouds provided by the ToF system. This is only one reason explaining why ToF systems are expensive and complex to build in endoscopy.

3D reconstruction in endoscopy using structured light systems. Structured light systems recover the 3D surface information of an object with a principle which is close to that of stereoscopy (passive vision stereoscopy is described in Subsection 1.2.1.2), but using an artificial light pattern. Structured light systems project a known pattern (often grids or horizontal bars) into the scene. These structured light systems exploit the parallax existing between the optical axes of the camera and of the projector. As sketched in Fig. 1.6, triangulation techniques are used to reconstruct the position of 3D points.

The structured light system must first be calibrated to perform 3D measurements. The calibration is usually based on the mathematical modeling of the projection of a light pattern in the 3D space, the equations being expressed in the camera coordinate system. The camera/projector system can be calibrated with the method described in [BHSD+13]. These authors proposed a calibration proce-

dure for active vision systems projecting different kinds of point patterns since their calibration method does not depend on the number, color and spatial distribution of the projected points. Besides that, no positioning device is required because the projector geometry can be determined in the camera coordinate system using only unknown positions of the calibration board. The principle of such a 3D reconstruction of small bladder surface parts was also described in [SFS+12] and validated on phantoms.

Inspired from [SFS+12], the CRAN laboratory has proposed a method to extend the 3D field of view of endoscopic bladder scenes [BDS16]. For each sensor position

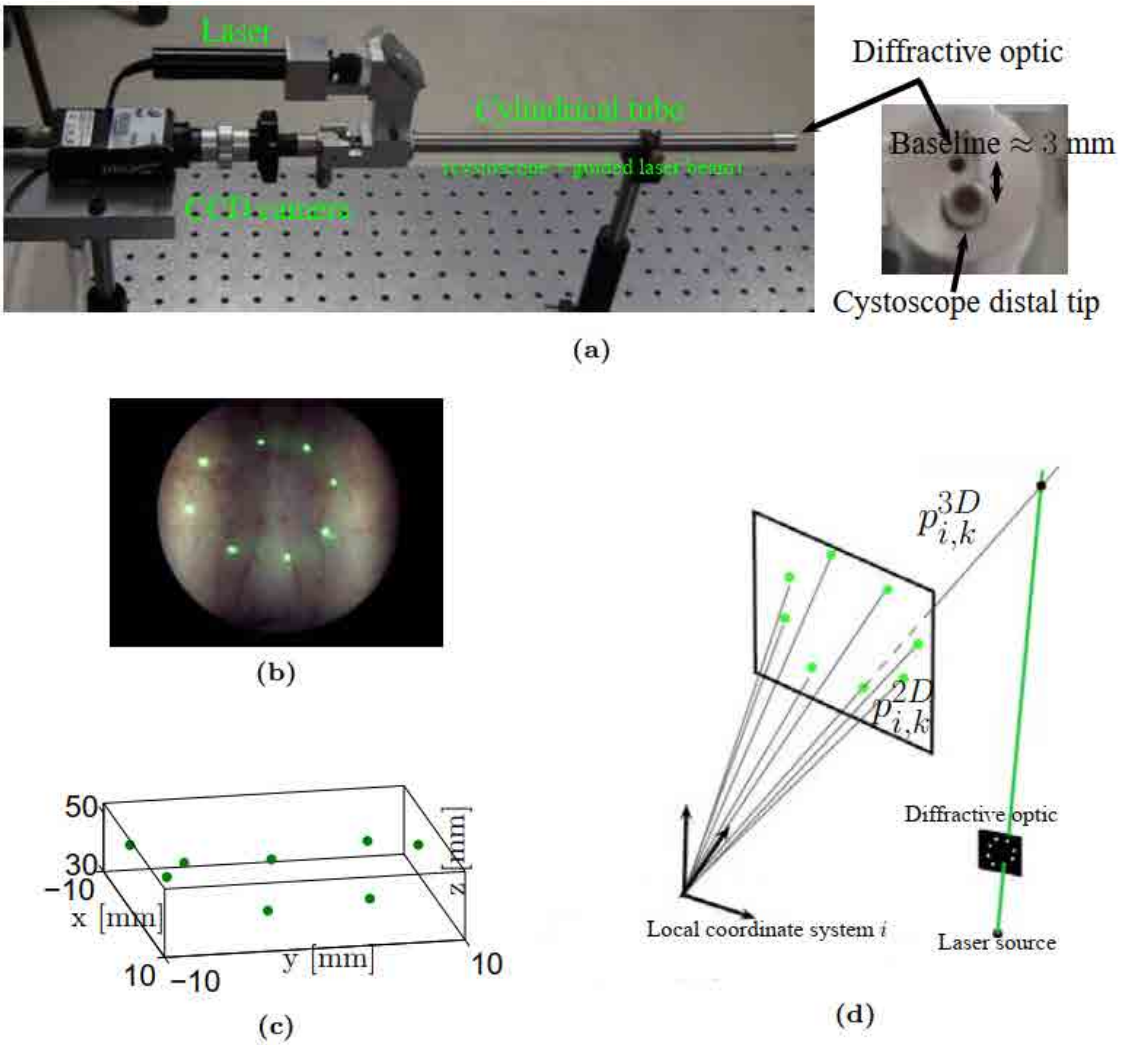


Figure 1.6: Active stereo-vision principle of the laser cystoscope prototype developed at the CRAN laboratory. (a) The prototype consists of a cylindrical tube, through which a cystoscope is inserted. (b) A set of eight laser rays is projected onto the color image. The small dot area allows to capture most of the bladder's texture (c) Reconstructed points corresponding to the laser dots of (b). (d) A triangulation technique is used to compute the 3D coordinates of the laser dot centers. To simplify the reading of the drawing, only one of the eight diffracted laser beams was drawn. Images visualization are borrowed from [BHSD+13].

it exists a set of 3D points known in the coordinate system of the camera and reconstructed with the principle described in Fig. 1.6. Since the camera moves between two acquisitions (rigid 3D transformation $\mathbf{T}_{3D}$) two point clouds are known in two different coordinate systems. An image is available for each viewpoint, and the assumption is made that between two consecutive acquisitions the geometrical link between homologous pixels can be precisely modelled by a homography $\mathbf{H}_{2D}$. To find the camera displacement between two images $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$, the authors in [BDS16] iteratively optimize the three translation and rotation parameters of $\mathbf{T}_{3D}^{i,i+1}$ so that the homography $\mathbf{H}_{2D}^{i,i+1}$ which superimposes the homologous pixels of the images maximizes a similarity measure (joint entropy [LDB+08]) between the overlapped parts of $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$. The maximum of this similarity measure gives the rigid 3D transformation $\mathbf{T}_{3D}^{i,i+1}$ which allows to place the two point clouds of the viewpoints of images $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$ in a same coordinate system. Applying this procedure along the video-sequences allows to increase the field of view. Test made in [BDS16] on realistic bladder phantoms have shown the feasibility of such an active vision system.

The advantages of this structured light approach are the high speed, accuracy and robustness of the 3D point reconstruction of featureless objects. However, the point positioning errors in a common coordinate system accumulate along the endoscopic trajectory so that a global surface correction is needed. However, the most limiting aspect of structured light systems is that too significant hardware changes are requested. In comparison to ToF systems, structured light based solution are more accurate in terms of 3D reconstruction. However, the hardware changes involve higher endoscope prizes.

1.2.1.2 Passive vision systems in 3D endoscopy

Passive vision approaches only exploit the content of images and have the advantage, in comparison to active systems, that they do not imply hardware changes of the existing endoscopic systems. All passive 3D reconstruction methods are based on the disparity of the scenic information seen in images acquired from different viewpoints.

3D reconstruction based on stereoscopy Stereoscopy exploits the parallax between two lines issuing from two homologous pixels located in two images acquired from different viewpoints and intersecting themselves onto the 3D coordinates of the scene point which project itself in the two images. Ideally (i.e. without numerical errors), the 3D lines effectively intersect themselves and the coordinates of the intersection point gives the 3D point position. In other words, the lines issuing from the homologous pixels in the images (and which have to be computed) correspond exactly to the projection trajectory of the 3D point onto the two images planes. For a precise triangulation, the angle between the 3D lines has to be large. Classical 3D reconstruction algorithms based on stereoscopy consists of four steps [SS02, HMB+13]:

- camera calibration (the calibrated parameters allow to compute the coefficients of the 3D lines issuing from the pixels),

- segmentation of particular points in the images acquired for different viewpoints,
- matching of the homologous points, and
- 3D point localisation in a scene coordinate system using triangulation techniques.

Stereoscopy was used in several reconstruction methods dedicated to laparoscopy [Sto12, RBS+12, SSPY10]. In [SSPY10], Stoyanov et *al.* proposed a technique for building a semi-dense real-time reconstruction of the operating field in minimally invasive surgery. This technique relies on the propagation of a sparse set of stereo correspondences into a semi-dense 3D structure by using a best-first principle growing scheme [LQ00]. The authors in [SSPY10] validated the effectiveness of their approach using phantom data with known ground truth. However, the method was not tested with patient data.

Stereoscopy has been early shown as a feasible technique for the determination of 3D surfaces using in-vivo laparoscopic images. It is currently the most widely tested technique in clinical practice since stereoscopic hardware is already implemented in some laparoscopes[1]. Such laparoscopes acquire simultaneously images from two viewpoints and display them on two screens or are used together with dedicated glasses. In these systems, it is the human brain which reconstructs the 3D information. But obviously, the image pairs of such systems can also be used to compute the 3D information.

However, multiple light sources, complex organ appearance, and procedure-dependent surgical devices impose stereoscopy algorithms which are specially designed for particular minimally invasive surgery situations. It is also noticeable that, when a 3D scene part can be reconstructed with such stereo-techniques, the extent of the 3D surfaces remains limited. Besides that, not all stereoscopes can be used for a 3D reconstruction using stereoscopy, as some laparoscopes acquire stereoscopic images using beam splitters and do not have a baseline between the two cameras. Moreover, all intrinsic parameters (distortion coefficients, focal length, etc.) of laparoscopes have to be calibrated before the surgical intervention using a simple and flexible procedure. Even if such a calibration procedure would be available, the focal length must be fixed during the examination since the position of the reconstructed 3D points, as well as the values of the distortion parameters, depend on the focal length. But, in laparoscopy, the distance between the endoscope's distal tip and the tissue often change so that the focal length has to be adjusted to ensure the acquisition of focussed images [HMB+13].

3D reconstruction based on Shape from Shading (SfS). The concept of SfS was introduced by Horn [Hor75] in the early 1970*s* and explains how a single image

[1] Some stereo systems are either already available or in development. One can mention the da Vinci® surgical System from the Intuitive Surgical company (CA, USA), the laparoscopes from Karl Storz GmbH (Tuttlingen, Germany), or the laparoscope from the Richard Wolf GmbH (Knittlingen, Germany).

can be used to recover the three dimensional shape of an object. SfM techniques are based on models giving the relationship between pixels intensities and their corresponding normals on the surface. Zhang et *al.* [ZTCS99] classified SfS techniques into four large groups:

- *minimization approaches* which obtain the solution by optimizing an energy function [ZC91],
- *propagation approaches* which propagate the shape information from a set of surface points (e.g., singular points) to the image [BP92],
- *local approaches* which determine a 3D shape based on the assumption of a surface type (i.e. using a priori knowledge about the surface, for instance, the spherical assumption as in [LR85]) and
- *linear approaches* which compute the solution based on the linearization of the reflectance map [Pen88].

The most common assumptions of existing methods are: (*i*) the scene contains only a single light source, whether directional or proximal, (*ii*) the scene's reflectance is Lambertian (the light is reflected equally in all directions), (*iii*) the shape's albedo is constant or known (the surface albedo is the reflectance coefficient that gives the fraction of light energy reflected by the surface as a function of the wavelength of the incident light), and (*iv*) the shape is continuously differentiable so that its projection in the image does not create discontinuities [HMB+13].

SfS techniques were successfully used in endoscopy to reconstruct bone structures [WNJ10], for a realtime visualisation of 3D surfaces in monocular laparoscopic videos [CA12], to perform 3D reconstructions in capsule endoscopy videos [PFFK12] and for the 3D reconstruction of other endoscopic scenes [OD97, FT00, YTY99]. Yeung et *al.* [YTY99] proposed a SfS method which is based on the identification of singular points (e.g minima, maxima and saddle points) in a distance map. The method in [YTY99] uses these singular points and a level set propagation algorithm to obtain maps giving the distance from each surface point to the light source. The method merges then the distance maps based on the knowledge of homologous singular points. Finally, it projects the distance map back to the 3D coordinates to get the depth map of the object. Experimental results obtained on simulated and real data have shown that a quite realistic surface reconstruction is possible with this approach. Besides that, Forster et *al.* [FT00] obtained promising results for the reconstruction of single internal stomach images. It is assumed that a spherical projection model accurately represents the camera projection geometry and the image distortion parameters are estimated using a set of calibration images. To obtain a Lambertian surface close to that of image, a dichromatic model is used as the Lambertian surface assumption to remove the specular reflection component present in that image. However, the authors in [FT00] did not give an assessment of the algorithm accuracy.

The advantage of SfS methods over active vision techniques (structured light and ToF approaches) is that they can be used with almost any hardware (usually they

do not require the modification of the standard endoscope hardware). However, their main drawback is that only small surfaces can be reconstructed due to the "restrictive" assumptions (known and constant light source characteristics, surface reflectance, etc.) made by SfS algorithms.

3D reconstruction based on structure from motion (SfM). The aim of SfM methods is to recover 3D structures of a stationary (non-deformable) surface using a set of 2D images. SfM methods require the knowledge of numerous and accurate point correspondences between images. The correspondence of homologous points is usually given in the form of point-tracks which are used to simultaneously estimate the camera trajectory and a 3D point cloud located on the surface to be recovered. SfM is a widely employed technique that is able to reconstruct a great variety of scenes using only images acquired from different viewpoints [JR12, COSH13].

In [SPS12], Soper et *al.* proposed a SfM based surface reconstruction of the bladder wall using cystoscopic video frames acquired with an ultrathin and highly flexible endoscope coupled with a robotic steering mechanism. The acquisition conditions are perfectly controlled since the robotic system places the optical axis perpendicular to the surface to be reconstructed and the images are acquired along a spiral shaped trajectory which allows to scan the complete surface and ensures numerous and large image overlaps. Besides the fact that such an acquisition system is not used in clinical practice, well controlled acquisition conditions cannot be ensured when using standard rigid of flexible cystoscopes. However, surface reconstruction tests in [SPS12] were successfully conducted on a pig bladder phantom. Although no test on human data was performed, the results obtained in this work show the potential of SfM in cystoscopy when numerous homologous points can be matched.

In [LAZ+17], Lurie et *al.* proposed a SfM-method for the dense 3D reconstruction of the bladder wall using white light cystoscopy videos that are aquired with standard clinical cystoscopic systems. However, the method is based on the assumption that a significant amount of homologous points can be extracted and matched using the scale invariant feature transform (SIFT) method [Low04] for almost all images. This assumption is often incorrect in endoscopic scenes since large image regions may be without textures due to radiotherapy or surgical intervention for lesion removal for instance.

The works in [SPS12, LAZ+17] prove the feasibility of the reconstruction of endoscopic scenes using SfM methods. However, the use of SfM with very particular hardware for controlling the acquisition conditions is not in accordance with real clinical conditions. Moreover, making the assumption that sufficient texture or structure information is available in endoscopic images is not always true in cystoscopy, and is even wrong in other endoscopic examinations. For instance, gastroscopic images are characterized by a lack of textures and strong illumination changes from one image to another. That makes existing feature detection and matching method, as well as other methods for determining point correspondences, inoperative. Consequently, the current state of the art of SfM algorithms cannot reconstruct 3D surfaces for the complex scenes of endoscopy such as gastroscopy.

3D reconstruction based on Simultaneous Localization And Mapping (SLAM). SLAM, sometimes referred to as online SfM, is a sequential and real-time technique for simultaneously estimating the 3D scene structure (mapping) and the camera pose (localization and orientation). It is a fundamental technique in robotics, since it provides crucial information for autonomous navigation of cars, drones and consumer robots [DRMS07]. SLAM was largely developed by the robotics community and some of the outstanding approaches that can be mentioned are MonoSLAM [DRMS07, GBC+14], PTAM (parallel tracking and mapping) [KM07], ORB-SLAM [AT17] and DTAM (Dense tracking and mapping in real-time) [NLD11], etc.

SLAM algorithms were successfully applied in endoscopy to recontruct 3D surfaces. For example, a visual SLAM algorithm based on EKF (Extended Kalman Filter) was validated with human in-vivo endoscopic videos [GBC+14]. Chen et *al.* [CBA+19] proposed a SLAM approach that incorporates depth predictions made by an adversarially-trained convolutional neural network (CNN) to produce dense surface models of ex-vivo porcine colon tissue. Mahmoud et *al.* [MCH+16] used ORB-SLAM to reconstruct a semi-dense map of soft organs. Their experimental results on in-vivo pigs, shows a robust endoscope tracking even with organs deformations and partial instrument occlusions. The use of ORB-SLAM in [QR18] also enabled to recover map points and trajectories of the endoscope in oral cavities.

SLAM is a mature approach and works in rigid (or almost rigid) environments like the colon. The accuracy of these algorithms rely on the avalaibility of long point tracks (numerous point correspondences). Although few results have been achieved on phantom pig data, applying SLAM-approaches on patient data like stomach or colon videos remains a real challenge due to large changes in lighting conditions, specular reflections, partial occlusion and tissue surfaces without textures. These acquisitions conditions make that the feature point tracks that are usually required in SLAM approaches cannot be robustly and accurately determined.

3D reconstruction based on a combination of techniques. SfS techniques alone are not the most efficient for cystoscopic or gastroscopic scenes in which the illumination conditions drastically change with the viewpoint. Several works [KW08, ZPN+16, WPZ+17] have associated SfS with SfM methods in order to simultaneously exploit shading and feature information for the reconstruction of surfaces from endoscopic images. Common points can be mentioned for these works. A SfS algorithm is first used to reconstruct the 3D geometry of small inner surface parts seen in each images. Feature point tracks are also determined on the images and a SfM method produces a sparse 3D points cloud and more importantly estimates the camera positions (rotation and orientation) for each image. The camera positions are used to fuse the multiple partial surfaces given by the SfS step. The optimization and the small surface outlier rejection algorithms are also integrated into the process of fusion of multiple partial surfaces to achieve a final 3D surface rendering.

One of the main challenges when using SfM and SfS is that all individual reconstructions (SfS methods reconstruct the surface for individual images) are only

partially overlapped due to the constantly changing camera viewpoint. Moreover, the final (fused) surface may have missing data (holes) due to surface parts that can be occluded. In addition, with the joint use of the SfS and SfM methods, it is still very difficult to produce a full surface reconstruction using endoscopic data, mainly due to too poor shape priors, arbitrary surface reflectance and strong illumination changes.

In [MWP+19], the authors proposed a real-time reconstruction of small colon parts by combining a SfM and a SLAM method. The authors used the popular COLMAP SfM algorithm to build a large database of colon surface parts. These 3D reconstructions were then used to train a recurrent neural network (RNN) to predict depth maps and camera poses of consecutive images. The predictions are exploited in real-time by a deep learning driven-SLAM algorithm to construct small colon surface parts. One strong point of this method lies in the real-time visualization of colon surface parts which notably informs clinicians whether some regions of interest (which potentially include lesions) were completely scanned of not. However, the length of the reconstructed colon parts remains rather limited since the predicted depth maps and camera displacements quickly lead to divergent diameters of the colon. This is probably due to the fact that the colon surfaces reconstructed by the SfM method are taken as ground truth during the learning stage of the neural network. Colon images include few texture information and are affected by strong illumination changes. As shown in Subsection 1.2.4, feature based approaches of standard SfM algorithms lead to very inaccurate surface reconstructions for such endoscopic data. Even if the approach proposed in [MWP+19] is very elaborated and has a high potential in endoscopy, its main limitation are due to the inaccurate ground truth information provided by the SfM algorithm.

1.2.1.3 Global discussion about 3D endoscopy

The aim of this thesis is to propose a 3D mosaicing algorithm for endoscopic scenes under the assumption that the observed surfaces are (almost) rigid. The selection of an appropriate reconstruction approach is crucial when the aim is to construct extended FoVs in a robust and accurate way.

Active vision techniques are based on the controlled projection of an appropriate (e.g., invisible or contrasted) light into the scene. These vision systems can usually be accurately calibrated and the reconstruction process is not affected by missing textures or changing illumination conditions. Even if in industrial applications such techniques lead to robust and accurate 3D reconstruction solutions, obtaining an accurate 3D reconstruction for a clinical scene is difficult since the implementation of an active vision principle in an endoscope is challenging. The attempts to build laser- or ToF-based medical endoscopic systems highlighted the technical difficulties to meet an accurate and robust solution. Moreover, an active vision system leads to significant and expensive hardware changes and they reconstruct only small epithelial surfaces with data from one acquisition. The extension of the surface remains also an unresolved issue with active vision solutions. For all these reasons, the choice was made in this thesis to focus rather on a passive vision solution.

Passive vision systems reconstruct 3D information only with 2D images and most of the techniques are able to increase the extent of the surfaces. Although numerous attempts were made in endoscopy to reconstruct organ surfaces, the proposed solutions remain often inaccurate (or even fail) when few textures are available or can only work in particular and constant scene conditions. For instance, SfS algorithms can be used to reconstruct the 3D structure of tissue surfaces from a single image including few textures, but they also rely on strong assumptions (Lambertian surface, single light source and stable illumination) which significantly restrain the scene types for which such an approach is usable, especially in endoscopic examination with varying illuminations condition, changing viewpoints and specular reflections.

Stereoscopy was one of the earliest tested solutions in endoscopy. Even if stereo-laparoscopes are the only commercially available 3D medical endoscopes, they are limited to particular and very controlled scene conditions.

SfM or SLAM approaches (SLAM is a particular case of SfM) are the techniques which can naturally extend the surfaces, and their accuracy and robustness depend only the availability of textures and/or structures in the images. When numerous feature points can be segmented in the images and when many homologous points can be found between the images, then the structure (shape) of surfaces can be recovered robustly and accurately, even for uncalibrated cameras. This is the reason why in numerous applications SfM-based solutions led to efficient algorithms. Also in endoscopy (especially in urology [LAZ+17, SPS12]), SfM gave promising results. Moreover, when features points can be segmented and matched, there is no interest to use a SfM technique in combination with other 3D reconstructions methods since SfM techniques already reconstruct dense point clouds of extended FoVs.

Analyzing globally all the advantages and drawbacks discussed in this section, the choice was made to study a SfM based solution for the reconstruction of the inner wall of hollow organs like the bladder and the stomach. This choice will also be discussed from the scientific and medical point of view at the chapter end. It is also worth noticing that it exists a particular class of SfM methods referred to as deformable SfM (DSfM). DSfM methods deal with surface deformations. DSfM was notably used in laparoscopy [MBC11, BGCC12] to reconstruct small soft tissue surfaces. However, DSfM methods reconstruct a surface for each viewpoint (each image), but they cannot be used to extend a surface. This is the reason why such techniques did not appear in the bibliography of this section.

1.2.2 Generic SfM-based surface construction pipeline

As sketched in Fig. 1.7(a), SfM-based pipelines classically use five sequentially chained parts to construct surfaces from a set of 2D images acquired by a camera with calibrated or unknown intrinsic parameter values.

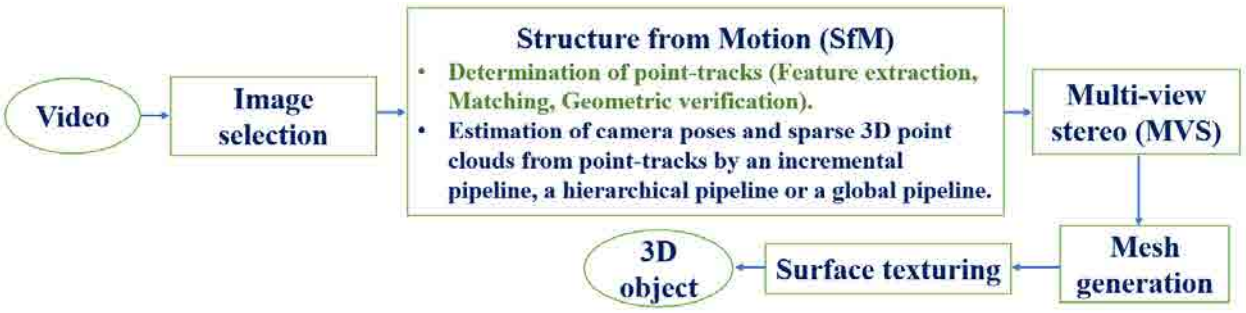
Pre-processing: In a video-sequence (as for instance acquired in cystoscopy or gastroscopy), the image quality depends strongly on the camera speed which varies (and produces motion blur), on the varying distance between the camera and the surface to be reconstructed (which causes defocusing and refocusing of images) and on the instrument orientation (which can lead to reflections). The aim of pre-

processing step is to remove images with low quality since only the most informative images should be used to build the hollow organ surface [AZB+20].

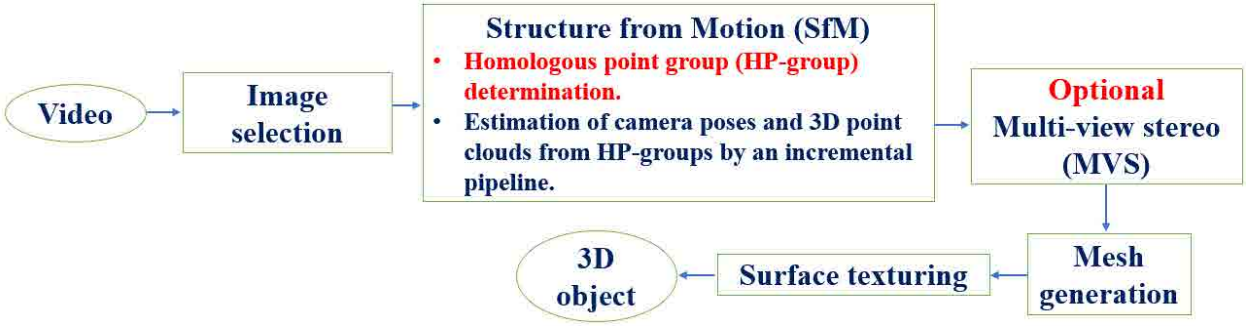
Moreover, the classical pinhole camera model (described in Chapter 2) consists of optical parameters (intrinsic parameters) which can be used to mathematically describe the projection of a 3D point onto the 2D image plane. Some of these parameters are related to a perspective projection, whereas other parameters describe the barrel distortions which affect images acquired with short focal length camera systems. The barrel distortion parameters can be computed, either with methods which calibrate the whole set of intrinsic parameters ([Zha00] describes such a state-of-the-art algorithm), or with methods which register non distorted images (computer-generated patterns printed on sheets of paper) and distorted images (acquisition of the patterns on the paper sheets) to obtain only the barrel distortion parameter values [MBD+04]. Whatever the calibration method, the barrel distortion model with known parameter values is used to compensate the image distortions.

According to the scene type, image constrast may also be enhanced to facilitate feature detection and matching.

SfM: Using the selected and undistorted images, SfM algorithms simultaneously determine the relative camera poses (extrinsic parameter values corresponding to a camera displacement between two acquisitions) for each image, and the scene structure represented by a sparse 3D point cloud lying more or less close to the surface to be constructed. For uncalibrated cameras, the SfM step also determines the intrinsic parameters. Numerous effective SfM pipelines have been proposed



(a)



(b)

Figure 1.7: Comparison of (a) the traditional SfM-based 3D reconstruction pipeline and of (b) the pipeline used in this contribution.

in the literature: *(i)* incremental pipelines [Wu13, SF16, NSR06], *(ii)* hierarchical pipelines [FFG09a, HTP10], and *(iii)* global pipelines [CT15, MMM13].

The common point of all SfM approaches lies in the initial information used to construct a cloud of points located all in a common coordinate system: all methods first determine the camera displacement (relative camera position and orientation changes) between two viewpoints (or two image acquisitions). For each image pair, a camera position corresponding to the first image acts as a reference and the computed extrinsic parameters correspond to the relative position of the second camera in the coordinate system of the reference camera. 3D point positions are also known in the reference coordinate systems and are determined using the homologous points which can be successfully extracted and matched for two images (this step is mathematically described in Chapter 2). A SfM method belongs to a given class of SfM pipeline depending on the way the relative camera positions are used to construct the surface.

In the next sections "camera" refers to a camera viewpoint and all the related data, whereas the term "registered camera" means that the position of a camera and the 3D points related to this camera were added to the surface and placed in a common world coordinate system. This terminology is usually used by the SfM community.

The SfM pipeline is referred to as incremental if cameras are registered one by one with the surface (i.e. with the point cloud) which is gradually growing by successively adding 3D points from the current camera. An incremental pipeline starts by selecting, among all camera pairs with known relative positions, the image pair with the most homologous points. This camera pair is used to perform an initial reconstruction (the result is a small point cloud which is the seed for the surface growing). At each iteration, the camera, whose image shares the most homologous points with the images of the already used cameras, is selected and registered with the surface. This process is iterated until all the cameras were added to the surface. Since the incremental approach is an iterative process, errors accumulate themselves during the surface growing. In order to reduce those drifting errors, most of the existing SfM methods use bundle adjustment techniques to refine both the camera poses and the 3D points by minimizing re-projection error [TMHF99]. In incremental methods, two types of bundle adjustments are usually done according to the number of involved cameras [Wu13, SF16]. If the surface has grown by at least a given percentage after a given number of images were registered, a global bundle adjustment (involving all cameras) is performed to refine the data (the point cloud and the camera poses). On the contrary, if the surface growing remains under the surface growing threshold, only a local (or intermediate) bundle adjustment is performed on the data of few cameras (the currently added camera and the cameras sharing common homologous points). Although the local bundle adjustment help to improve the robustness towards noisy data and incorrect relative poses, their frequent usage is computationally expensive. In addition, a bundle adjustment method is a non-linear optimization problem whose solution is sometimes trapped by the wrong local minima. This is especially the case for local bundle adjustments. That makes the reconstruction less accurate. For this reason, global adjustments are performed

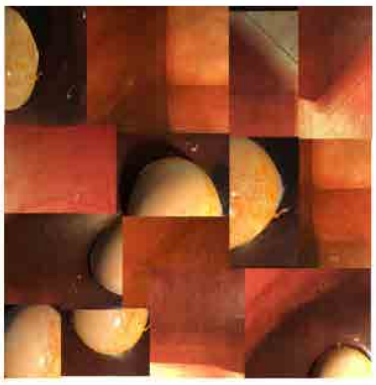
regularly.

For global SfM pipelines, the data of all cameras (camera poses and 3D points) are simultaneously registered using the data of all camera pairs with known relative motion [CT15]. To do so, these approaches first recover the global rotations of all cameras (this step is referred to as rotation averaging) and then determine global translations of all cameras (this step is referred to as translation averaging). Global SfM algorithms consider all geometrical links between camera pairs together and a bundle adjustment is only used one time to refine the final camera poses and 3D point cloud. Although global SfM algorithms could have a high potential to be more accurate than incremental approaches, their use remain challenging since the rotation and translation averaging steps are very difficult.

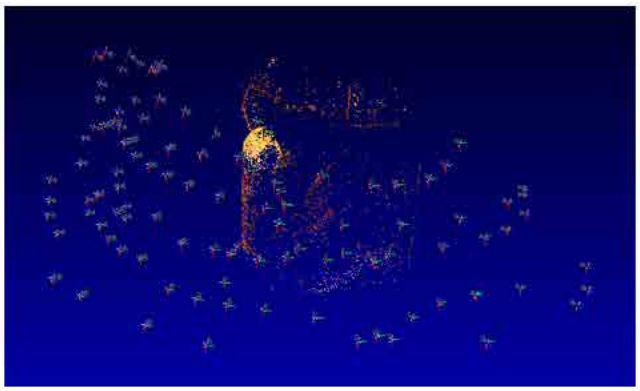
In hierarchical pipelines (or partitioning methods) the surface construction is reduced into smaller and better conditioned subproblems which can be effectively optimized [TGFF15, GFF10]. The hierarchical algorithms usually first determine the point matches between all image pairs, then they organize the available images into a hierarchical cluster tree by using a measure of overlap given by the distance which is described in [FFG09b, TGFF15]. Finally, the reconstruction proceeds hierarchically along this tree from the leaves to the root where images are stored in the leaves and each internal node (a cluster) performs a partial reconstruction for several cameras. Two-view reconstructions must be performed in a cluster and a view is added to a cluster if it observes existing scene points in the current reconstruction. Similarly to the incremental the incremental SfM pipeline, surface reconstructions are improved by using a local bundle adjustment algorithm, while the determination of the external camera parameters (orientation and rotation) of each view point is based on feature point tracks. Two clusters have to be merged when they share a common camera. Since the two sets of camera poses were determined in different coordinate systems, the cameras of one cluster must be registered with the cameras of the other cluster using a 3D similarity transformation. After this camera registration, the 3D points are then all re-computed using a triangulation technique applied to any point tracks that become visible after the cluster merging. This newly reconstructed point cloud and the camera poses are finally refined with a bundle adjustment method. Although hierarchical-based algorithms seem to be promising reconstruction solutions, their use remains very limited since existing algorithms as [TGFF15, GFF10] have shown that, on the one hand, this class of SfM methods is not robust when numerous cameras have to be treated and, on the other hand, the 3D reconstruction is time consuming.

The SfM algorithms based on the incremental approach have proven the ability to reconstruct a great variety of scenes [COSH13, JR12]. Incremental pipelines are mature paradigms and widely used due to their robustness and ability to deal with numerous cameras. Thus, this thesis will focus on the incremental approach to reconstruct extended epithelium surfaces from endoscopic videos.

Multi-view stereo (MVS): The goal of the MVS step is to determine dense 3D point clouds by jointly exploiting the sparse point cloud provided by the SfM step and a set of images taken from known camera viewpoints. MVS usually requires an accurate knowledge of the intrinsic and extrinsic camera parameters for each



(a) 2D sequence of undistorted images .



(b) SfM: the trajectory of the camera and a sparse 3D point cloud are computed in this step.

(c) Dense 3D point cloud given by the MVS step.

(d) Mesh generation and refinement.

(e) Texture mapping.

Figure 1.8: Illustration of a generic SfM-based surface construction performed in this thesis. Gastroscopic images were printed on paper sheets that were glued onto the cylinder which carries a sphere. The camera moves close to the cylinder surface so that small field of view images are acquired, as in endoscopy. The images acquired from different viewpoints present large overlaps.

image used to compute the dense point cloud. The intrinsic parameters do not depend on the camera's position and correspond to the focal length, the principal point in the image, the skew value and the distortions parameters. The external parameters are, in this step, the (absolute) position (3D translation) and orientation (3D rotation) given in a world coordinate system for each camera viewpoint along the sensor trajectory. It is well recognized that SfM methods allow to determine very accurate camera parameters, as well as precise sparse 3D point clouds [Vu11]. For this reason, MVS can effectively deliver dense clouds of accurate 3D points.

There exists numerous state-of-the-art MVS methods, such as Patched-MVS [FP10], CMPMVS [JP11], or MVS [SZFP16]. A comparison and evaluation of MVS reconstruction algorithms can be found in [YH15, SCD+06]. An example of the passage from a sparse point cloud provided by a SfM method to a dense point cloud

using a MVS method is shown in Figs. 1.8(b) and 1.8(c).

Meshed surface computation and refinement: The dense point cloud computed in the MVS step is used to construct a meshed surface. The meshed surface consists of triangular facets defined by three vertices. The meshed surface computation starts with a refining of the point cloud through a statistical-outlier removal [RBG+19]. Then, the Poisson surface reconstruction algorithm detailed in [KBH06] generates a mesh using the surface normals and the coordinates of the 3D points. The representation of the meshed surface is improved using a refinement step [VLPK12] leading to the final surface. An example of meshed surface computation after the refinement step is shown in Fig. 1.8(d).

Multiple view mesh texturing: Texture mapping is an important final part of the SfM-based pipeline to obtain 3D surfaces with coherent colors and structures. The superimposition of the 2D image texture information onto the meshed surface is crucial to obtain visually coherent scene rendering.

However, texture mapping on meshed surfaces is a challenging task due to the noise affecting the depth data, the geometrical reconstruction errors of surface parts [FYY+18], the large variability of the scale of the numerous available images, image blur, exposure variations from one image to another, and occluded surface parts [WMG14]. To tackle those difficulties, several algorithms based on per-vertex colors were proposed. In the algorithm described in [WMG14], the authors give a complete texturing framework for large 3D surfaces which is based on image registration (the pixels correspondence between registered images facilitates the surface texturing task). This texturing method, together with efficient SfM and MVS steps, led to the reconstruction of precise extended 3D surfaces with an impressive coherence in terms of texturing. Even already published in 2014, [WMG14] is still one of the most effective texture rendering methods. That is the reason why in this thesis one uses it for mapping the textures on extended 3D epithelium surfaces. A labeled texture example can be seen in Fig. 1.8(e) (triangular facts with the same colour are labeled with the same input image).

1.2.3 Overview on a standard SfM algorithm

Most of the current state-of-the-art SfM algorithms share the following main steps:

1. *Determination of the point tracks:* The objective of the first phase of a SfM method is to find homologous points between images, such "connected" points being usually referred to as "point tracks" (see Fig. 1.9 for a illustration of a point track along three images). A point track consists of a set of 2D image points which correspond all to the same 3D scene point acquired from different viewpoints. These homologous points are usually tracked along consecutive and non-consecutive images of a video-sequence. Most of the existing SfM approaches determine point tracks with three sequentially chained algorithms. First, feature points (e.g., corners) are classically detected in the images. Then, feature vectors (or descriptors) are computed in small regions centred on the detected points and are used to match the points between images. Finally, a

geometric check is performed with the RANSAC (RANdom SAmple Consensus) method [FB81] to rejected outlier points which were erroneously chosen as homologous points.

The accuracy of the homologous point localization, as well as the length of point-tracks play a key role in the SfM algorithms (only an accurate point localisation and long point tracks enable SfM approaches to reconstruct precise 3D point clouds). In numerous (textured) scenes, fast and robust feature detectors and descriptors (e.g., as SIFT [Low04], SURF [BETV08] or KAZE [ABD12]) enable to establish long tracks of accurately localised feature points.

2. *Reconstruction phase consisting of a projective reconstruction and a refinement:* In this SfM-phase, the point tracks, associated with a classical camera perspective projection model and a triangulation algorithm [HZ04] enable the simultaneous estimation of 3D point positions and of the relative camera poses. The relative camera displacements (3D translation and rotation), known for all image pairs of consecutive acquisition, are then used in one of the three SfM strategies described in Subsection 1.2.2. In this thesis, one uses the incremental SfM strategy to place all points in a common scene coordinate system. This approach uses a final and global bundle adjustment algorithm [TMHF99] to refine the camera parameter values (intrinsic parameters, extrinsic parameters, and distortion coefficients) and the 3D positions of the sparse point cloud.

1.2.4 Standard SfM approaches in the context of medical endoscopic scenes

As discussed in Subsection 1.2.3, standard SfM methods enable a precise and robust point cloud computation for numerous scene types in which textures and/or structures are available. The 3D point cloud construction is a well mastered issue under the assumption that homologous points can be tracked using classical feature detection and matching methods. However, there is a class of medical scenes in which the use of feature detection and matching in SfM approaches is not appropriate, or at least not optimal.

As shown in Fig. 1.10 for two pyloric antrum images, only few homologous points were found when associating the SIFT algorithm [Low04] to the RANSAC outlier rejection method [FB81]. Besides the lack of textures, homologous point determination is also impeded by the strong illumination changes between two acquisitions and inhomogeneous lighting due to viewpoint changes and vignetting effects of endoscopes, respectively. Specular reflections also favor false point correspondences. Such few and partially wrong matches are not appropriate for a 3D reconstruction using SfM approaches.

The acquisition conditions and the scene characteristics of medical applications are significantly different from those of the applications for which SfM has been proven efficient. First, the reconstruction of 3D points is more accurate when homologous points can be acquired from very different viewpoints. In classical SfM applications (e.g., manufactured part or monument surface construction), the ac-

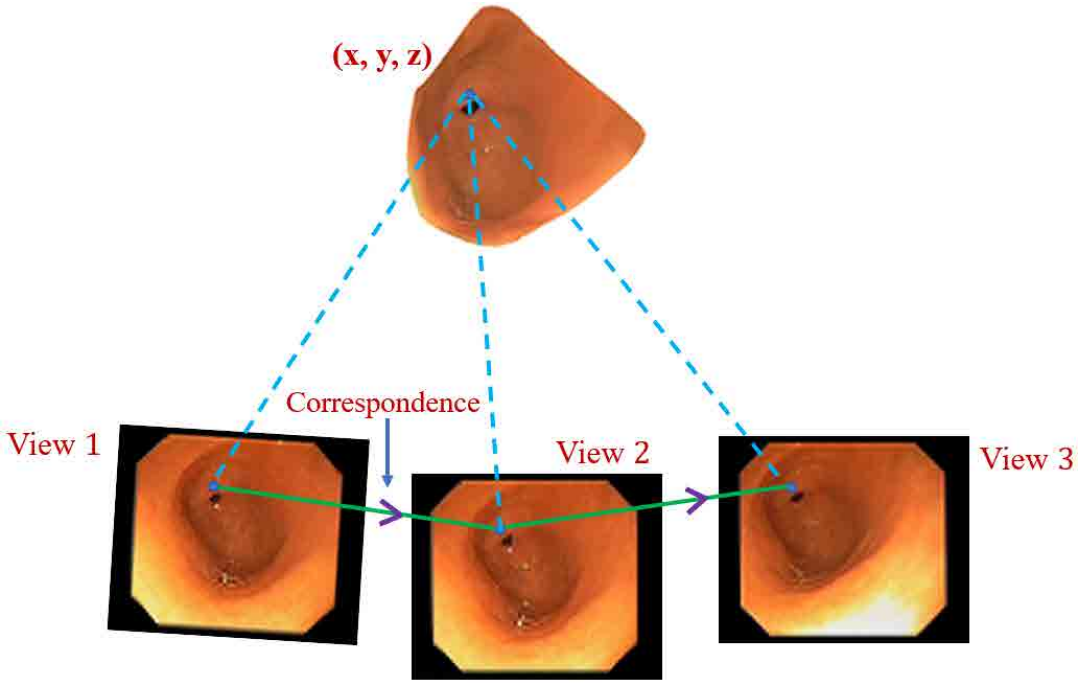


Figure 1.9: An illustration of a point track. Each point track corresponds to a 3D point in the scene.

quisition conditions are controlled in the sense that scene parts can effectively be acquired from very different viewpoints. In endoscopy, the camera trajectory is quite difficult to control. Obtaining images of the same organ part from very different viewpoints is a difficult task in endoscopy. Secondly, images of natural scenes or manufactured parts usually include image primitives (structures like corners, line segments, etc.), contrasted textures and/or a great variation in terms of colours. On the contrary, the color variations are smaller in endoscopy than in numerous other scenes, while in gastroscopy most images are with very few and weakly contrasted textures and structures.

These challenging scene and acquisition conditions make the classical SfM-based methods often inoperative when extended surfaces of hollow organs have to be constructed.

1.3 Thesis objectives

The main challenge of the 3D epithelium surface reconstruction lies in the determination of point tracks providing numerous point correspondences between images (first step of SfM), see Fig. 1.7(a). The proposed SfM approach is based on the fact that in scenes, where feature detectors are unusable, dense optical flow (DOF) can be used for the point correspondance establishment. If feature matching methods can detect a minimum number of feature points and achieve at least a certain number of point matches between two images, then DOF can be combined to feature matching in the SfM step for generating large 2D point groups. Thus, according to the scene content, the first step of the proposed SfM pipeline can be performed

either only with a DOF based matching method, or with a combination of DOF and feature data. This section points out the issues that need to be addressed in terms of scientific and medical objectives.

1.3.1 Scientific objectives of the thesis

In hollow organs, SfM methods were mostly tested in the specific case of cystoscopy. In [SPS12], the authors replaced the cystoscope by a non-standard system which acquires image sequences using an ultrathin fiber whose trajectory is controlled by a robotic steering system. The spiral shaped camera trajectory ensures numerous image overlaps and controlled camera viewpoints which favors robust SfM. Surface reconstruction tests were successfully conducted on pig bladders. Although no test on human data was performed, the results achieved in [SPS12] show the feasibility of SfM in cystoscopy. The method in [LAZ+17] confirmed this potential on clinical data. This method is based on the assumption that a significant amount of matching points can be determined using SIFT features for almost all images. However, this assumption is not always true. On the one hand, there is no warranty to obtain contrasted textures in all images (these textures are due mainly to blood vessels). Indeed, it is difficult to control the cystoscope trajectory so that the distance between the instrument's distal tip and the inner epithelial surface, as well as the endoscope speed can quickly change. These uncontrolled acquisition conditions lead to defocussing and motion blur, respectively. On the other hand, large image regions may be without textures due to surgical intervention for lesion removal for instance. This was the motivation in [ADWB13] for switching automatically between SURF feature extraction and optical flow (OF) for 2D bladder mosaicing.

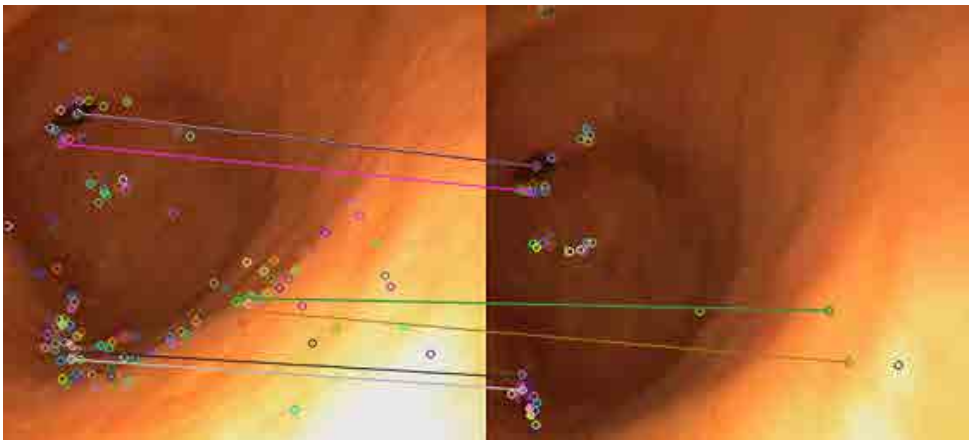
The literature shows that the epithelium surface construction from 2D image sequences remains an open issue (notably in cystoscopy) and is a real challenge in gastroscopy (at the best of our knowledge, no publication deals with 3D stomach cartography).

As mentioned in previous section, the accuracy and robustness of the homologous image point determination plays a key role in the SfM step of the surface construction pipeline. In scenes including rich information (images with numerous and contrasted structures and textures), feature based methods such as SIFT [Low04], SURF [BETV08] or KAZE [ABD12] are able to determine long tracks of accurately localized points, even in sets of images which are temporally unordered and acquired from different viewpoints (i.e., leading to illumination and scale changes for instance). For such image type (covering a large variety of scenes), object surfaces can be effectively reconstructed based on SfM. However, how to recover a 3D structure using a set of images for which feature matching techniques are inoperative?

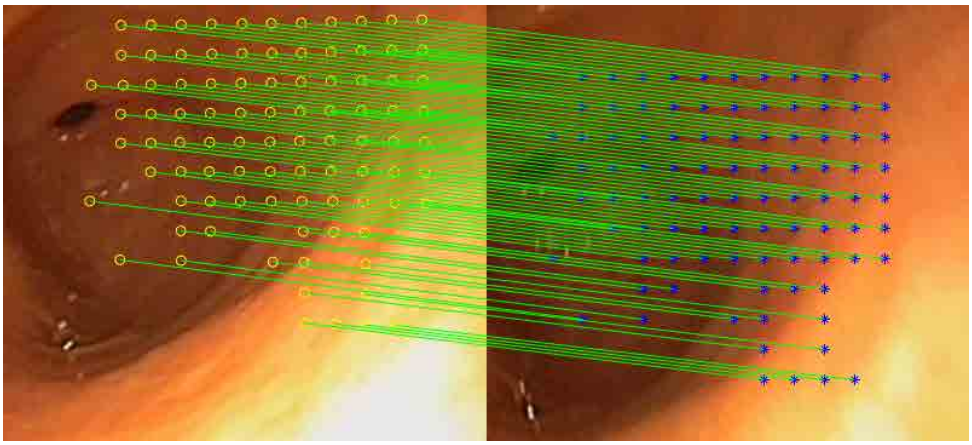
In particular in endoscopic imaging, feature-based techniques are often unable to establish robust point correspondences (see Fig. 1.10(a)). DOF may be a viable solution to cope with the poor information of endoscopic scenes. As shown in Fig. 1.10(b), the DOF approach described in this work is able to deal with scenes including only few textures and structures. Although a dense homologous point correspondence between two overlapping images can be obtained using DOF, point

matching using DOF was rarely used in SfM up to now. To understand the reason for this, let us consider following situation.

Suppose that $\mathbf{I}_i$ and $\mathbf{I}_j$ (with $j \neq i \pm 1$) are two non (temporally) consecutive images in a video-sequence and that they share a common scene part. If $\mathbf{I}_i$ and $\mathbf{I}_j$ include contrasted structures and or textures, feature detectors and descriptor (e.g., SIFT) can effectively determine (both in quantity and quality) the homologous points between them. The advantage of the feature matching methods is



(a)



(b)

Figure 1.10: Comparison of a feature based method and a dense optical flow approach taken as point matching algorithms in the context of gastroscopy. (a) Two consecutive gastroscopic images of a video-sequence (epithelium of the internal stomach wall). Numerous SIFT feature points (circles) were detected, but only six homologous point pairs were matched using the SIFT descriptors and the RANSAC outlier rejection method (segment lines). Among them, the green and light brown segment lines link false homologous points and correspond to wrong matches. Only few matches can be established mainly due to the fact that features points were most often only found in one image (their homologous point was not detected in the second image). In the lower left image corners some false feature points are due to specular refections. (b) Numerous homologous points found with the optical flow method proposed in this thesis for two gastroscopic images in (a).

that the points detected by detector (keypoints) and their descriptors are often invariant to geometric (e.g. scale or in plane rotations) and photometric (e.g., illumination) changes. Thus, point-tracks determined from images $\mathbf{I}_i$ to $\mathbf{I}_j$ by feature based methods lead often to subpixel accuracy in terms of localisation of matched points. On the contrary, if an OF-based tracking method is used to find homologous points between $\mathbf{I}_i$ and $\mathbf{I}_j$, flow fields $\mathbf{F}_{k,k+1}$ (with $k = i, i+1, \ldots, j-1$) have to be computed for consecutive image pairs $(\mathbf{I}_k, \mathbf{I}_{k+1})$ from $\mathbf{I}_i$ to $\mathbf{I}_j$. With a starting point A_i in $\mathbf{I}_i$, the sequence of tracked points $(A_i, A_{i+1}, \ldots, A_j)$ is determined, with $A_k = A_{k-1} + \mathbf{F}_{k-1,k}(A_{k-1})$ with $k = i+1, \ldots, j$, and A_j is supposed to be the homologous point of A_i:

$$A_i \xrightarrow{\mathbf{F}_{i,i+1}} A_{i+1} \xrightarrow{\mathbf{F}_{i+1,i+2}} A_{i+2} \ \ldots \ A_{j-1} \xrightarrow{\mathbf{F}_{j-1,j}} A_j.$$

Two issues are related to this way to track homologous points.

1. First, for an image pair, even if a very accurate OF method providing a dense flow field between images is used, it is impossible to reach the subpixel accuracy of feature matching methods.

2. Second, for an image sequence, although the errors affecting the OF vectors linking points in consecutive images are weak, these errors accumulate themselves along the sequence and become quickly large when the length of the point track increases. Therefore, A_i and A_j are often wrong (or at least very inaccurate) homologous points when the temporal distance $|j-i|$ is large. This lack of accuracy and of robustness explains why DOF is rarely used in SfM approaches.

The aim of the work of Trinh et *al.* described in [TDBL18] was to build 2D mosaics of the stomach wall. This work was notably based on a robust OF determination between images pairs. Indeed, in [TBD17, TD19], Trinh et *al.* have shown that variational OF using illumination invariant descriptors is one of the most effective approaches for the determination of homologous points between images without significant textures and with strong illuminations changes. However, the ability of finding robustly numerous homologous points between images pairs is not sufficient to build 2D panoramic images without data misalignments and illumination discontinuities. For this reason the authors in [TDBL18] developed a shortest path strategy to use consecutive and non-consecutive images which allowed the 2D mosaic to iteratively grow by minimizing the number of OF fields required for increasing the field of view. Although such OF techniques associated with a shortest path algorithm led to a robust and accurate 2D mosaicing algorithm of gastroscopic scenes, the extent of the mosaics without significant distortion (due to the 3D surface projection onto the mosaicing plane) remained limited.

Similarly to previous work dealing with 2D image mosaicing, the aim of the this thesis is to define

- (i) the best strategy to group the images with common scene parts in sets such that numerous homologous points can be found without a simple tracking along the image sequence, and

- (*ii*) to define new illumination invariant descriptors which lead to OF data-terms appropriate for images with few textures.

The strategy in point (*i*) must be adapted to the homologous point matching phase of SfM approaches: homologous point groups must be as large as possible, homologous points must be seen from different viewpoints (i.e. they must be extracted at least partly from non-consecutive images) and homologous points should be linked by as small as possible OF vector sequences. The aim of point (*ii*) is to find image region descriptors that are able to efficiently encode weak texture information under strong illumination changes. To sum up, the aim of this work is to extract in an optimal way the information of the 2D images to feed the 3D reconstruction part of a SfM algorithm that has already proven itself in the literature.

As mentioned in Subsection 1.2.1, classical SfM methods (based on features extraction and matching) deliver only sparse point clouds. MVS methods have to be used to densify the point clouds to facilitate the surface meshing step (see Fig. 1.7(a)). One scientific and practical objective of the DOF matching and image grouping strategy is also to obtain directly a dense point cloud with the SfM step (i.e. without MVS), as seen in Fig. 1.7(b).

Previous scientific objectives have been set for scenes in which almost no structure or texture can be detected and matched with feature based methods (see, for example, the gastroscopic scene in Fig. 1.10). In other endoscopic scenes, as for instance in cystoscopy, structures and textures are available, but are not systematically sufficient in all bladder parts for a robust 3D reconstruction. Another scientific objective is to propose a SfM approach which can jointly use DOF and feature point information to find the best comprise between the robustness, accuracy and computation time. In other words, the objective is to take advantage as much as possible of the accuracy of the detection and feature matching methods, while ensuring robustness when too few textures or structures are available.

The aim of this thesis is the construction of extended three-dimensional hollow organ surfaces. To reach this goal, the endoscopic scenes used to reconstruct 3D surfaces need to be (almost) rigid. Indeed, when a surface is deformable, only the set of 3D points corresponding to a same surface state (or shape) can be jointly used to reconstruct a surface part [BHB00]. It means that for each viewpoint (i.e., for each image) a set of 3D points to construct small (partial) surface parts which are difficult to combine for obtaining an extended 3D surface.

1.3.2 Medical objectives of the thesis

Different tasks which must be done in clinical routine (e.g., lesion diagnosis, patient follow-up, computer-assisted surgery, surgical planning, etc.), can be facilitated by accurate 3D models of the regions of interest of the stomach or the bladder.

The extended FOV surface achieved by the 3D reconstruction allows for a visualization of large hollow organ surface parts which include both whole lesions and anatomical landmarks. Such 3D representations enable a second diagnosis after the examination. This second diagnosis, performed either by the endoscopist

who acquired the data or by one of its colleagues, is usually very difficult to be effected on the video-sequence (for this reason, the video-sequences are usually not recorded during an examination). The extended 3D surface has not to be available during the examination itself, so that a real-time reconstruction is not required. SfM algorithms are usually time-consuming, but constructing an extended 3D map in about an hour remains an appropriate objective for a second lesion diagnosis or concertation between specialists of one or more medical fields.

Panoramic views given by a SfM algorithm can represent an offline solution offering an exchange media which clearly shows the appearance of regions of interest including abnormalities (e.g., polyps) or inflammations (e.g., inflammations in the pyloric antrum region of the stomach) that are detected in various image modalities. In gastroscopy and urology, endoscopic examinations are standardly performed under white light (WL). But carcinoma in situ are earlier detected in fluorescence (FL) cystoscopy, whereas inflammations (which can lead to ulcers or cancers) are easier to detect in the narrow band imaging (NBI) modality. For this reason, the proposed algorithm should work for these different images modalities (WL, FL and NBI), ideally with the same parameter values.

Two 3D mosaics computed for a same patient with image sequences acquired at an interval of some weeks or months enable, for instance, a visual assessment of a lesion evolution or of the remission of tissue after surgery. The SfM algorithm should lead to comparable surfaces even if the image sequences are not the same for the different examinations of a same patient (obviously the image sequences have to represent the same organ part). The SfM algorithm must be accurate enough to systematically lead to coherent surfaces (coherent surface shape and textures/structures without discontinuities).

During an examination, it is not easy for a urologist or a gastroenterologist to scan a complete region of interest without omitting some tissue parts. In the 3D mosaic, these “omissions” appear as “holes” on the 3D surfaces. Such a representation allows to check whether a region of interest (with potential lesions) was completely scanned. The presence of holes should be an indication that some regions were not scanned by the endoscopist. However, when locally (for some organ regions) only few 2D correspondences can be found, it may also happen that a SfM algorithm reconstructs surfaces with small holes. The homologous point group determination step of the SfM method should be robust enough to determine numerous correspondences even in regions with few overlapping images. Missing parts in surfaces should be due to the image acquisition performed by the endoscopists while being as much as possible independent of the image overlap quality.

Meeting all these application related conditions would lead to an improvement in terms of endoscopic data exploitation and examination traceability.

1.4 Conclusion

This chapter provided a general overview on the medical context and gave the thesis objectives. The SfM approach was selected for reconstructing the internal epithe-

lium surfaces of hollow organs using endoscopic videos, the targered organs being the stomach and the bladder. The first SfM-step is crucial for the feasibility and precision of the structure recovery. Most existing SfM methods are based on the assumption that point correspondences can be established by detecting and matching feature points. However, this assumption is rarely valid for endoscopic scenes. For instance, gastroscopic image sequences (see Fig. 1.10) are characterized by a lack of textures, strong illumination changes, specular reflections and small deformations of the stomach tissues. All these scene characteristics make the feature detection and matching algorithms inappropriate. Therefore, although 3D reconstruction have many steps, this work focus on the determination of homologous point groups that make SfM inoperative in endoscopy. Finally, this chapter gives some challenges when using optical flow for homologous point groups determination. The next chapter details all required and general computer vision aspects relating to the 3D reconstruction of scenes (e.g., cameral models, geometry in multi-view, structure from motion) and all theoretical aspects of the differents SfM steps.

Chapter 2

Classical Structure from Motion

Contents

This chapter gives a detailed overview of the incremental SfM pipeline which has shown, in the last years, to be an effective solution for the construction of extended surfaces in very different application fields. We start with a short introduction to the pinhole camera model and to the related calibration techniques since SfM methods exploit the camera parameters whose values were either calculated prior the surface reconstruction, or are to be determined during the point cloud reconstruction process. Then, the two-view SfM part aims to find the geometrical link between camera viewpoint pairs and to reconstruct first approximated 3D point positions into local 3D coordinate systems using the homologous point of two images. These results between image pairs are then exploited by the multi-view SfM step which consists of two main parts.

The first part lies in the determination of point tracks giving the correspondence between homologous points seen in numerous images. As described in Subsection 2.3.1, these point tracks are constructed with traditional feature matching methods.

The second part consists of the SfM algorithm itself which exploits the point track information. A small initial surface is constructed using the two-view SfM algorithm whose principle is described in Section 2.2. The incremental SfM method, whose principle is presented in Subsection 2.3.2, takes then the initial surface as a seed to perform an iterative growing of the 3D point cloud.

The complete SfM algorithm delivers both a sparse 3D point cloud and the camera poses along the sensor trajectory, all these results being obtained using only the 2D images acquired from different viewpoints.

2.1 Geometrical camera modeling

This section describes the projective (or perspective) model which makes the geometrical link between a point in the 3D scene and its projection onto the image plane. The pinhole camera model is given to explain all geometrical aspects of the image capturing process. The perspective projection is mathematically expressed by the projection matrix which includes the intrinsic and extrinsic parameters of the camera. Calibration algorithms for the estimation of the values of these camera parameters are presented in this section. The points and lines on a 2D plane and in the 3D space are expressed using the homogeneous coordinate convention. Three subsections present all aspects of the camera modeling topic: fundamentals on homogeneous points and lines, the pinhole camera model and camera calibration principles are introduced in Subsections 2.1.1, 2.1.2 and 2.1.3, respectively.

2.1.1 Homogeneous points and lines

Homogeneous points or vectors play an important role in the description of the projection model. The homogeneous convention represents N-dimensional coordinates with $N+1$ numbers. The homogeneous coordinate convention facilitates the calculus when computer graphics or 3D computer vision tasks involve perspective models.

2D lines. The equation of a line in a plane is given by $ax+by+c=0$ where the parameter triplet (a,b,c) completely defines a line. Thus, a line is defined by vector $(a,b,c)^T$ [HZ04, Sze11]. For any non-zero k, the vectors $(a,b,c)^T$ and $k\,(a,b,c)^T$ represent the same line.

Let us consider some basic results (proven in [HZ04]) which are helpful in the following sections

- Point $\mathbf{a}$ lies on line $\mathbf{l}$ if and only if $\mathbf{a}^T\mathbf{l}=0$
- The intersection of two lines $\mathbf{l}$ and $\mathbf{l}'$ is defined by point $\mathbf{a}=\mathbf{l}\times\mathbf{l}'$.
- The line through two points $\mathbf{a}$ and $\mathbf{a}'$ is given by $\mathbf{l}=\mathbf{a}\times\mathbf{a}'$.

Symbol $\times$ stands for the cross product (or vector product). If $\mathbf{a}=(a_1,a_2,a_3)^T$ and $\mathbf{b}=(b_1,b_2,b_3)^T$, then the cross product $\mathbf{a}\times\mathbf{b}$ leads to vector

$$(a_2b_3-a_3b_2,a_3b_1-a_1b_3,a_1b_2-a_2b_1)^T.$$

Cross products are also mathematically linked to skew-symmetric matrices through following equation:

$$\mathbf{a} \times \mathbf{b} = [\mathbf{a}]_\times \mathbf{b},$$

where $[\mathbf{a}]_\times = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}$. Note that skew-symmetric matrix $[\mathbf{a}]_\times$ is singular [HZ04].

2D points. Considering $\mathbb{R}^2$ as a vector space, a point in the plane is represented by column vector $(x, y)^T$. A point with coordinates $(x, y)^T$ lies on the line $(a, b, c)^T$ when

$$ax + by + c = 0.$$

Adding a third coordinate with value 1 to point $(x, y)^T$, leads to triplet $(x, y, 1)^T$ which satisfies following dot product condition:

$$(x, y, 1)\,(a, b, c)^T = ax + by + c = 0.$$

Thus, $(x, y, 1)^T$ in $\mathbb{R}^3$ represents the same point as $(x, y)^T$ in $\mathbb{R}^2$. For any non-zero value k, all points $(kx, ky, k)^T$ lying on a same 2D line are equivalent to $(x, y, 1)^T$ [HZ04]. Therefore, for any non-zero values of k, the set of vectors $(kx, ky, k)^T$ can be a representation for the point $(x, y)^T$ in $\mathbb{R}^2$. On the contrary, a homogeneous vector $\tilde{\mathbf{x}} = (\tilde{x}, \tilde{y}, \tilde{z})^T$ allows to retrieve an inhomogeneous point $(\tilde{x}/\tilde{z}, \tilde{y}/\tilde{z})^T$ by dividing the $\tilde{x}$ and $\tilde{y}$ components by last component $\tilde{z}$, for any $\tilde{z} \neq 0$. Therefore, point $\tilde{\mathbf{x}} = (\tilde{x}, \tilde{y}, \tilde{z})^T$ has the 2 degrees of freedom up to scale $\tilde{z}$. The 2D projective space is defined by

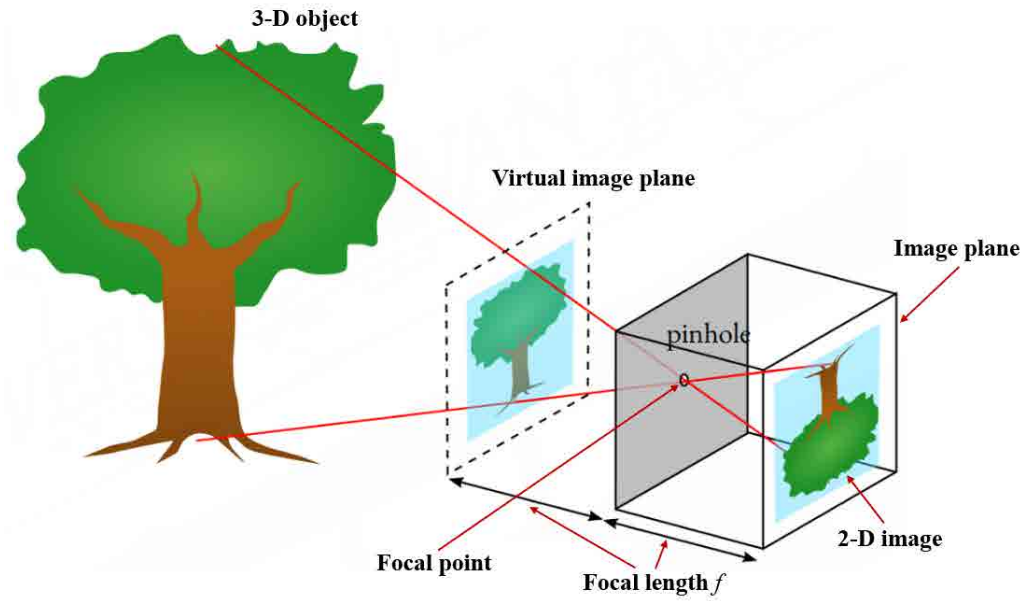
$$\mathbb{P}^2 \equiv \mathbb{R}^3 - (0, 0, 0)^T,$$

where $-(0, 0, 0)^T$ indicates that the vector $(0, 0, 0)^T$, which does not correspond to any line, is excluded [Sze11].

3D points. Similarly to 2D point coordinates, the coordinate of 3D points can be written using inhomogeneous coordinates $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$ or homogeneous coordinates $\tilde{\mathbf{x}} = w\,(x, y, z, 1) \in \mathbb{P}^3$, for any $w \neq 0$.

2.1.2 Pinhole model

This subsection describes the projection of 3D scene points onto a 2D image plane. The simplest way of modeling this process is to use the pinhole model which is a simple camera for which the lens is replaced by a single small aperture that is often designated by the terms "focal point" (see the left part of Fig. 2.1), or "optical centre" (see the left part of Fig. 2.2(a)). Obviously, the pinhole camera model does not account for lens distortion because such an ideal camera does not have a lens. However, according to their importance, distortions have to be taken into account for a correct reconstruction of the 3D data when using images acquired with real cameras. The camera parameters include intrinsic parameters, extrinsic parameters, and distortion coefficients (the latter are also often seen as intrinsic parameters).



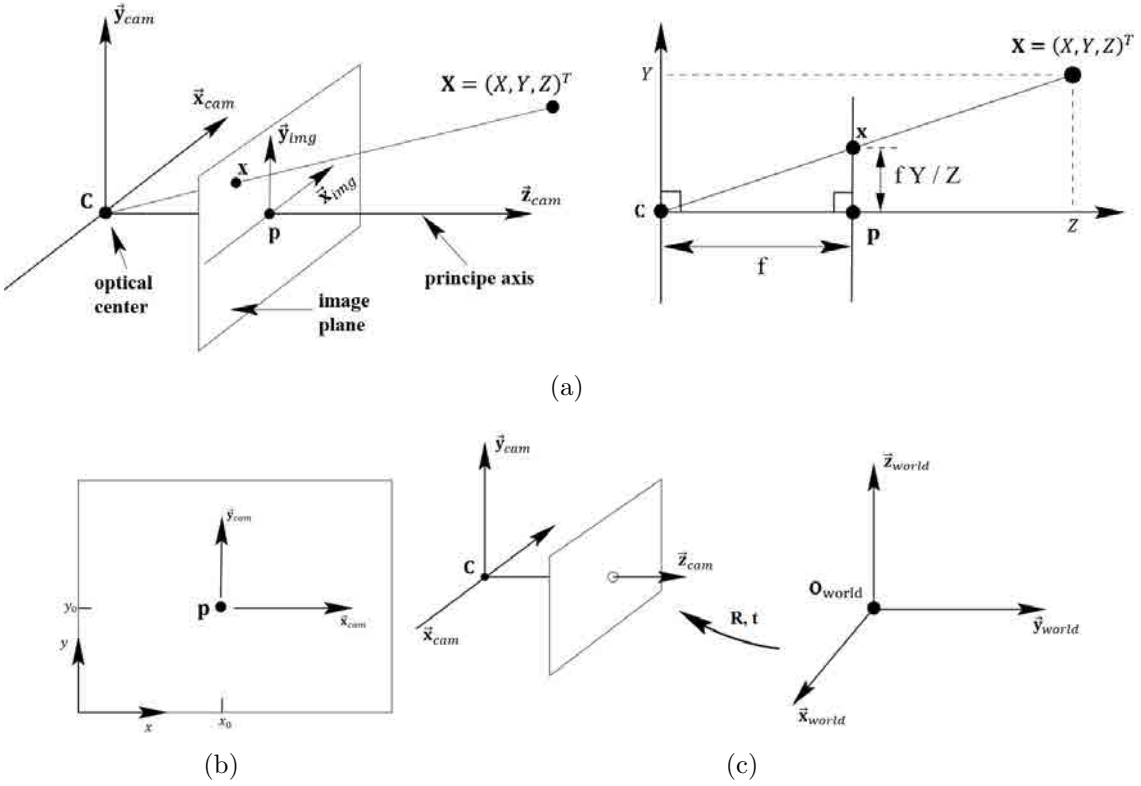
(a)

Figure 2.1: Distorsion free camera projection geometry. In the pinhole model, the true image plane corresponds to the plane including the CCD (charge coupled Device) sensor matrix of the camera. The virtual image plane is the non inverted image plane. The real and virtual images are placed on both sides of the focal point (the pinhole) at a distance corresponding to the focal length. The illustration is taken from [Mat].

2.1.2.1 Use of homogeneous coordinates in the camera model

The straight line defined by a 3D scene point and the camera optical centre (referred to as **C** in Fig. 2.2(a)) corresponds to a perspective projection trajectory in the 3D space. Since **C** is in a fixed position, the 2D coordinates of the point projection in the image plane depends only on the 3D coordinates of the scene point. It is noticeable that all 3D points lying on a same straight line project themselves on the same image point. As shown in the right part of Fig. 2.2(a), point $\mathbf{X} = (X, Y, Z)^T \in \mathbb{R}^3$ projects itself on point $\mathbf{x} = (x, y, f)^T$ on the image plane. Focal length f corresponds to the distance between principal point **p** (as illustrated by the right part in Fig. 2.2(a), **p** is the projection of the optical centre into the image plane) and the optical center **C**. The components x and y of the point **x** are defined by $x = f\frac{X}{Z} + p_x$ and $y = f\frac{Y}{Z} + p_y$ (see Fig. 2.2(b)), where (p_x, p_y) are the coordinates of the principal point **p**. Thus, point $\mathbf{X} = (X, Y, Z)^T$ is projected on the coordinates $\left(f\frac{X}{Z} + p_x, f\frac{Y}{Z} + p_y, f\right)^T$ on the image plane. The third homogeneous image coordinate can be ignored when passing from the Euclidean $\mathbb{R}^3$ space to the Euclidean $\mathbb{R}^2$ space:

$$(X, Y, Z)^T \mapsto \left(f\frac{X}{Z} + p_x, f\frac{Y}{Z} + p_y\right)^T.$$



(a)

(b) (c)

Figure 2.2: Geometrical representation of Fig. 2.1. (a) On the left side, the virtual image plane is considered as being the real image plane in this model. The camera centre **C** is here placed at the coordinate origin. Both left and right sides of this figure visualize the projection of a 3D point **X** onto the point **x** on the image plane. (b) Coordinate systems of the image plane (x, y) and the camera (x_{cam}, y_{cam}) and their relation with the principal point **p**. (c) The Euclidean transformation between the world coordinate system and the camera coordinate system. The illustrations are taken from [HZ04].

The perspective projection based homogeneous convention is defined by a linear relationship between 3D and 2D homogeneous coordinates:

$$\begin{pmatrix} fX + Zp_x \\ fY + Zp_y \\ Z \end{pmatrix} = \begin{bmatrix} f & 0 & p_x & 0 \\ 0 & f & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$$

which equales to

$$Z \begin{pmatrix} f\frac{X}{Z} + p_x \\ f\frac{Y}{Z} + p_y \\ 1 \end{pmatrix} = \begin{bmatrix} f & 0 & p_x & 0 \\ 0 & f & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}. \tag{2.1.1}$$

2.1.2.2 Complete perspective projection model

The homogeneous coordinate convention allows for a matrix formulation of the complete projection process which consists of two successive geometrical transformations

(from the world coordinate system to the camera coordinate system, and from the latter to the image coordinate system):

$$\lambda \mathbf{x} = \mathbf{P}\mathbf{X}, \tag{2.1.2}$$

where $\mathbf{P}$ is the camera matrix which projects the 3D scene points into the image plane, $\mathbf{X}$ is a 3D scene point defined by a homogeneous 4×1 vector $(X, Y, Z, 1)^T$, $\mathbf{x}$ is an image point represented by a homogeneous 3×1 vector $\mathbf{x} = (x, y, 1)^T$, and λ is an unknown scale factor in $\mathbb{R}_+^*$. The latter is known as the projective depth of the scene point $\mathbf{X}$ corresponding to the image point $\mathbf{x}$. The camera projection matrix is a 3×4 matrix with rank 3 and has the form $\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$, where $\mathbf{K}$ and $[\mathbf{R}|\mathbf{t}]$ correspond to the intrinsic matrix (a 3×3 upper triangular matrix) and the extrinsic matrix (a 3×4 matrix), respectively. The correspondence between points in the world and points in the image can be described by a simple model (see Eqs. (2.1.1) and (2.1.2)) in which the projection matrix can be decomposed as follows:

$$\mathbf{P} = \begin{pmatrix} \gamma f & s & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{pmatrix} [\mathbf{R}|\mathbf{t}], \tag{2.1.3}$$

with:

$$\mathbf{K} = \begin{pmatrix} \gamma f & s & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.1.4}$$

Matrix $\mathbf{K}$ is defined by the focal length f, the principal point $\mathbf{p} = (p_x, p_y) \in \mathbb{R}^2$, the skew coefficient s which is usually set to 0, and the aspect ratio parameter γ. For cameras equipped by a matrix of non-square CDD sensors the aspect ratio is used to model a different scale along the x- and y-image axes. In this case, the aspect ratio γ takes a value different from 1.

Matrix $\mathbf{R} = (r_{ij})_{1 \leq i,j \leq 3}$ and translation vector $\mathbf{t} = (t_x, t_y, t_z)^T$ are included in the extrinsic matrix which corresponds to a rigid transformation defining the position and orientation of the camera. This matrix gives the geometrical link between an arbitrary 3D world cooordinate system in which the positions of the scene points are known and the 3D camera coordinate system whose origin is the camera optical centre and whose z-axis corresponds to the the principal axis (see Fig. 2.2(c)). Let $\mathbf{X}^w = (X_w, Y_w, Z_w)^T$ be the 3D point in the world coordinate system having a position $\mathbf{X}^c = (X_c, Y_c, Z_c)^T$ in the coordinate system of the camera. This relationship between point positions may be written in homogeneous coordinates as follows:

$$\begin{pmatrix} \mathbf{X}^c \\ 1 \end{pmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mathbf{X}^w \\ 1 \end{pmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \mathbf{X}^w \\ 1 \end{pmatrix}. \tag{2.1.5}$$

The rotation matrix $\mathbf{R} = \mathbf{R}_x(\theta)\,\mathbf{R}_y(\phi)\,\mathbf{R}_z(\psi)$ can be decomposed into three matrices corresponding to three successive rotations around the axes $\overrightarrow{\mathbf{x}}_{world}$, $\overrightarrow{\mathbf{y}}_{world}$, and $\overrightarrow{\mathbf{z}}_{world}$ of the world coordinate system. These three matrices, from which the

parameters r_{ij} of Eq. (2.1.5) can be calculated, are defined as follows:

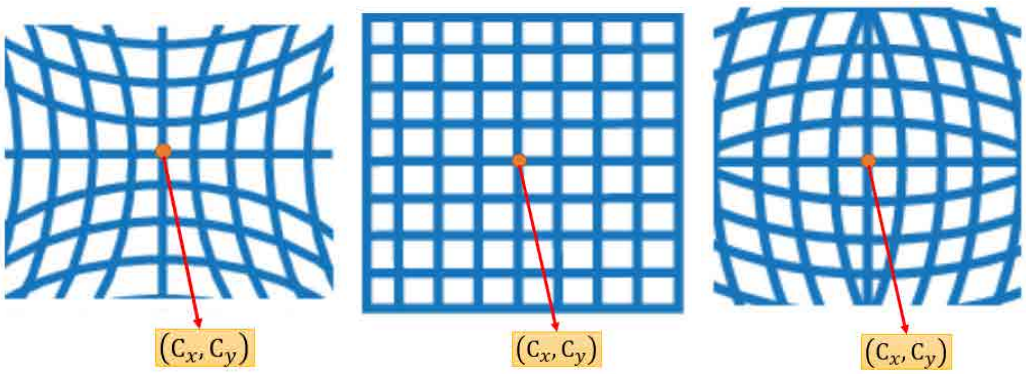
$$\mathbf{R}_x(\theta) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}, \quad \mathbf{R}_y(\phi) = \begin{pmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{pmatrix},$$
$$\mathbf{R}_z(\psi) = \begin{pmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The values of θ, ϕ, and ψ give the angles of rotation around the axes $\overrightarrow{\mathbf{x}}_{world}$, $\overrightarrow{\mathbf{y}}_{world}$, and $\overrightarrow{\mathbf{z}}_{world}$, respectively.

Matrix $\mathbf{R}$ belongs to $SO(3)$ which is the group of all rotations in the 3D space: $\mathbf{R} \in SO(3)$. Since rotation matrices are orthogonal, their inversion is equivalent to a transposition, e.g., $\mathbf{R}^{-1} = \mathbf{R}^T$. A more detailled description of matrix $\mathbf{P}$ can be found in [HZ04].

2.1.2.3 Image distortion model

A distortion is radial when the trajectories of light rays bend differently near the edges of a lens than close to the optical center. The effects of this optical phenomenon are illustrated in Fig. 2.3. Let (x, y) be the ideal distortion free pixel coordinates, and (x_{dist}, y_{dist}) the corresponding real (distorted and observed) image coordinates. The distortion can be mathematically formulated using the principal point coordinates $(\mathbf{C}_x, \mathbf{C}_y)$ and the distance r between $(\mathbf{C}_x, \mathbf{C}_y)$ and the coordinates



(a) Negative radial distortion (b) No distortion (c) Positive radial distortion

Figure 2.3: Radial image distortion. The distortion centre with coordinates $(\mathbf{C}_x, \mathbf{C}_y)$ superimposes the principal point (projection of the optical center onto the image plane). (a) Negative radial distortions (or pincushion) are observable for long focal length values. (b) Images without visible distortions. (c) Positive radial distortion (or barrel distortion) appear for short focal length values (i.e., $f < 12\ mm$). Images of endoscopes are usually affected by barrel distortions.

(x, y) of the point without distortions:

$$\begin{aligned} x_{dist} &= x + (x - \mathbf{C}_x)\left(1 + k_1 \times r^2 + k_2 \times r^4 + k_3 \times r^6 + \ldots + k_n \times r^{2n} + \ldots\right), \\ y_{dist} &= y + (y - \mathbf{C}_y)\left(1 + k_1 \times r^2 + k_2 \times r^4 + k_3 \times r^6 + \ldots + k_n \times r^{2n} + \ldots\right), \\ r &= \sqrt{(x - \mathbf{C}_x)^2 + (y - \mathbf{C}_y)^2}. \end{aligned} \tag{2.1.6}$$

where k_1, k_2, ..., k_n are the radial distortion coefficients of the lens. These distortion coefficients can be estimated in the camera calibration process. The radial distortion is referred to as pincushion distortion (see Fig. 2.3(a)) if the magnification increases when moving towards the image borders and is called barrel distortion (see Fig. 2.3(c)) if the magnification decreases close to the image borders.

For most conventional optics (as in endoscopy), the distortions are mainly radial. Tangential distortions are due to the fact that the lens and image (CCD matrix) planes are not perfectly aligned (i.e., parallel). Due to the resulting image tilting effect, objects seem to be closer or more far away from the camera according to their location in the image. In endoscopy, tangential distortions are negligible. For more details about this type of distortion, the reader can refer to [Zha00, Mat, HZ04].

The next subsection introduces popular camera calibration techniques which are often used to estimate the intrinsic and extrinsic camera parameters.

2.1.3 Camera calibration

Camera calibration also referred to as camera resectioning, is an important issue of the surface construction pipeline since different steps of a SfM method require the knowledge of the camera parameters. For instance, in the preprocessing step, distortion coefficients have to be known to correct the images, intrinsic parameters as the focal length are required for the 3D point position estimation using triangulation methods, intrinsic and extrinsic parameters have to be known to project the textures onto the surfaces. Some calibration approaches are dedicated to the determination of particular coefficients (e.g., the distortion parameters) or compute only the intrinsic parameters of a camera. Other approaches determine the whole intrinsic or extrinsic parameter set when one or several cameras are in a fixed position in a world coordinate system.

Numerous calibration methods were published in the last three decades. It is possible to mention approaches such as photogrammetric calibration [Tsa87, Fau93], self-calibration [MRG98, MC99], targetless camera calibration [PMSE12], calibration based on vanishing points for orthogonal directions [CT90] or calibration from pure rotation [Ste95]. Among them, pattern (e.g., checkerboards [Zha00]) based calibration methods and self-calibration algorithms [MRG98] are often used in SfM approaches. Especially, calibration algorithms with checkerboards led to accurate, robust and easy to implement solutions [Zha00, Bou, HZ04]. Numerous recent SfM pipelines [SF16, SSH+15] use the intrinsic camera parameter values available in the Exif tags[1] [EXI, NSR06] which are first used as approximated values and then re-

[1] EXIF stands for "Exchangeable Image File Format". Some cameras write automatically the intrinsic parameter values (as the pixel size, focal lenght, etc.) in the header of the image files.

fined by an optimization algorithm during the SfM step. This way to proceed do not require any calibration steps and lead to effective SfM methods based on numerous images acquired with different cameras (multiple or different camera types) [NSR06].

Camera calibration based on checkerboards or other patterns. The use of a calibration pattern is one of the most reliable ways to estimate the camera parameters. Commonly, camera calibration is done with checkerboard patterns printed on a paper sheet with a standard printer and attached to a flat surface as done in Fig. 2.4. The camera parameter estimation is based on the positions of 3D points and their corresponding 2D image point locations given in the world and image coordinate systems, respectively. These correspondences can be determined using images of a calibration pattern (such as the checkerboard in Fig. 2.4) acquired from different viewpoints [Zha00, MGV09, MBD+04]. The knowledge of these point correspondences is sufficient to estimate the camera parameters. While the intrinsic parameters are viewpoint independent, the extrinsic parameters are given with respect to the chosen world coordinate system usually defined by some checkerboard pattern points, see Fig. 2.4.

The accuracy of a camera calibration method with patterns can be assessed in two main steps [HZ04, Zha00]: (*i*) the 3D positions of calibration pattern points (i.e., the corners of the checkerboard squares in Fig. 2.4) are reconstructed for each viewpoint, and (*ii*) these 3D points are re-projected into the images and the mean Euclidean distance between these re-projected points and their corresponding segmented image points is the criterion which allows to assess the calibration accuracy.

Such calibration procedures are used in numerous applications. For instance, in industrial applications, such as the dimensional analysis of manufactured parts,

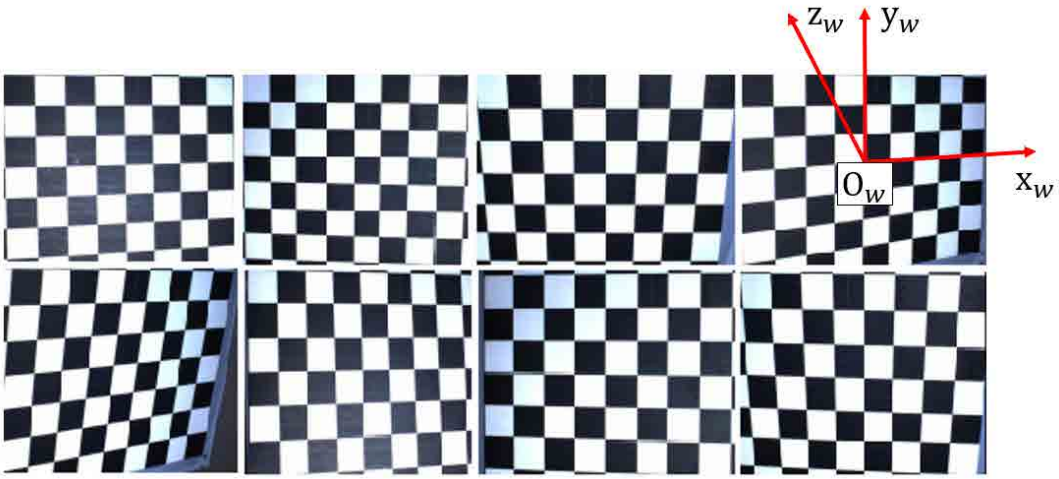


Figure 2.4: Calibration pattern acquired from various viewpoints. For the state-of-the-art method of Zhang, it is recommended to take numerous (at least 10-15) images from very different camera distances and orientations. Tests performed in the frame of this thesis have shown that a significant variation of the camera orientation is of particular importance to ensure an accurate calibration with Zhang's approach.

cameras can be calibrated (and re-calibrated) from controlled viewpoints and the calibration time is usually not a critical issue since the camera remains in a fixed position. It is noticeable that, in this situation, the extrinsic camera parameters have to be known precisely. In medical applications, like endoscopy, the extrinsic parameters are not of importance since the 3D points have not to be reconstructed in a known world coordinate system defined by a calibration plate (this would be difficult in a clinical situation). Moreover, in standard endoscopic examinations, the intrinsic camera parameters and the distortion coefficients are never calibrated.

Self-calibration (auto-calibration). Self-calibration (or auto-calibration) methods are employed in situations where cameras cannot be calibrated using calibration patterns. Self-calibration methods determine the intrinsic camera parameters directly with the content of the images acquired for the 3D scene reconstruction. Some early methods determine the values of constant intrinsic parameters using only the acquired image sequence, while some other methods were conceived for cameras for which the value of one or several intrinsic parameters is time (image) dependent.

In 1992, a general theory on multi-view camera self-calibration was published in a precursor work [FLM92]. Numerous self-calibration approaches [FLM92, MRG98, Tri97, HZ04] attempt to determine the camera parameter values by exploiting constraints relating to the intrinsic parameters themselves (e.g., skew-less camera or principal point located on the image centre, see s and (p_x, p_y) in Eq. (2.1.4)), to a priori knowledge about the camera motion (e.g., purely rotating cameras, planar motion, degenerate motions), or to the scene (e.g., planar scenes or scenes with depth relief). Changing focal length values (due to zooming to focus on scene details or to sharpen the image contrast), was one of the earliest reasons why the self-calibration task has been extended to varying intrinsic parameters [HÅ97].

Without any initial information about the camera parameters (the intrinsic parameters, as well as the extrinsic parameters relating to the camera positions) the relative positions of 3D points located on an object surface can be determined, but the absolute object position and orientation in the 3D space, as well as the true scale of the object, cannot be recovered.

In endoscopy, Barreto et *al.* [BRSF09] proposed an automatic calibration for endoscope cameras with lens distortions. This method only requires a single image of a planar chessboard pattern acquired for an uncontrolled (i.e., approximative) viewpoint. Although this method is a promising solution in computer-aided surgery (laparoscopy), it is difficult to adapt to other endoscopic examinations.

Exif tags. An alternative to pattern-based calibration methods and self-calibration approaches is to read the camera Exif (exchangeable image file format) tags [EXI] used to initialize some intrinsic parameters. The latter are refined using bundle adjustment approaches exploiting simultaneously all images (as seen in the coming section, such bundle adjustments are standardly employed in SfM [NSR06]).

Exif is a standard that specifies the format of images, sound, and ancillary tags used by digital cameras (including smartphones), scanners and other systems handling image and sound files recorded by digital cameras. The metadata tags defined

in the Exif standard includes recording date and time information, the settings of camera such as camera model reference, aperture, shutter speed, focal length, metering mode, and ISO speed information, etc. Exif tags provide focal length estimates in millimeters and these estimates can be converted to pixel units by applying the simple algorithm in [Sna08].

2.2 Two-view SfM principle

Algorithms which recover 3D information from two different viewpoints classically exploit the epipolar geometry which describes the mathematical link between 3D points and their two projections on image planes. This two-view reconstruction task is also an important step in multi-view SfM methods. For this reason, the 3D reconstruction with two images using two-view SfM has also be extensively studied [MRG98, HZ04]. This section shows how to estimate the poses of a calibrated camera using two images in order to reconstruct the 3D structure of a scene up to an unknown scale factor.

For cameras with known intrinsic parameters and delivering distortion free images, the two-view SfM algorithm consists of the following steps:

i. determination of homologous points using matching techniques,

ii. estimation of the fundamental matrix (and of the related essential matrix) which geometrically links the pixels of the two images,

iii. computation of the complete camera matrix P (including the extrinsic parameters) for each camera position, and

iv. use of triangulation methods to find 3D points.

The intrinsic camera parameters allow to compute the rays issuing from an image point and passing through the 3D point to be found. This line equation is given in the camera coordinate system taking the optical centre as origin. The determined extrinsic parameters give the geometrical transformation between two camera positions and thus allow to place all line equations into a common coordinate system. Usually, the pose of one camera defines the reference coordinate system and the relative pose of the second camera is given by the 3D rotation and translation between the two viewpoints.

This section begins with the description of the geometric relationship between two views (Subsection 2.2.1). Then the traditional feature matching methods to obtain point correspondences between those two views are presented in Subsection 2.2.2. Lastly, the results of two steps are used by a triangulation algorithm to estimate the 3D point cloud and camera poses (Subsection 2.2.3).

2.2.1 Two-view geometry

This section gives a brief overview of the two view geometry and the related geometric transformations such as the homography, the essential matrix, and the fun-

damental matrix that are required in the 3D point reconstruction process.

2.2.1.1 Homography

A 2D homography is a projective transformation in $\mathbb{P}^2$ to itself that preserves lines in a projective space. Let us consider two points $\mathbf{x}_i = (x_i, y_i, 1)^T$ and $\mathbf{x}_j = (x_j, y_j, 1)^T$ located in two images $\mathbf{I}_i$ and $\mathbf{I}_j$, and corresponding to the same 3D scene point (see Fig. 2.5). A homography between $\mathbf{I}_i$ and $\mathbf{I}_j$ is a non-singular 3×3 matrix, denoted by

$$\mathbf{H}_{i,j} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$$

and which defines a linear relationship between the coordinates of homologous points:

$$\begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_j \\ y_j \\ 1 \end{bmatrix}. \tag{2.2.1}$$

A homography gives the relationship between two images acquired by a same camera which was in two different positions (see the left part of Fig. 2.5) or provides the geometrical relationship between two images of a same scene acquired by two cameras from different viewpoints (see the right part of Fig. 2.5). In both situations, a homography gives the link between points lying in a plane. Mathematically, this geometrical relationship is written as follows:

$$\begin{aligned} -h_{11}x_j - h_{12}y_j - h_{13} + (h_{31}x_j + h_{32}y_j + h_{33})\, x_i = 0, \\ -h_{21}x_j - h_{22}y_j - h_{23} + (h_{31}x_j + h_{32}y_j + h_{33})\, y_i = 0. \end{aligned} \tag{2.2.2}$$

Matrix $\mathbf{H}_{i,j}$ contains 9 entries, but is defined only up to a scale. Thus, the number of degrees of freedom of $\mathbf{H}_{i,j}$ is 8. Since each point correspondence provides 2 equations

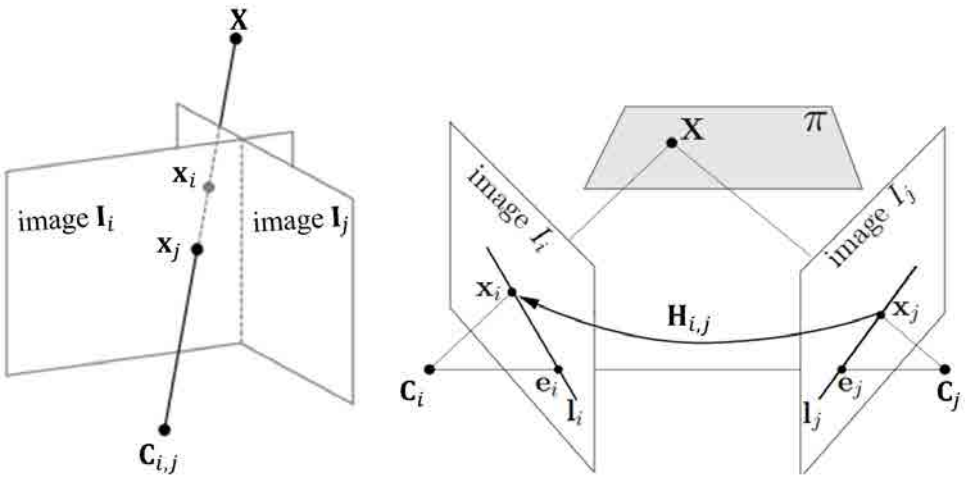


Figure 2.5: A homography describes the two-view geometry between two images of a purely rotating camera for instance (left) or by two cameras with an arbitrary viewpoint difference and capturing a planar scene (right). [Sch18, Ali16].

(see Eq. (2.2.2)), $\mathbf{H}_{i,j}$ can be estimated from at least 4 point correspondences between two views. The normalized DLT (Direct Linear Transform) algorithm in [HZ04] is an algorithm classicaly used to find the parameter values of a homography matrix.

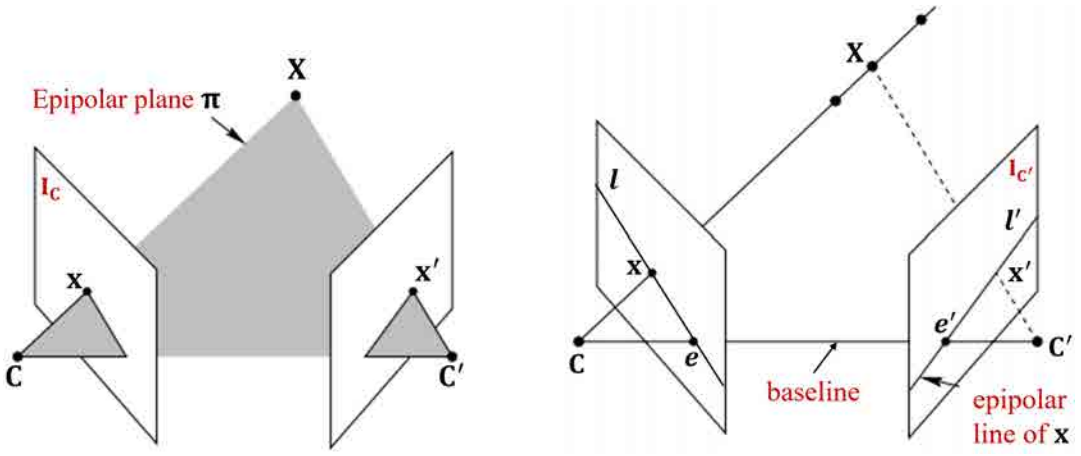
2.2.1.2 Epipolar geometry

In comparison to a homography, the epipolar geometry gives a more general link between the pixels of two images since the scene acquired by a moving camera has not to be planar. This geometry is independent of the scene structure and only relies on the intrinsic parameters and the relative camera poses [Sch18, HZ04]. Let us consider point $\mathbf{X}$ which is projected on two points $\mathbf{x}$ and $\mathbf{x}'$ located in images $\mathbf{I_C}$ and $\mathbf{I_{C'}}$, respectively (see Fig. 2.6). $\mathbf{C}$ and $\mathbf{C}'$ are two optical centre positions of the moving camera. The line segment that connects them is referred to as "baseline". The three 3D points $\mathbf{C}$, $\mathbf{C}'$ and $\mathbf{X}$ define the epipolar plane. The projection of the optical center $\mathbf{C}$ ($\mathbf{C}'$) in $\mathbf{I}_C$ ($\mathbf{I}_{C'}$) is denoted by $\mathbf{e}$ $\left(\mathbf{e}^{'}\right)$. Points $\mathbf{e}$ and $\mathbf{e}^{'}$ are the epipoles of images $\mathbf{I}_C$ and $\mathbf{I}_{C'}$ and are located on the baseline. They satisfy the following equations:

$$\mathbf{e} = \mathbf{P}\mathbf{C} \qquad \text{and} \qquad \mathbf{e}' = \mathbf{P}'\mathbf{C}', \tag{2.2.3}$$

where $\mathbf{P}$ and $\mathbf{P}'$ are the camera projection matrices of first and second view, respectively. An epipolar line $\mathbf{l}$ ($\mathbf{l}'$) is the intersection of the epipolar plane with the image plane $\mathbf{I}_C$ ($\mathbf{I}_{C'}$).

Fundamental matrix. The fundamental matrix $\mathbf{F}$ gives an algebraic representation of the epipolar geometry. $\mathbf{F}$ is a 3×3 matrix of rank 2. The concept of



(a) (b)

Figure 2.6: Epipolar geometry for an arbitrary point $\mathbf{X}$. (a) The optical centres camera centres $\mathbf{C}$ and $\mathbf{C}^{'}$, 3D point $\mathbf{X}$, and the projections $\mathbf{x}$ and $\mathbf{x}'$ of $\mathbf{X}$ in $\mathbf{I}_C$ and $\mathbf{I}_{C'}$ lie in a common plane $\boldsymbol{\pi}$. (b) The lines emanating from $\mathbf{x}$ and $\mathbf{x}'$ and passing through $\mathbf{C}$ and $\mathbf{C}'$ respectively intersect themselves at point $\mathbf{X}$. These lines are coplanar and are lying in $\boldsymbol{\pi}$ [HZ04]. Projection $\mathbf{x}'$ of $\mathbf{X}$ in $\mathbf{I}_{C'}$ must be located on epipolar line $\mathbf{l}'$ of projection $\mathbf{x}$ in $\mathbf{I}_C$.

fundamental matrix is detailed in [FLM92, Har92] for uncalibrated cameras. If 3D point $\mathbf{X}$ projects itself in $\mathbf{x}$ in the first view, and in $\mathbf{x}'$ in the second view, then these homologous image points satisfy the following equation:

$$\mathbf{x}'^T \mathbf{F} \mathbf{x} = 0. \tag{2.2.4}$$

Fig. 2.6 shows for an image pair that for each point $\mathbf{x}$ in one image, there exists a corresponding epipolar line $\mathbf{l}'$ in the other (second) image. If point $\mathbf{x}'$ of the second image is an homologous projection of $\mathbf{x}$, then $\mathbf{x}'$ must lie on the epipolar line $\mathbf{l}'$. Consider a set of 3D points $\mathbf{X}_i$ lying in the epipolar plane. The set of all points $\mathbf{x}_i$ in the first image and the set of the corresponding points $\mathbf{x}'_i$ in the second image are projectively equivalent, since they are each projectively equivalent to the planar point set $\mathbf{X}_i$. Thus, there exists a 2D homography $\mathbf{H}_{\boldsymbol{\pi}}$ mapping each $\mathbf{x}_i$ to its corresponding $\mathbf{x}'_i$ [HZ04].

Epipolar line $\mathbf{l}'$ passing through $\mathbf{x}'$ and epipole $\mathbf{e}'$ are mathematically linked by $\mathbf{l}' = \mathbf{e}' \times \mathbf{x}' = [\mathbf{e}']_\times \mathbf{x}'$, where $\times$ represents the cross product and $[\mathbf{e}']_\times$ is a skew-symmetric matrix. Matrix $[\mathbf{e}']_\times$ is defined as follows with the homogeneous coordinates $(a_1, a_2, a_3)^T$ of $\mathbf{e}'$:

$$[\mathbf{e}']_\times = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}.$$

Moreover,

$$\mathbf{x}' = \mathbf{H}_{\boldsymbol{\pi}} \mathbf{x}$$

leads to

$$\mathbf{l}' = [\mathbf{e}']_\times \mathbf{x}' = [\mathbf{e}']_\times \mathbf{H}_{\boldsymbol{\pi}} \mathbf{x}. \tag{2.2.5}$$

Since point $\mathbf{x}'$ belongs to line $\mathbf{l}'$ one can write the following scalar product:

$$\mathbf{x}'^T \mathbf{l}' = 0.$$

By identifying (or comparing) previous equation with Eq. (2.2.4), one can see that $\mathbf{l}' = \mathbf{F}\mathbf{x}$. Thus, the fundamental matrix $\mathbf{F}$ is given by:

$$\mathbf{F} = [\mathbf{e}']_\times \mathbf{H}_{\boldsymbol{\pi}}.$$

The most used properties of the fundamental matrix are the following:

- *Link between homologous points.* If $\mathbf{x}$ and $\mathbf{x}'$ are corresponding image points, then $\mathbf{x}'^T \mathbf{F} \mathbf{x} = 0$.
- *Link between a 2D point and its epipolar line.* The epipolar lines of $\mathbf{x}$ and $\mathbf{x}'$ are $\mathbf{l}' = \mathbf{F}\mathbf{x}$ and $\mathbf{l} = \mathbf{F}^T \mathbf{x}'$, respectively.
- *Epipoles.* Since epipole point $\mathbf{e}'$ lies on epipolar line $\mathbf{l}' = \mathbf{F}\mathbf{x}$, one have $\mathbf{e}'^T (\mathbf{F}\mathbf{x}) = 0$, which is equivalent to $\left(\mathbf{e}'^T \mathbf{F}\right) \mathbf{x} = 0$. Thus $\mathbf{e}'^T \mathbf{F} = 0 \ \forall \mathbf{x} \in \mathbf{l}$. Similarly $\mathbf{F}\mathbf{e} = 0$.

Futher geometrical properties involving the fundamental matrix can be found in [HZ04].

A solution can be determined for the 3×3 matrix $\mathbf{F}$ of rank 2 by writing Eq. (2.2.4) for at least 7 corresponding point pairs $(\mathbf{x}, \mathbf{x}')$. Each point pair defined by $\mathbf{x}_i = (x_i, y_i, 1)^T$ and $\mathbf{x}'_i = (x'_i, y'_i, 1)^T$, leads to an equation in which the point coordinates are linearly linked by the unknown entries of $\mathbf{F}$:

$$\begin{bmatrix} x'_i & y'_i & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = 0. \tag{2.2.6}$$

With at least 7 points, Eq. (2.2.6) leads to following system of equations:

$$\begin{bmatrix} x'_1x_1 & x'_1y_1 & x'_1 & y'_1x_1 & y'_1y_1 & y'_1 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_nx_n & x'_ny_n & x'_n & y'_nx_n & y'_ny_n & y'_n & x_n & y_n & 1 \end{bmatrix} \begin{bmatrix} f_{11} \\ f_{12} \\ f_{13} \\ f_{21} \\ f_{22} \\ f_{23} \\ f_{31} \\ f_{23} \\ f_{33} \end{bmatrix} = 0. \tag{2.2.7}$$

Matrix $\mathbf{F}$ can only be determined up to a scale with this set of homogeneous equations. A solution can be determined using linear methods like the normalized eight-point algorithm [HZ04].

Essential Matrix. The essential matrix was first introduced by Longuet-Higgins [LH81]. This matrix is the adaptation of the fundamental matrix to the case of normalized image coordinates and calibrated cameras. Consider a camera matrix defined by $\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$ and let $\mathbf{x} = \mathbf{PX}$ be a point in the image, where $\mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^{3\times1}$ are the rotation matrix and translation vector between two viewpoints, respectively. Let us denote by $\hat{\mathbf{x}}$ the image point with normalized coordinates. This normalized coordinates are determined with:

$$\hat{\mathbf{x}} = \mathbf{K}^{-1}\mathbf{x} = [\mathbf{R}|\mathbf{t}], \tag{2.2.8}$$

where $\mathbf{x}$ is the image coordinates before the normalization and $\mathbf{K}$ is the calibration matrix with known intrinsic parameters. In such conditions, the essential matrix $\mathbf{E}$ is used to represent the geometrical constraints between homologous points $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}'$ of two images:

$$\hat{\mathbf{x}}'^T\mathbf{E}\hat{\mathbf{x}} = 0. \tag{2.2.9}$$

With equations (2.2.4), (2.2.8) and (2.2.9), it follows that the relationship between the fundamental and essential matrices is:

$$\mathbf{E} = \mathbf{K}'^T\mathbf{F}\mathbf{K}. \tag{2.2.10}$$

In addition, as proven in [HZ04, MRG98], an essential matrix can be decomposed in a product of matrices:

$$\mathbf{E} = [\mathbf{t}]_\times \mathbf{R},$$

where

$$[\mathbf{t}]_\times = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}$$

is a skew symmetric matrix and $\mathbf{R}$ is the orthonormal matrix corresponding to the rotation matrix included in projection matrix $\mathbf{P}$. Essential matrices are always of rank two (consequently they include one zero singular value) and have two equal non-zero singular values. An essential matrix is defined by the values of six parameters, namely the three translations components of $\mathbf{t}$ and the three angles (roll, pitch, and yaw) defining the entries of $\mathbf{R}$. However, since the essential matrix is defined up to a scale, it has only five degrees of freedom [HZ04].

According to [GL96, HZ04], skew-symmetric matrix $[\mathbf{t}]_\times$ may be written as

$$[\mathbf{t}]_\times = \beta \mathbf{U}\mathbf{B}\mathbf{U}^T,$$

where $\mathbf{U}$ is an orthogonal matrix, β is a scale factor and $\mathbf{B}$ is following matrix:

$$\mathbf{B} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Since $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$ and the scale is arbitrary, $\mathbf{E}$ can be written as:

$$\mathbf{E} = \mathbf{U}\,\mathrm{diag}\,(1,1,0)\left(\mathbf{B}\mathbf{U}^T\mathbf{R}\right) = \mathbf{U}\,\mathrm{diag}\,(1,1,0)\,\mathbf{V}^T, \tag{2.2.11}$$

where $\mathbf{V}^T = \mathbf{B}\mathbf{U}^T\mathbf{R}$. It is assumed that the two non-null singular values of $\mathbf{E}$ equal both 1. This assumption can be made for two reasons. On the one hand, the singular values can be equal since their decomposition (SVD) is not unique. On the other hand, their values can be set to 1 since the scale is arbitrary. Matrix $\mathbf{E}$ can directly be computed from Eq. (2.2.9) using normalized image coordinates, or can be determined with the fundamental matrix when the calibration matrix $\mathbf{K}$ is known (see Eq. (2.2.10)).

If the SVD of $\mathbf{E}$ is $\mathbf{U}\,\mathrm{diag}\,(1,1,0)\,\mathbf{V}^T$ and when ignoring the signs, there are two possible factorizations of $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$ (see [HZ04]):

$$\begin{cases} [\mathbf{t}]_\times &= \mathbf{U}\mathbf{B}\mathbf{U}^T \\ \mathbf{R} &= \mathbf{U}\mathbf{B}^T\mathbf{V}^T \end{cases} \quad \text{or} \quad \begin{cases} [\mathbf{t}]_\times &= \mathbf{U}\mathbf{B}\mathbf{U}^T \\ \mathbf{R} &= \mathbf{U}\mathbf{B}\mathbf{V}^T \end{cases} \tag{2.2.12}$$

2.2.2 Feature detection and matching methods

In previous subsections, it was supposed that the homologous points $(\mathbf{x}, \mathbf{x}')$ of two images are known to estimate a 2D geometric transformation. The parameters of homographies, fundamental matrices, or essential matrices can only be accurately assessed with robust matching methods which minimize wrong point matches. This

subsection presents a popular and robust approach for establishing the point correspondence between two images. Such a correspondence is obtained by detecting feature points localizing particular image textures and by computing feature point descriptions (descriptors) to match the points. The feature descriptor based matching method classically uses the geometrical model (e.g., the homography or the fundamental matrix) to verify the validity of point matches.

Feature detection and matching is an important task in many computer vision applications, such as SfM, image retrieval, or object detection. These algorithms consists in three main steps:

i. The aim of the detection step is to identify 2D interest points.

ii. In the feature characterization step, feature descriptors are determined with pixels belonging to a small region centered on the interest points. This descriptor is usually a vector with binary or real-valued components.

iii. The third step consists of a preliminary feature matching aiming to find as much as possible corresponding texture or object points seen in two images. It is a preliminary step in the sense that the false matches are discarded in second time.

2.2.2.1 Feature detection

Features are specific structures in the image [GHT11, Dec].

- *Corners/interest points.* The terms "corner" and "interest point" are interchangeably used and refer to as point-like features which have a local two dimensional structure. A rather salient corner is defined by the intersection of two edges with different orientations, while a rather "less pronounced" corner is defined by a point with two little different edge directions in its neighborhood. An interest point is a pixel which has a well-defined position and can be robustly detected. Some well-known detectors of such features are the Harris-Stephens Corner detector [HS88], the corner detector that uses the minimum eigenvalue algorithm [ST94], FAST (Features from Accelerated Segment Test, [RD06]) detector, BRISK (Binary Robust Independent Elementary Features, [SCS11]) detector, ORB (Oriented FAST and Rotated BRIEF, [RRKB11]) detector, and BRIEF (Binary Robust Independent Elementary Features, [CLSF10]) detector.

- *Edges.* An edge corresponds to a boundary between two colour homogeneous image regions. In general, an edge contains at least two interest points linked by a junction. Edge detectors are usually conceived by exploiting constraints relating to grey-level transitions. For instance, the shape or the smoothness of transitions are constraints which impose the unicity of the response of gradient-based edge detectors like that of Canny [Can86] or of the SUSAN detector [SB97].

- *Blobs/regions of interest points.* Blobs either refer to image regions with significant textures or correspond to objects. The position of blobs are usually associated to a certain points (e.g., the center of mass of the blob, the local maximum of an operator response, or the blob's center of gravity). Different blob detection algorithms were published: SURF (Speeded-Up Robust Features, [BETV08]), SIFT (Scale-Invariant Feature Transform, [Low04]), or KAZE [ABD12], for instance.

- *Ridges.* Ridges are used to describe elongated objects, and they occur normally along the center of those objects [Lin98]. However, ridges are not commonly used since they are by far more difficult to extract than corners or blobs.

- *Affine-invariant region detectors.* These region detectors, proposed by [MS04, Low04], are invariant to local affine changes in small image regions. They are able to identify (and locate) similar regions in images taken from different viewpoints, these regions being geometrically linked by transformations including scale changes, rotations and shearing.

In the litterature, corners and blobs are often referred to by the same name and are both simply seen as interest points.

Numerous local feature information are usually used to describe each point of interest and to decide whether it should be selected as a keypoint or not (points of interest are candidates and keypoints are effectively used for the matching step). A feature detector is considered as efficient when it is able to locate keypoints under changing illumination conditions and at different scales. A feature detector should also be able to determine the dominant orientation of the keypoint (the dominant orientation helps to avoid mismatching during the homologous search). An efficient detector has to exhibit repeatability and reliability [RWdS+19, Kie]. Repeatability means that the same keypoint can be detected in different images. Reliability means that the keypoint detected should be discriminant enough so that the number of its matching candidates is small.

The type of the detector used to find keypoints depends on the aim of the application and/or on the characteristics of the scene. For example, for images of bacteria cells, it is more appropriate to use a blob detector than a corner detector. On the contrary, in images of aerial views of a city, a corner detector can be used to find man-made structures [Fea]. An evaluation of the efficiency of different keypoint detectors can be found in [GMBR10, SMB00].

Since more than a decade, SIFT [Low04] has arguably been the most popular keypoint detection and matching method. The SIFT algorithm proposed by Lowe consists of a keypoint detection which is independent towards image rotations, scale changes, affine transformations, intensity variations, and viewpoint changes. The SIFT algorithm consists of four basic steps.

- In the first step, grey-level extrema are searched in a scale space given in the form of a Difference of Gaussian (DoG) image pyramid.

- Then, in the second step, an accurate position is determined for the interest

points and the keypoint are obtained by discarding the interest points with low contrast.

- In the next step, an orientation is assigned to the keypoints.
- Finally, in the last step, a descriptor vector is computed using the local image gradient magnitudes and orientations taken in a region centered on the keypoint coordinates at the different scales of the pyramid.

SIFT detectors were integrated in many 3D reconstruction pipelines and contributed to the robustness of the algorithms [SF16, NSR06, GFF10, CT15, SSH+15]. A result obtained by the SIFT detector for endoscopic images is illustrated in Fig. 2.7(b).

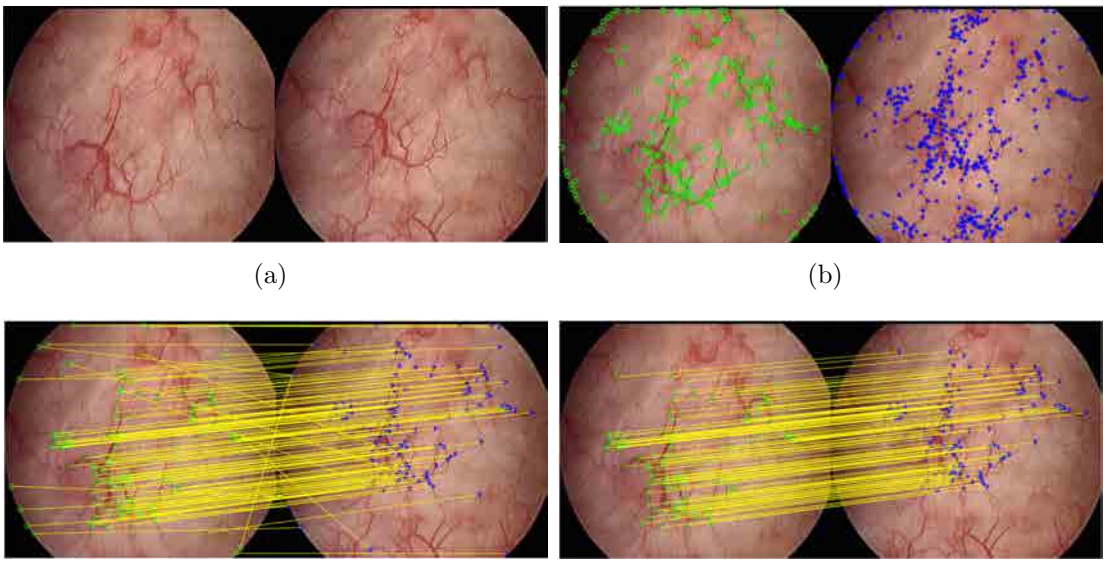
Recently, the ability of neural networks to learn keypoint representations from image data led to significant progress in the field of computer vision, notably in the tasks of object detection and recognition [KSH12, RASC14]. Neural networks have also been applied to the problem of descriptor learning [BRPM16, YTLF16, STF+15] to derive more discriminative representations for local features. A comparative evaluation of hand-crafted features such as SIFT [Low04], SURF [BETV08], DSP-SIFT [DS15] and learned keypoint features can be found in [SHSP17]. The reported results show that the learned keypoint features can be a promising solution for the feature detection step in SfM algorithm.

2.2.2.2 Feature matching

A descriptor is a vector whose components correspond to numerous characteristics computed with the pixel values of a image region (a patch) centered on the feature point. Some descriptors, such as SIFT or SURF, rely on local pixel values exploited at different scales. Binary descriptors, such as BRISK or ORB, rely on the local intensity differences, whose signs are encoded into a binary vector. The descriptors associated with the keypoints of two images are used to match the homologous points irrespective of image rotations, scale changes, and illumination variations. The component values of the descriptor vectors should be discriminant enough to minimize the number of homologous keypoint candidates during the matching step.

Descriptors can be categorized in two classes:

- Local descriptors are determined for a patch whose content differs from that of the surrounding image regions in terms of textures, colors and/or intensities. Local descriptors are designed to have component vector values representing a small neighborhood around a keypoint. For this reason, they are especially suitable for the encoding of patch information in the context of point matching. SIFT detectors are popular local descriptors able to encode local information for different scenes.
- The vector components of a global descriptor encodes the content of the whole image. Such descriptors exhibit generally a limited robustness since a change of a part of the image content may have a too significant impact on the descriptor components [Sch18, Tya].



(a) (b)

(c) (d)

Figure 2.7: Illustration of the whole procedure of keypoint detection and matching based on the SIFT and RANSAC algorithms. (a) Two bladder images extracted from a cystoscopic video-sequence. (b) Keypoints (or feature points) detected by the SIFT algorithm on image $\mathbf{I}_i$ (green points) and image $\mathbf{I}_j$ (blue points). Descriptors are available for all detected feature points. It is visible that the detected points are mainly on texture points corresponding to blood vessels and on the spherical field of view of the endoscope. (c) Preliminary feature point matching. The displacement of the endoscope between $\mathbf{I}_i$ and $\mathbf{I}_j$ mainly consists of a translation. The yellow lines which represent the point correspondence should ideally all be almost parallel. (d) The RANSAC outlier rejection algorithm discarded all wrong matches. The model used was a homography (the small area of the epithelial surface seen through the reduced field of view of the endoscope allows to consider that the surface is planar).

Once both the feature points and the feature descriptors were determined for two images, the preliminary feature matching can be performed to find candidates for the corresponding feature point pairs. The similarity of two feature descriptors is measured through the Euclidean distance computed with the components of the two vectors ($\mathbb{L}_2$ norm of the vector difference, [Low04, BETV08]). Two points are considered as homologous point candidates when their similarity value is smaller than a given threshold. Let $\mathbf{G}_i(\mathbf{I}_i)$ denote the set of the feature points detected in image $\mathbf{I}_i$. For every pair of images $\mathbf{I}_i$ and $\mathbf{I}_j$, one have to consider each feature point $g \in \mathbf{G}_i(\mathbf{I}_i)$ and to find its nearest neighbor (i.e., the closest vector) $g_{nn} \in \mathbf{G}_j(\mathbf{I}_j)$ so that

$$g_{nn} = \underset{g' \in \mathbf{G}_j(\mathbf{I}_j)}{\arg\min} \|\mathbf{g}_d - \mathbf{g}'_d\|_2,$$

where $\mathbf{g}_d$ and $\mathbf{g}'_d$ are the descriptor vectors of feature points $g \in \mathbf{G}_i(\mathbf{I}_i)$ and $g' \in \mathbf{G}_j(\mathbf{I}_j)$, respectively. A commonly used algorithm to search for the nearest neighbor can be found in [AMN+98]. This algorithm uses a kd-tree data structure to efficiently compute the nearest neighbors. This way to determine the candidate pair (g, g_{nn})

for feature descriptor g is repeated for all $g \in \mathbf{G}_i$ to obtain a preliminary set of matched points for images $\mathbf{I}_i$ and $\mathbf{I}_j$.

Due to numerous reasons (blur, lack of contrast in some image regions, reflections, occlusions, etc.) the preliminary matching procedure is imperfect and false correspondences are often established. Fig. 2.7(c) shows wrong matching examples after the determination of the preliminary point correspondences. As illustrated in Chapter 1, wrong matches affect significantly the 3D point reconstruction quality of a SfM method, mainly due to the strong impact of the matching errors on the correctness of the essential or fundamental matrix. It is a standard procedure to use the so-called RANSAC [FB81] or MSAC [TZ00] algorithms to remove the wrong or inaccurate matches (outliers) and to keep the accurate matches (inliers).

RANSAC is an iterative method that simultaneously estimates the parameters of a model (e.g., a homography) from a set of observed data (e.g. feature points) and finds the data which is in accordance with the model. When data are in accordance with the model they are referred to as "inliers" and the outliers are discarded so that they do not influence the parameter values of the model. The RANSAC algorithm can also be seen as an outlier rejection method.

In the frame of a SfM algorithm, the model used in the RANSAC geometry consistency test to improve the set of matched feature points is a homography, an essential matrix, or a fundamental matrix [ZDFL95, Sna08, HZ04]. Algorithm 1, which will be detailed in the next sections, shows how the RANSAC method can be used to select inliers (true homologous points) using a homography which acts as a geometrical model for the outlier rejection. The principle of the RANSAC algorithm remains the same when a fundamental matrix (or an essential matrix which is a particular case of the fundamental matrix) has to be determined. The only difference lies in the minimal number of homologous point pairs which is required to compute the model parameters (4 pairs for a homography and 8 pairs for a fundamental matrix).

The RANSAC algorithm uses the set of the matched points delivered by the preliminary matching step. In step 1 of Algorithm 1, it is assumed that the homography $\mathbf{H}$ which was computed with 4 randomly selected point pairs corresponds to a correct geometrical link between the two images. Homography $\mathbf{H}$ is then used to exploit one of the epipolar geometry constraints ($\mathbf{x}_i - \mathbf{H}\mathbf{x}_i^{'}$ is ideally null, where $\mathbf{x}_i$ and $\mathbf{x}_i^{'}$ are supposed to be homologous). Distance d_i defined by Eq. (2.2.13) in Step 1 of Algorithm 1 should be small for inliers. Threshold ρ is used to decide whether a pair $\left(\mathbf{x}_i, \mathbf{x}_i^{'}\right)$ is an inlier or an outlier. At each iteration of step 1, this threshold is used to count the number of point pairs which are accurately linked by current homography $\mathbf{H}$. Step 1 is repeated until the number of N_s iterations was reached. Among the N_s homographies computed in this way, the algorithm selects the homography $\mathbf{H}$ which led to the greatest number of inliers.

The aim of Step 2 is to refine the value of homography $\mathbf{H}$ provided by Step 1. A minimization algorithm uses all pairs $\left(\mathbf{x}_i, \mathbf{x}_i^{'}\right)$ of images $\mathbf{I}_i$ and $\mathbf{I}_j$ to determine the refined (most accurate) homography $\hat{\mathbf{H}}$ and the final set of accurate inliers points pairs $\left(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i^{'}\right)$. The cost function used in Eq. (2.2.14) not only leads to an accurate homography $\hat{\mathbf{H}}$, but allows also to select the subset of pairs $\left(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i^{'}\right)$ allowing for the

Algorithm 1 Homography matrix determination using the RANSAC algorithm

Input: The set of matched point pairs $\left(\mathbf{x}_i, \mathbf{x}_i^{'}\right)$ of images $\mathbf{I}_i$ and $\mathbf{I}_j$ found in the preliminary matching step, iteration number N_s and Euclidean distance threshold ρ which can be adjusted according the methods in [FB81, HZ04].

/* ***Step 1***: *RANSAC-based homography matrix* $\mathbf{H}$: */*
for $i = 1$ to N_s **do**
(a) Select a random sample of 4 point pairs.
(b) Estimate homography $\mathbf{H}$.
(c) Compute Euclide distance d_i for all point pairs $\left(\mathbf{x}_i, \mathbf{x}_i^{'}\right)$ using

$$d_i = \left\|\mathbf{x}_i - \mathbf{H}\mathbf{x}_i^{'}\right\|_2 + \left\|\mathbf{x}_i^{'} - \mathbf{H}\mathbf{x}_i\right\|_2 . \tag{2.2.13}$$

(d) Count the number of inliers which fulfill $d_i < \rho$.
end for
Choose the homography matrix $\mathbf{H}$ with the largest number of inliers.

/* ***Step 2***: *Optimal estimation* $\mathbf{H}$: */*
Re-estimate $\mathbf{H}$ from all the inliers in Step 1 by minimizing the error function $\mathcal{D}$

$$\begin{aligned} \min_{\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i^{'}} \quad & \mathcal{D}\left(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i^{'}\right) = \sum_i \left(d\left(\mathbf{x}_i, \hat{\mathbf{x}}_i\right)^2 + d\left(\mathbf{x}_i^{'}, \hat{\mathbf{x}}_i^{'}\right)^2 \right), \\ \text{subject to} \quad & \hat{\mathbf{x}}_i^{'} = \hat{\mathbf{H}}\hat{\mathbf{x}}_i, \ \forall i. \end{aligned} \tag{2.2.14}$$

Output: Matrix $\hat{\mathbf{H}}$.

determination of a precise geometrical link.

The Maximum Likelihood estimation method [HZ04] is used to minimize cost function $\mathcal{D}$ defined in Eq. (2.2.14). An illustration of point matches achieved using the RANSAC algorithm taking a homography as model is given in Fig. 2.7(d). As justified caption of this figure, the wrong initial matches (notably those corresponding to interest points located on the spherical border of the field of view limits) were eliminated and only correct keypoint pairs remain.

Other solutions than the joint use of the SIFT and RANSAC algorithms were also proposed to find homologous point pairs. Thus, Agarwal et *al.* [ASS+09] employed an image retrieval method [NS06] to match the points of same or similar objects seen either in different scenes, or in a same scene with changing illumination conditions or viewpoints. In their approach, images are selected and treated by the approximate nearest neighbor feature matching method to find homologous points. Schönberger et *al.* [SBF15] determine point matches with a new approach for quickly predicting whether two images contain a common scene part. In the frame of a SfM method, Havlena et *al.* [HS14] proposed an efficient way to match points between all image pairs without the need to exhaustively test each individual image pair. They directly use the assignments of individual feature points to visual words in a vocabulary tree

[NS06, PCI+07] as verified point correspondences.

2.2.3 Camera poses and 3D point cloud computation

Let us consider two images $\mathbf{I}_1$ and $\mathbf{I}_2$ acquired with two calibrated cameras whose known intrinsic matrices are $\mathbf{K}_1$ and $\mathbf{K}_2$, respectively. $\mathbf{K}_1$ and $\mathbf{K}_2$ can be computed with the camera calibration methods described in Subsection 2.1.3. The determination of point matches, as well as the computation of essential and fundamental matrices are detailed in previous Subsections 2.2.2 and 2.2.1. Thus, this chapter part deals with the issues (iii) and (iv) of the SfM algorithm for two-view as introduced at the beginning of this section. Issue (iii) relates to the computation of the projection matrix $\mathbf{P}$ of each camera, whereas issue (iv) concerns the simultaneous determination of the camera poses and the 3D point cloud using the homologous points seen in $\mathbf{I}_1$ and $\mathbf{I}_2$. This section details the determination of the projective camera matrices under the assumption that the intrinsic parameter matrix and the essential matrix are known. Then, the 3D point cloud can be obtained using a triangulation method, while the camera poses can be estimated using the relative orientations and positions between the two viewpoints.

The camera of one of the two viewpoints is usually taken as reference, i.e., to define a reference coordinate system. For instance, the optical center $\mathbf{C}_1$ of the first camera can be taken as the origin of the reference coordinate system ($\mathbf{C}_1 = (0, 0, 0)$) and the position of the optical center of the second camera is defined by a translation $\mathbf{t}$. In the same way, the relative orientation of the two cameras is defined by a rotation $\mathbf{R}$ [Sze11, HZ04].

2.2.3.1 Computation of the projective camera matrices

The point matches obtained with a feature matching method are used to estimate the essential matrix $\mathbf{E}$ defined by Eq. (2.2.9). Rotation matrix $\mathbf{R}$ and the components (t_x, t_y, t_z) of translation vector $\mathbf{t}$ between two viewpoints are then computed with Eq. (2.2.12). The common technique often used to compute essential matrix $\mathbf{E}$ and the relative camera pose using at least 5 corresponding points is detailed in [Nis04]. It is assumed that, optical center $\mathbf{C}_1$ of camera 1 overlaps the origin of the reference coordinate system, and that the optical axis of this camera is also the z-axis of the reference coordinate system. Thus, for camera 1, rotation matrix $\mathbf{R}$ is the identity and it's translation vector is null. Therefore, for the two view configuration, the camera matrices are $\mathbf{P}_1 = \mathbf{K}_1 [I|0]$ and $\mathbf{P}_2 = \mathbf{K}_2 [\mathbf{R}|\mathbf{t}]$, respectively, where I is a 3×3 identity matrix.

However, for projection matrix $\mathbf{P}_2$ of camera 2 there are four possible solutions due to the arbitrary signs for translation and rotation (see Fig. 2.8). It is recalled that the signs are unknown since the camera positions are not known in an absolute world coordinate system. The solution to be chosen is that which ensures that the reconstructed 3D points are in front of both cameras (the other three solutions place the points behind at least of one of the cameras [HZ04]).

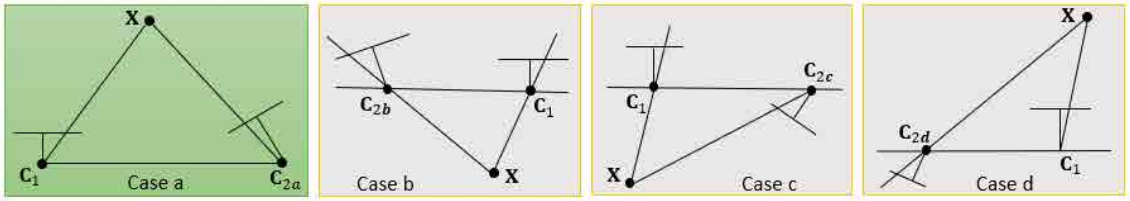


Figure 2.8: Illustration of the four possible reconstruction solutions with a known matrix $\mathbf{E}$ and a calibrated intrinsic matrix $\mathbf{K}$. A test with a single point $\mathbf{X}$ to determine if it is in front of both cameras is sufficient to decide between the four different solutions for the camera matrix $\mathbf{P}_2$.

2.2.3.2 3D point triangulation

Once the projection matrices of the cameras were computed, the 3D points can be estimated from measured point projections (i.e., the feature point positions) seen in the two images. The 3D point positions can be recovered by computing the rays issuing from the measured points, each ray having an equation in the 3D space which is derived from the camera projection matrices and the relative camera poses. The intersection of two rays emanating from homologous points gives the 3D point localisation in the reference coordinate system (see Fig. 2.9).

Due to various effects that affect the image quality (digitization noise, unperfect distortion correction, etc.), the true positions of homologous points $\mathbf{x}_1$ and $\mathbf{x}_2$ are never exactly measured. These errors affect also the accuracy of the back-projected rays which do not intersect themselves into the 3D point to be reconstructed. Let $\mathbf{x}_1, \hat{\mathbf{x}}_1, \mathbf{x}_2, \hat{\mathbf{x}}_2$ be the measured ($\mathbf{x}_i$) and the predicted ($\hat{\mathbf{x}}_i = \mathbf{P}_i\mathbf{X}$) homologous image positions of a 3D point $\mathbf{X}$ in two views $\mathbf{I}_1$ and $\mathbf{I}_2$, respectively. The aim of the triangulation algorithm is to estimate a 3D point $\hat{\mathbf{X}}$ whose position is ideally the same as the true position (that of $\mathbf{X}$). To approach this goal, one have to minimize the reprojection errors of point $\hat{\mathbf{X}}$. In other words, the predicted (reprojected) points $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ should ideally exhibit an exact geometrical link through the fundamental matrix[2] between images $\mathbf{I}_1$ and $\mathbf{I}_2$. To reach this goal, one seeks the points $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ that minimize function $\mathcal{C}(\mathbf{x}_1, \mathbf{x}_2)$:

$$\begin{aligned} \min_{\mathbf{x}_1,\mathbf{x}_2} \quad & \mathcal{C}(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{x}_1, \hat{\mathbf{x}}_1)^2 + d(\mathbf{x}_2, \hat{\mathbf{x}}_2)^2, \\ \text{suject to} \quad & \hat{\mathbf{x}}_1 \mathbf{F} \hat{\mathbf{x}}_2 = 0, \end{aligned} \tag{2.2.15}$$

where $d(.,.)$ is the Euclidean distance between points [HS97, HZ04]. Supposing that the measurement noise follows a Gaussian error distribution, accurate image point correspondences $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ can be found with the Maximum Likelihood Estimates (MLE) method [HZ04, HS97]. Once $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ are known, the corresponding rays should intersect themselves in the 3D space:

$$\begin{cases} \hat{\mathbf{x}}_1 & = \mathbf{P}_1 \hat{\mathbf{X}}, \\ \hat{\mathbf{x}}_2 & = \mathbf{P}_2 \hat{\mathbf{X}}. \end{cases} \tag{2.2.16}$$

[2] In this section, the fundamental matrix is used instead of the essential matrix to make the established formulas more general and applicable to both calibrated and uncalibrated cameras.

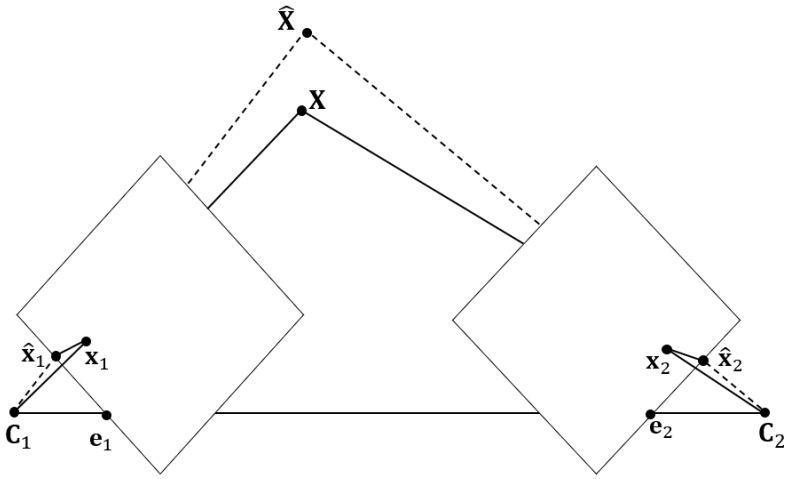


Figure 2.9: Illustration of the effect of noise on the triangulation accuracy. The term triangulation refers to the determination of a point in the 3D space given its projections onto two or more images. In order to solve this problem it is necessary to know the parameters of the projective matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ in the two-view case. While $\mathbf{x}_1$ and $\mathbf{x}_2$ are the images measurement given by a feature matching method; $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ are the predicted image points. Point $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ correspond to an accurate 3D point $\hat{\mathbf{X}}$ only when they are are located on their respective epipolar lines (these 2D points have to satisfy the epipolar constraint defined by the fundamental matrix).

With $\hat{\mathbf{x}}_1 = (x_1, y_1, 1)$ and $\hat{\mathbf{x}}_2 = (x_2, y_2, 1)$, Eq. (2.2.16) can be rewritten as follows:

$$\begin{cases} x_1\left(\mathbf{p}_1^{3T}\hat{\mathbf{X}}\right) - \left(\mathbf{p}_1^{1T}\hat{\mathbf{X}}\right) & = 0 \\ y_1\left(\mathbf{p}_1^{3T}\hat{\mathbf{X}}\right) - \left(\mathbf{p}_1^{2T}\hat{\mathbf{X}}\right) & = 0 \\ x_1\left(\mathbf{p}_1^{2T}\hat{\mathbf{X}}\right) - y_1\left(\mathbf{p}_1^{1T}\hat{\mathbf{X}}\right) & = 0, \end{cases} \text{ and } \begin{cases} x_2\left(\mathbf{p}_2^{3T}\hat{\mathbf{X}}\right) - \left(\mathbf{p}_2^{1T}\hat{\mathbf{X}}\right) & = 0 \\ y_2\left(\mathbf{p}_2^{3T}\hat{\mathbf{X}}\right) - \left(\mathbf{p}_2^{2T}\hat{\mathbf{X}}\right) & = 0 \\ x_2\left(\mathbf{p}_2^{2T}\hat{\mathbf{X}}\right) - y_2\left(\mathbf{p}_2^{1T}\hat{\mathbf{X}}\right) & = 0, \end{cases} \quad (2.2.17)$$

where $\mathbf{p}_1^{iT}$ and $\mathbf{p}_2^{iT}$ are the i^{th} vector row of matrices $\mathbf{P}_1$ and $\mathbf{P}_2$, and $1 \leq i \leq 3$. With matrix

$$\mathbf{A} = \begin{bmatrix} x_1\mathbf{p}_1^{3T} - \mathbf{p}_1^{1T} \\ y_1\mathbf{p}_1^{3T} - \mathbf{p}_1^{2T} \\ x_2\mathbf{p}_2^{3T} - \mathbf{p}_2^{1T} \\ y_2\mathbf{p}_2^{3T} - \mathbf{p}_2^{2T} \end{bmatrix},$$

one obtains $\mathbf{A}\hat{\mathbf{X}} = 0$. Thus, the 3D position of point $\hat{\mathbf{X}}$ can be estimated with a linear triangulation method which is analogous to the direct linear transformation (DLT) method [HS97, HZ04].

2.3 Multi-view SfM principle

A multi-view SfM algorithm aims to simultaneously estimate a sparse 3D point cloud, the successive camera poses along the acquisition path (the extrinsic parameters provide the camera position and orientation in a common coordinate system), the intrinsic camera parameters (when these parameters were not calibrated) and

the distortion coefficients (when these values are not available for short or long focal lengths). The principles presented in previous section for cameras with calibrated intrinsics parameters and for two viewpoints can be extended to uncalibrated cameras and to multi-view SfM. In the multi-view SfM algorithms, point correspondences across image sequences (videos or numerous images taken individually from different viewpoints) are most often given in the form of point tracks. Usually, point tracks are obtained from pairwise point correspondences which are determined with feature matching method. The common images in different pairs allow to determine feature tracks. The point tracks are determined in the first part of the multi-view SfM method and act as input for the second part, namely the incremental reconstruction part.

Fig. 2.10 gives an overview on the iterative incremental reconstruction process which is based on five steps. The first step, which is performed only once, selects two images used for an initial 3D point reconstruction by a two-view SfM algorithm (see Subsection 2.2.1). This initialization step provides a first 3D point cloud, the camera projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ of two viewpoints and the poses (a pose is defined by a 3D position and 3D orientation) of the cameras in a world coordinate system. Then the 3D point cloud grows iteratively by performing a loop over four steps. At iteration n, the "matching" step selects an additional viewpoint $n+1$ and determines the homologous points between image $\mathbf{I}_{n+1}$ and the already selected images. The resulting point correspondences are used to determine projection matrix $\mathbf{P}_{n+1}$ and the camera pose for viewpoint $n+1$. Then, the triangulation step uses a multi-view method to determine additional 3D points having projections seen both in image $\mathbf{I}_{n+1}$ and in images $\mathbf{I}_i$ with $i \in [1, n]$. After this point cloud growing, depending on the incremental SfM implementation, local or global bundle adjustment methods refine the 3D point positions, some or all projection matrices $\mathbf{P}_i$, and some or all camera poses. In the final step (outlier filtering) of the iteration loop, 3D outlier points are rejected to smooth the surface represented by the point cloud. This loop is iterated until all viewpoints have been exploited.

Obviously, the SfM algorithms achieve optimal reconstruction results in terms of robustness and accuracy when the images are of high quality (e.g., well contrasted), acquired from different viewpoints under similar illumination conditions, include rich textures/structures, and are significantly overlapped.

In this section, the set of n images and their triangulated scene points are denoted by $\mathcal{I} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_n\}$ and $\mathcal{X} = \{\mathbf{X}_k \in \mathbb{R}^3 | \; k = 1, \ldots, m\}$, respectively. m stands both for the number of 3D points and for the number of point tracks. This section presents first the point track determination methods and focusses then on the incremental reconstruction pipeline.

2.3.1 Determination of point tracks

A point track consists of a set of 2D matched points extracted from the images of a video sequence or from a set of images acquired "frame by frame" from different viewpoints. Each set of points matched across multiple images should correspond to a single 3D scene point. It means that each individual point in a track should be

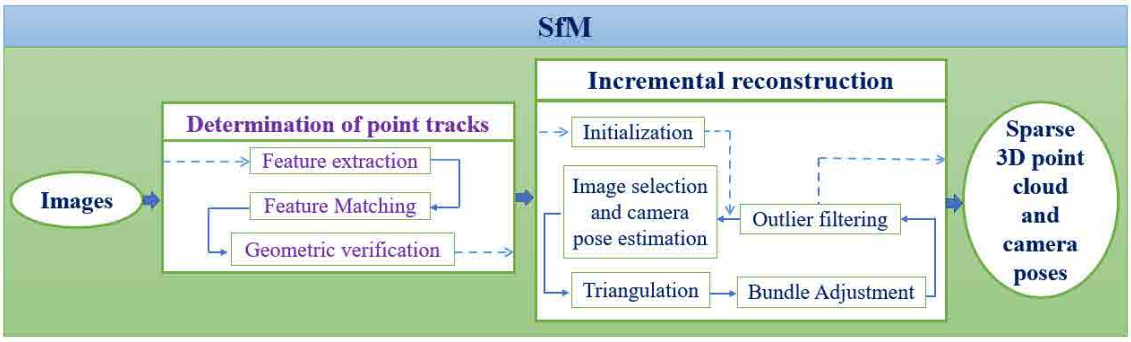


Figure 2.10: Overview of a classical SfM pipeline with two parts. This illustration is taken from the state-of-the-art paper published by Schönberger and Frahm [SF16].

the projection of the same 3D point. Tracks may contain points from images widely distributed throughout the data set and, in a video-sequence or for a frame-by-frame acquisition, it can link temporally consecutive images, as well as non temporally consecutive images. In SfM, the accuracy of the point positions and the length of point tracks impact significantly the surface reconstruction performances. Indeed, small 2D point localization errors lead to important 3D point reconstruction errors, and the precision and robustness of the 3D reconstruction are also maximized when the viewpoint differences are significant.

Point tracks are often determined in three main steps: (*i*) detection of feature points in each image, (*ii*) matching of feature descriptors between pairs of image, and (*iii*) exploitation of the pairwise matches to form point tracks across multiple images [Sna08]. Steps (*i*) and (*ii*) use the feature detection and matching methods which are described in Subsection 2.2.2. The principle of step (*iii*) is based on the following observation: an initial track consists of two corresponding points and a new 2D point of another keypoint pair is added to the track when its homologous points is already present in the track. Point tracks lead to SfM methods with high performance when textured scenes are acquired from different viewpoints under relative constant illumination conditions. However, in scenes with almost no textures or for weakly contrasted image textures acquired under changing illumination conditions, the performances of the keypoint detection and matching methods strongly decreases, and only few and short (or none) point tracks can be built.

The Kanade-Lucas-Tomasi (KLT) tracker is often used in video-sequences to establish point correspondences between consecutive frames. The algorithm detects a set of keypoints in a frame and then, each of them are tracked in the next frames. Each keypoint of an image that does not have a corresponding point in the previous image corresponds potentially to a new track. One of the central issues of the KLT method is to find the appropriate feature type which optimizes the tracking results. In SfM, one of the most commonly used feature type for the keypoint tracking is corners which are detected by the minimum eigenvalue algorithm [ST94]. The KLT tracker is widely used for small baseline (distance between camera optical centers) matching or for video-sequences acquired with a small camera motion [TK91, ST94]. However, even in well contrasted scenes, finding long tracks using the popular KLT methods is still challenging due to occlusions, illumination changes, noise, and large

motion of object, which may interrupt the chain of tracked points.

2.3.2 Incremental reconstruction pipeline

This subsection describes the 5 steps of an incremental SfM pipeline.

2.3.2.1 Initialization step

The incremental SfM pipeline initializes the reconstruction with a carefully selected image pair to ensure a robust and accurate 3D point cloud determination (an inappropriate choice for the initialization can even make fail the surface reconstruction). Appropriate image pair candidates are those having a large number of point matches. The image pair candidate should also be acquired from two camera positions with a large baseline value (i.e., for a significant viewpoint change) so that the initial two-view reconstruction is robust and accurate [BS06].

The work in [Sna08] confirms that the selected initial image pair must provide as much as possible matches, but also highlights the fact that these matches cannot be well modeled by a homography for numerous scene types. It is especially the case when images visualize scene parts with a non-planar surface lying in a large depth interval. For a non-planar surface with a constant depth interval, the appropriateness of a homography increases when the acquisition distance becomes larger. However, in the general case (for all types of scenes), a homography is most often inappropriate.

A video-sequence or an image set acquired "frame by frame" often consists of a mix of panoramic images (including large parts of the scene) and of local images (corresponding to small scene parts and including numerous fine details). The work

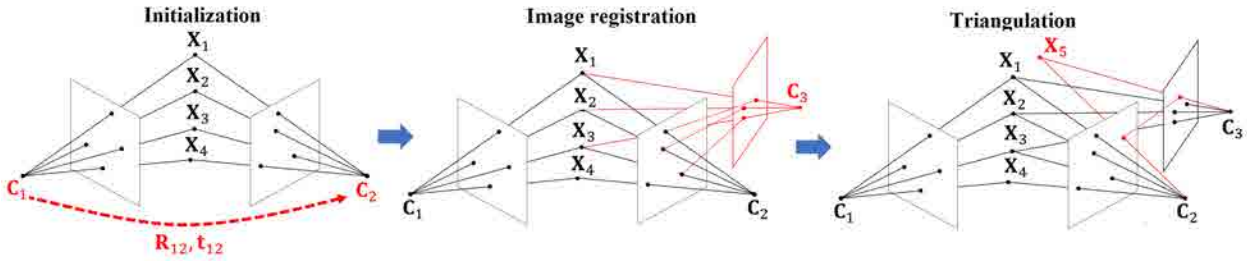


Figure 2.11: Illustration of the first three steps of an incremental SfM pipeline. (a) Initialization step. The SfM algorithm starts by selecting the images of two viewpoints and reconstructs an initial 3D point cloud using a two-view geometry and triangulation method. The camera pose and displacement is also known between the two viewpoints (b) At each iteration of the incremental SfM pipeline, a new viewpoint is selected and the keypoints of the new image are linked to homologous keypoints of the already selected images. The camera pose and projection matrix are also determined for this viewpoint. (c) Multi-view triangulation. The 3D positions of the keypoints of the new image are reconstructed using a multi-view triangulation technique which implies all points of a keypoint track.

in [SF16], has shown that an initialization with local images leads to a more efficient (accurate and detailed) surface reconstruction than choosing panoramic views as initial images. This is mainly due to the fact that with local images with more fine details, the number of starting matches, as well as the number and length of the points tracts are often the more important.

The output of this initialization step consists of a first set of 3D points (i.e. the initial point cloud), the two projections matrices of the cameras, and the camera displacement, all this information being expressed in a world coordinate system (typically that of the first camera pose). The results of this step are obtained with the two-view reconstruction approach described in Section 2.2.

2.3.2.2 Image selection and camera pose estimation

In the incremental SfM reconstruction process, the data of individual viewpoints are iteratively added to the current 3D point cloud to extend the surface (after iteration n, the point cloud was computed using $n + 2$ images, $n = 0$ before the first iteration) corresponding to the first point cloud obtained with the two images selected in the initialization step). At each iteration, the SfM algorithm aims to find the most appropriate image (or viewpoint) for extending the surface. The camera viewpoint to be used to extend the surface at iteration n is given by the image including the greatest number of keypoints belonging to the point tracks used to reconstruct the current 3D point cloud. For this viewpoint, the camera pose (camera position and orientation in the world coordinate system) can be estimated by solving the Perspective-n-Point (PnP, [LNF09]) problem using the correspondence information between the keypoints of the selected image and their homologous 2D projections of already estimated 3D points (such a correspondence makes the link between a keypoint of the selected image and an already known 3D point). The PnP algorithm is a camera calibration method that can be used to determine the position and orientation of a camera, given known intrinsic parameters and a set of correspondences between 3D points and their 2D projections. For uncalibrated cameras, the PnP algorithm also estimates the intrinsic parameters.

The camera pose is characterized by six degrees-of-freedom (DoF) corresponding to three rotation parameters (roll, pitch, and yaw) and 3D translation parameters which are all defined in the world coordinate system. Since the 2D-3D correspondences are affected by errors (some of the correspondences can be false or inaccurate), the pose of calibrated cameras is usually estimated using the RANSAC algorithm (taking the projective matrix as a model) and a minimal pose solver [LNF09]. An accurate camera pose estimation is essential as the point triangulation (which adds new points at the current surface, see coming Subsection 2.3.2.3) may fail when the camera parameters are inaccurate. Therefore, the choice of the appropriate image to increase the size of the point cloud is crucial. Recently, Schönberger et *al.* [SF16] proposed a novel and robust "next best image selection method" for an accurate pose estimation by using a large number of 2D-3D correspondences with a uniform distribution of points. This uniform distribution of points (in the 2D images) makes the triangulation step more accurate and reliable.

The output of this image selection step consists of the input data ($n+1$ camera poses, n camera displacements defined by a 3D rotation and a 3D translation, $n+1$ projective camera matrices and the 3D point cloud after $n-1$ iterations), the camera projection matrix of viewpoint $n+2$ and the camera pose of viewpoint $n+2$ in the world coordinate system which were determined at current iteration n.

2.3.2.3 Triangulation step

The triangulation method described in Subsection 2.2.3.2 for two images can be extended to more viewpoints. On the one hand, multiview triangulation increases both the robustness and accuracy of the 3D point localization due to the data-redundancy. On the other hand, it allows to add 3D points to an existing 3D surface. It exists numerous multi-view triangulation methods [HS97, SF16, AAT12, ASS08].

The keypoints of the newly "registered" image (i.e., the image selected in the step described in previous Subsection 2.3.2.2), must fulfill two conditions to be usable in this triangulation step. Firstly, such a keypoint must be a homologous point of the projection of one of the 3D points belonging to point set $\mathcal{X}$ of the already reconstructed cloud. Secondly, this keypoint must also have a homologous point which is still not a projection of one of the 3D points in set $\mathcal{X}$. All the keypoints which satisfy these two conditions are used to reconstruct the new 3D points which extend the size of the point cloud (i.e., set $\mathcal{X}$).

The output of this step consists of $n+2$ camera poses, $n+1$ camera displacements (3D rotations + 3D translations), $n+2$ projective camera matrices, and the 3D point cloud which was extended at current iteration n.

2.3.2.4 Bundle adjustment step

The aim of the bundle adjustment is to simultaneously refine the 3D coordinates of the scene points, the extrinsic camera parameters and, when not calibrated before the SfM step, the intrinsic parameters and the radial distortion coefficients. As sketched in Fig. 2.12, this step aims to adjust the bundle of rays passing through the camera optical centres $\mathbf{C}_i$ and the set of reconstructed 3D points $\mathcal{X}$ [TMHF99].

The camera parameters and the 3D point positions are refined by an optimization method which minimizes a criterion based on the Euclidean distances between the segmented (measured) features points acting as ground truth and the reprojections of their corresponding 3D points. Let $\mathbf{x}_{ij}$ be the point position measured in image i and belonging to track j. As illustrated in Fig. 2.12, the predicted points correspond to the 3D points $\mathbf{X}_j$ which are projected by camera matrix $\mathbf{P}_i$ onto image plane $\mathbf{I}_i$. For n views and m tracks, objective function g can be written as:

$$g\left(\mathbf{P},\mathbf{X}\right)=\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}\left\|\mathbf{x}_{ij}-\mathbf{P}_i\mathbf{X}_j\right\|^2, \tag{2.3.1}$$

where w_{ij} is an indicator variable: $w_{ij}=1$ when a point of track j was detected in image i and $w_{ij}=0$ otherwise. It is noticeable that numerous $\mathbf{x}_{ij}$ are inexistent since often not all images contain points of all tracks. Euclidean norm $\|\mathbf{x}_{ij}-\mathbf{P}_i\mathbf{X}_j\|$

quantifies the reprojection error of 3D point $\mathbf{X}_j$ in image i. Such a reprojection error relates both to the imperfect values of the 3D point coordinates and of the parameters of the projection matrix $\mathbf{P}_j$.

The bundle adjustment involves a non-linear least square problem since, for points $\mathbf{X}_j$, camera matrices $\mathbf{P}_j$ lead to an overdetermined system of non-linear equations. Thus, the minimization of objective function g can be achieved using classical non-linear least-squares algorithms such as the Gauss-Newton, Levenberg-Marquardt, or preconditioned conjugate gradient (PCG) methods. Among these methods, the Levenberg-Marquardt algorithm [Mor78] is often used in SfM algorithms since it is easy to implement and it uses an effective damping strategy that ensures a quick convergence towards the solution from a wide range of initial guesses. The Levenberg-Marquardt algorithm provides a solution based on a system of linear equations by iteratively linearizing objective function g in the neighborhood of the current estimate.

However, non-linear least squares algorithms, such as Levenberg-Marquardt, only find local minima. Consequently, there is no guarantee of convergence to the optimal solution from an arbitrary starting point. This observation is also true for the large and high dimensional parameter spaces induced by bundle adjustment methods in SfM algorithms. For such parameter spaces, even the Levenberg-Marquardt method is prone to get trapped in wrong local minima [Sna08, SF16]. Therefore it is important to provide initial estimates of the parameters which are close to the solution. Much research in the field of 3D reconstruction deals with easily computable non-optimal solutions that can be used as a starting point for bundle adjustment

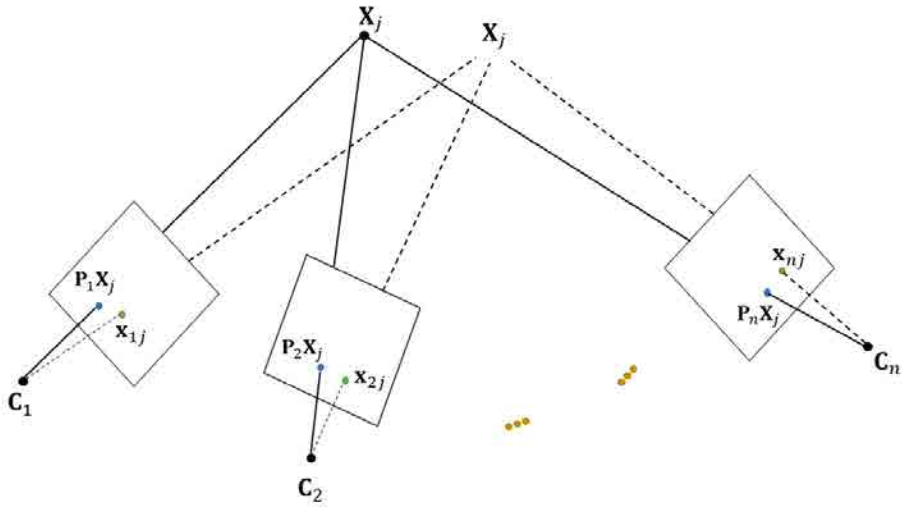


Figure 2.12: Illustration of the reprojection errors. These errors are defined with respect to the keypoints provided by the SIFT detector (point $\mathbf{x}_{ij}$, where $i = 1, 2, \ldots, n$ is the viewpoint number and j stands for the 3D point number). The error corresponds to the Euclidean distance between the back-projected point $\mathbf{P}_i\mathbf{X}_j$ and its corresponding keypoint $\mathbf{x}_{ij}$. Inaccurately detected keypoints can be at the origin of non-null Euclidean distances since they impact the accuracy of the 3D point reconstruction.

methods [Sna08, Wu13].

It is necessary to frequently perform a bundle adjustment after the image registration and triangulation steps in order to avoid the accumulation of large errors which cannot be corrected. In [SF16, Wu13], a local bundle adjustment is performed on the data of the subset of images with the most interconnected (i.e., homologous) points. This local adjustment is done after each viewpoint registration and triangulation (i.e., at each iteration n). In [SF16, Wu13], global bundle adjustments (on the data of all viewpoints) are also performed when the surface has grown by a given area expressed as a percentage. The computation time relating to the bundle adjustment steps is the main factor making SfM time-consuming.

2.3.2.5 Outlier filtering step

At various steps of the complete multi-view SFM algorithm different geometric verifications (e.g., RANSAC, MSAC) are usually performed to minimize all types of outliers (false 2D point correspondences, points which are not in accordance with a homography or fundamental matrix, false 2D/3D correspondences through a camera projection matrix, etc.). Such outlier filtering is required at almost all steps of a SfM algorithm since the efficiency of the current step depends on the accuracy of the previous steps. Although for accurate matching, triangulation, and bundle adjustment steps, the growing cloud of 3D points must be filtered at each iteration to avoid large errors that accumulate during the iterative process. 3D point outliers may occur due to either other outliers that were "forgotten" by the filtering process of different SfM steps [SF16, Wu13, NSR06], or to noise [WKZ+16, RBG+19].

In the last years, numerous noise filtering methods have been proposed for point clouds. For example, Wolff et *al.* [WKZ+16] detected and removed noisy points and outliers by reconstructing a same surface part several times using different sets of viewpoints (the different surfaces should be the same and exploiting them jointly facilitates the outlier rejection). These authors do not only evaluate the geometric consistency of surfaces but also exploit the photometric consistency between the images of two viewpoints. 3D point filtering not only improves the data for the next SfM loop iteration, but the final 3D point cloud also leads to a more smooth surface after the surface meshing.

The four last steps are iterated until all viewpoints were exploited.

2.4 Conclusion

This chapter gives an overview of the different steps of an incremental SfM algorithm. The iterative point cloud construction performed by incremental SfM approaches leads to extended surfaces which can be reconstructed in a robust and accurate way in the frame of numerous applications involving different scene types [SF16, NSR06, Wu13].

However, these results are usually obtained for scenes and acquisition conditions which fulfill at least several of the following conditions.

1. The images contain contrasted textures and/or structures in every regions of interest.

2. The available viewpoints are well distributed: objects are seen from very different angles and distances. Thus, objects or object parts are seen under different orientations and at different scales.

3. The scene contains surface parts or objects with a "rich" 3D structure information (i.e., surface shape variations which favor the efficiency of the SfM method) and the images visualize these surface parts.

4. The illumination conditions can vary between the images but remain rather moderate.

Most of these assumptions are verified for numerous applications since in natural or industrial scenes for instance the images contain rich information in terms of colors, textures, and image primitives, while the acquisition conditions can usually be controlled, or are at least partially controlled. In endoscopy the acquisition conditions and scenes are very different:

1. In gastroscopy, the stomach wall is almost without textures and structures, and in cystoscopy textures are not systematically available in all regions of interest of the images.

2. The available viewpoints exhibit a small variability: all images are acquired with a small distance between the camera and the epithelial surface. The endoscopist also tries to maintain the endoscope axis perpendicularly to the surface so that the view-angles are limited.

3. The short acquisition distance and the smooth shape of the hollow organs (stomach and bladder) explain why the image contains surface parts with few 3D structure/shape variations.

4. In endoscopy, the light is "directed": the light emanating from the endoscope's tip forms a cone whose main axis orientation depends on the endoscope position in the organ. For this reason, the illumination conditions change significantly with the viewpoint and the internal surface wall orientation.

In endoscopy, not only the acquisition and scene conditions are complex, but additional artifacts (usually less present in other applications) affect the scene: many pixels visualize specular reflections and floating objects (bladder) or blood (stomach and bladder) can occlude scene parts.

Chapter 3

Dense Optical Flow based Structure from Motion

Contents

As detailed in Chapter 1, SfM-algorithms classically consist of a 2D homologous point determination stage followed by a 3D scene structure estimation stage. While the simultaneous 3D point cloud and camera pose determination is a well-mastered stage, determining large and numerous homologous point groups remains an open problem for textureless scenes. In endoscopy, this challenge is particularly complicated as complex lighting conditions come in addition to the lack of textures. This chapter presents a DOF-based SfM algorithm by focussing on the stage of homologous point detection, the second SfM stage being that of the state-of-the-art incremental SfM pipeline detailed in [SF16].

3.1 Overview of the algorithm principle and chapter structuration

The proposed algorithm consists of following parts:

(a) First, homologous points are tracked by computing the DOF fields along the video-sequence and the centres of all images are localized into a common 2D image trajectory plane. As shown in Fig 3.1(a), the images are all approximatively placed in this 2D plane by considering that the image centres are linked by translations.

(b) As sketched in Fig. 3.1(b), the image positions expressed in the 2D plane mentioned in point (a) are used to estimate a (rough) overlap area for all pairs of images $\mathbf{I}_i$ and $\mathbf{I}_j$ with $i \neq j$.

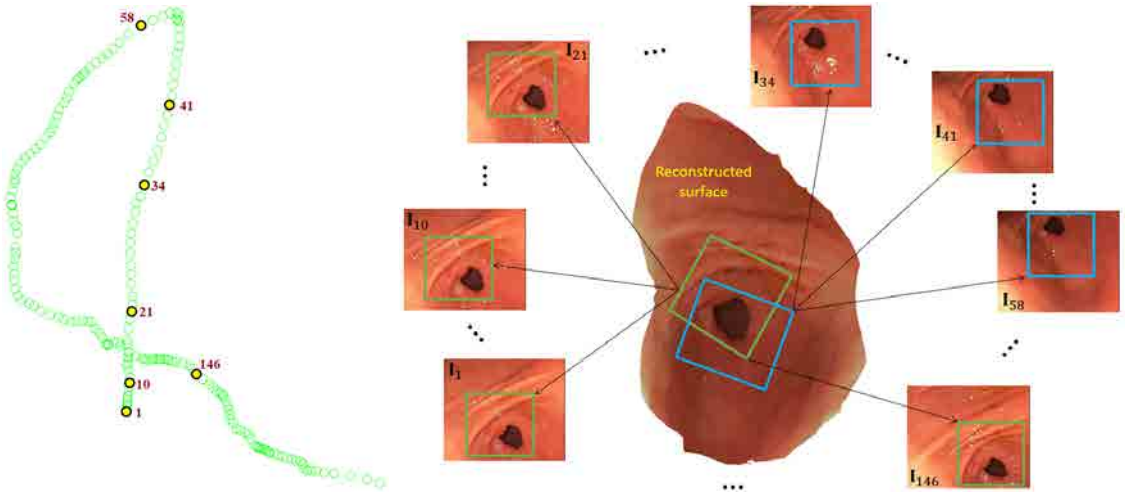
(c) Then, the set of all image pairs with a significant overlap (i.e. with an overlap area which is above a threshold) are used to search for reference images $\mathbf{I}_i^{ref}$. A reference image is an image which is overlapped by as much as possible images so that numerous and large homologous points groups can be determined with the set S_i of images overlapping $\mathbf{I}_i^{ref}$. The proposed algorithm maximizes the size of all image sets S_i and ensures that the chosen set of reference images $\mathbf{I}_i^{ref}$ leads to a 3D point cloud which covers the whole surface of interest. Since reference image $\mathbf{I}_i^{ref}$ has a common area with all images of its set S_i, no point tracking is required to compute homologous points. The point correspondence can be accurate since it is only determined between two images and not along an image sequence.

(d) A DOF approach is finally used to determine the homologous points included in the images of the sets S_i and required for a robust and accurate 3D point reconstruction using the SfM-algorithm.

After a brief introduction on OF-principles, this chapter describes a robust illumination-invariant OF method that delivers accurate correspondences even for weakly structured and textured images. Section 3.2 presents a new patch-based descriptor leading to an illumination invariant data-term which plays a central role in the proposed OF scheme. This OF method is used both in point (a) for the tracking of homologous points along the video-sequence (here a precise and dense flow field is not necessarily required) and in point (d) for the precise and dense homologous point group determination.

Section 3.3 describes the method used to assess the overlap area between image pairs (point (b)). It notably justifies why at this stage a rough positioning of the images using translations suits to the overlap area estimation. This section also details the image grouping strategy which maximizes the sizes of the homologous point sets.

After that, Section 3.4 explains how the OF scheme presented in Section 3.2 can be used to select only homologous points which effectively correspond to accurate point correspondences.



(a) (b)

Figure 3.1: Reconstruction example of a 3D scene using a sequence of 2D views (191 images) by exploiting the overlapping regions between images. (a) Image center trajectory in a 2D plane whose coordinate system is defined by the first image of the sequence images. (b) The overlapping region between image pairs was estimated by the displacement between the centres of those images in (a). As shown in Fig. 3.1(b), image $\mathbf{I}_1$ is overlapped by images $\mathbf{I}_{10}, \mathbf{I}_{21}, \mathbf{I}_{146}$. Similarly, image $\mathbf{I}_{41}$ is overlapped by images $\mathbf{I}_{34}$ and $\mathbf{I}_{58}$. Images having close center coordinates in the 2D plane of (a) are overlapped and include common scene parts. For instance, the sets $\{\mathbf{I}_1, \mathbf{I}_{10}, \mathbf{I}_{21}, \mathbf{I}_{146}\}$ and $\{\mathbf{I}_{34}, \mathbf{I}_{41}, \mathbf{I}_{58}\}$ have common parts which are represented by green and blue squares, respectively.

The complete algorithm for the homologous point group determination described in Sections 3.2, 3.3 and 3.4 is also illustrated in the video accessible at following link (file Supplementary_material_Algorithm.avi):

https://github.com/CRAN-BioSiS-Imaging/PR2020

Section 3.5 describes the method for adjusting the optimal parameter values, both for the computation of the flow fields and for the determination of the groups of homologous point. This adjustment is done by integrating the image grouping strategy in the incremental SfM pipeline given in [SF16] to generate 3D points of a scene with known ground truth.

Finally, the robustness of the DOF-based SfM scheme, the main contributions and the conclusion are given in Sections 3.6 and 3.7.

3.2 Dense optical flow for complex scenes

3.2.1 Introduction

The 2D displacement vector field describing the apparent motion of a scene between two images is popularly known as optical flow. A flow field corresponds to the projection of the motion of 3D points on the image planes. An OF can be computed by

exploiting pixel values which should remain either constant (brightness constancy) or with constant properties (e.g., gradient constancy) for corresponding points seen in different images. It is an essential part of various computer vision applications such as image registration [GRA13], object detection [TM04], motion segmentation [XCJ08], multiview stereo [LCHS03], image mosaicing [ADWB13, WDW+12, TBD17], etc.

Since the precursor work proposed by [HS81], numerous studies on variational OF have been proposed, e.g. [NE86, BBPW04, ZPB07, SRB10]. A comprehensive overview of the consequent work done by the OF researcher community, as well as the general principles can be found in [FBK15, SRB14]. Despite the numerous and effective methods published during the last decades, high accuracy OF estimation remains challenging in endoscopy due to the strong scene variability in terms of textures, illumination conditions, large instrument displacements, occlusions and the acquisition conditions as blur caused by camera motion and the defocus or refocus of the lens.

In [HS81], Horn and Schunck used a variational framework that minimizes an energy including a data-fidelity term and a regularizer. The data-term is based on a brightness constancy assumption (BCA). However, in numerous scenes (as outdoor or medical scenes) the BCA assumption does not hold since the illumination conditions change. Alternatively, the authors in [BBPW04, RMG+13] proposed a gradient constancy assumption (GCA) as a complementary term along BCA to deal with illumination changes between image pairs. However, a simultaneous use of BCA and GCA in all pixels is not optimal in terms of robustness (even if the global illumination changes modeled by these two assumptions are little bit more general, the methods based on the simultaneous use of BCA and GCA are often inaccurate for changing illumimation conditions).

Recently, learning-based OF methods (e.g., PWCNet [SYLK18]) have achieved impressive OF results. However, these methods are difficult to apply to endoscopic images due to the lack of ground-truth data for training. Besides that, the patch-matching methods (e.g., FlowFields [BTS15], CPM-Flow [HSL16]) determine point correspondences by matching textures seen in image patches. These approaches allowed to speed-up the OF computation and improved the flow filed accuracy for textured scenes. But, they are often inoperative for weakly structured/textured scenes. Recent contributions in OF have shown the appropriateness of variational methods [ADGB16, MRM+14] for scenes with weak textures and with strong illumination changes.

Ali et *al.* [ADGB16] have shown that the descriptor-based methods are able to preserve the accuracy of the OF under changing illumination conditions. Although this method was able to deliver precise flow fields for image mosaicing in endoscopy, no mathematical justification was given to prove the invariance towards illumination changes. To prove this invariance, Trinh et *al.* [TBD17, TD19] proposed a unified theoretical basis for defining illumination invariant descriptors and to facilitate their design. To do so, these authors proposed two general formulations of illumination invariant descriptors allowing to compute an OF with high and constant accuracy. The work in [TD19] has shown that a variational OF approach can deal

with complex endoscopic scenes observed through images including few textures and strong illumination changes. In [TDBL18], the data and regularization terms were designed to reduce as much as possible the effects of locally varying illuminations and specular reflections (SR) which usually occur in endoscopic scenes. Tests on complex endoscopic datasets confirmed the robustness of the descriptors proposed in [TD19, TDBL18].

This thesis exploits the robust OF variational scheme proposed in [TBD17] and makes the attempt to derive a new descriptor from the general illumination invariant descriptor formulations proposed in [TD19]. The aim of the new descriptor is to lead to a robust data-term allowing to capture enough information for the OF computation between few textured endoscopic images.

3.2.2 Optical flow estimation

The variational model for determining the flow field from source image $\mathbf{I}_s$ to target image $\mathbf{I}_t$ is defined as

$$\min_{\mathbf{u}} \left[E_{reg}\left(\mathbf{u}\right) + \lambda E_{data}\left(\mathbf{I}_s, \mathbf{I}_t, \mathbf{u}\right) \right], \tag{3.2.1}$$

where $\mathbf{u}\left(u_x, u_y\right)$ denotes the flow field, E_{reg} is a regularization term that assumes smoothness of solution $\mathbf{u}$, E_{data} is a data-term that measures the similarity of pixels in $\mathbf{I}_s$ and $\mathbf{I}_t$, and $\lambda > 0$ is a parameter controlling the relative importance between the two terms.

In endoscopic imaging, images are often affected by uncontrolled illumination variations and SR. Therefore, the data and regularization terms have to be appropriately designed. To this end, we follow the variational OF model given in [TDBL18] where SR pixels and the pixels surrounding SR regions (saturated pixels) are excluded from OF estimation while local illumination variations are controlled by using an illumination-invariant descriptor in the data-term.

Precisely, SR regions in images $\mathbf{I}_s$ and $\mathbf{I}_t$ are first segmented using the method described in [MKA+11, TDBL18] and based on the search of pixels having luminance components greater than their chromatic luminance. A binary mask M_{SR} is then computed as follows:

$$M_{SR} = \left(R_{\mathbf{I}_s} \oplus se\right) \cup \left(R_{\mathbf{I}_t} \oplus se\right), \tag{3.2.2}$$

where $R_{\mathbf{I}}$ denotes a binary image in which $R_{\mathbf{I}}(i,j) = 1$ when (i,j) corresponds to coordinates of a SR pixel in image $\mathbf{I}$, and $\oplus$ is the morphological dilation operator associated with a square structuring element se [Sze11]. Values at 1 in binary mask M_{SR} correspond to pixels located either inside SR regions in $\mathbf{I}_s$ or $\mathbf{I}_t$ (pixels at 1 before dilation) or close to reflections (pixels at 1 after dilation). After determining SR pixels and their neighbors, the data- and regularization terms in Eq. (3.2.1) are defined by:

$$E_{data} = \sum_{\mathbf{x}\in\Omega} \theta_{\mathbf{x}} \left\| \mathbf{D}\left(P_{\mathbf{I}_s}\left(\mathbf{x}\right)\right) - \mathbf{D}\left(P_{\mathbf{I}_t}\left(\mathbf{x}+\mathbf{u}_{\mathbf{x}}\right)\right) \right\|_2^2, \tag{3.2.3}$$

$$E_{reg} = \sum_{\mathbf{x}\in\Omega} \sum_{\mathbf{x}'\in\mathcal{N}_{\mathbf{x}}} \theta_{\mathbf{x}}\theta_{\mathbf{x}'}\omega_{\mathbf{x}}^{\mathbf{x}'} \left\| \mathbf{u}_{\mathbf{x}} - \mathbf{u}_{\mathbf{x}'} \right\|_1, \tag{3.2.4}$$

where Ω stands for the image domain and $\mathbf{u_x}$ is the displacement vector from pixel $\mathbf{x}$ in source $\mathbf{I}_s$. L_1-regularity is used in Eq. (3.2.4) because it is known to better preserve discontinuities compared to L_2-regularity [NE86, BBPW04]. Parameter $\theta_\mathbf{x}$ equals 0 for $M_{SR}(\mathbf{x}) = 1$, and $\theta_\mathbf{x} = 1$, otherwise. This ensures that saturated pixels and their close neighbors are not involved in the OF determination. In Eq. (3.2.3), symbol $P_\mathbf{I}(\mathbf{x})$ denotes a small patch[1] centered on pixel $\mathbf{x}$ in image $\mathbf{I}$, and $\mathbf{D}(P_\mathbf{I}(\mathbf{x}))$ is a descriptor vector computed with the colours of the pixels in $P_\mathbf{I}(\mathbf{x})$. If two pixels $\mathbf{x}$ and $\mathbf{x}+\mathbf{u_x}$ are homologous, then two vectors $\mathbf{D}(P_{\mathbf{I_s}}(\mathbf{x}))$ and $\mathbf{D}(P_{\mathbf{I_t}}(\mathbf{x}+\mathbf{u_x}))$ should ideally have the same components (i.e, the norm of their difference is null). In Eq. (3.2.4), $\mathcal{N}_\mathbf{x}$ is the set of neighbor pixels $\mathbf{x}'$ in a rectangular region centered on $\mathbf{x}$, and $\omega_\mathbf{x}^{\mathbf{x}'}$ is a weighting function which is used to define the mutual support between the pixels at positions $\mathbf{x}$ and $\mathbf{x}'$. The support-weight is computed based both on the color-similarities of pixels, and on their spatial distances:

$$\omega_\mathbf{x}^{\mathbf{x}'} = \exp\left(\frac{-\|\mathbf{x}-\mathbf{x}'\|_2^2}{\gamma_1} + \frac{-\|\mathbf{c}_{\mathbf{I}_s}(\mathbf{x}) - \mathbf{c}_{\mathbf{I}_s}(\mathbf{x}')\|_2^2}{\gamma_2}\right). \tag{3.2.5}$$

Vector $\mathbf{c}_{\mathbf{I}_s}(\mathbf{x}) = [L(\mathbf{x}), a(\mathbf{x}), b(\mathbf{x})]$ encodes the color of image $\mathbf{I}_s$ at pixel $\mathbf{x}$ in the CIELab space [YK06], while γ_1 and γ_2 are parameters controlling the importance of the colour similarity and the spatial distance.

The epithelial images often include few contrasted textures and structures and are affected by strong illumination changes due, for instance, to viewpoint changes between two image acquisitions. Descriptor vector $\mathbf{D}(P_\mathbf{I}(\mathbf{x}))$ of the data-term in Eq. (3.2.3) has to capture weak intensity variations, while being invariant to illumination changes between $\mathbf{I}_s$ and $\mathbf{I}_t$.

A new descriptor is proposed to perceive local and weak textures/intensity changes in patches $P_\mathbf{I}(\mathbf{x})$ having a size of 3×3 pixels and centered on $\mathbf{x}$. Twelve 3×3 convolution kernels K_1, K_2,...,K_{12} shown in Fig. 3.2 are used to capture intensity variations in patches $P_\mathbf{I}^g$ ($P_\mathbf{I}^g$ are grey-level patches computed with the original RGB image patches $P_\mathbf{I}$). Kernels K_d with $d = 1, \ldots, 8$ allow to encode gradient components approximating line segments with different orientations. Kernels K_d with $d = 9, 10, 11, 12$ are rather similar to corner detectors where the vertex of the detected corners are oriented in the direction of positive x-axis values ($d = 10$), of positive y-axis values ($d = 9$), of negative x-axis values ($d = 12$), or of negative y-axis values ($d = 11$). The descriptor vector with twelve components is defined as follows.

[1] The size of the descriptor patches relates to the illumination variation model detailed in [TD19]. To sum up, illumination changes between two small homologous rectangular regions of images $\mathbf{I}_s$ and $\mathbf{I}_t$ are modelled by an affine relationship between the colors. Both the multiplicative and the additive coefficients of the affine relationship are constant for all pixels of two homologous regions. Complex illumination changes can be modelled by choosing a size of 3×3 pixels for these regions (the illumination differences can by locally very strong since the values of the coefficients can vary for each small homologous region pairs of $\mathbf{I}_s$ and $\mathbf{I}_t$). The descriptor patches have the same size as the small regions in this illumination change model (3×3 pixels) and, as shown in this section, the values of the components of the descriptor vectors have to be independent of the values of the coefficients of the affine relationship between the colors.

$$K_1 = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad K_2 = \begin{bmatrix} 0 & -1 & -1 \\ 0 & 3 & -1 \\ 0 & 0 & 0 \end{bmatrix} \quad K_3 = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 3 & -1 \\ 0 & 0 & -1 \end{bmatrix} \quad K_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3 & -1 \\ 0 & -1 & -1 \end{bmatrix}$$

$$K_5 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3 & 0 \\ -1 & -1 & -1 \end{bmatrix} \quad K_6 = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 3 & 0 \\ -1 & -1 & 0 \end{bmatrix} \quad K_7 = \begin{bmatrix} -1 & 0 & 0 \\ -1 & 3 & 0 \\ -1 & 0 & 0 \end{bmatrix} \quad K_8 = \begin{bmatrix} -1 & -1 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$K_9 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & 0 & 0 \end{bmatrix} \quad K_{10} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 3 & -1 \\ 0 & -1 & 0 \end{bmatrix} \quad K_{11} = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 0 \end{bmatrix} \quad K_{12} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 3 & 0 \\ 0 & -1 & 0 \end{bmatrix}$$

Figure 3.2: Convolution kernels used to compute the components of the new illumination-invariant descriptor (vector $\mathbf{D}$ in Eq. (3.2.3)).

$$\mathbf{D}(P_\mathbf{I}) = \frac{\mathcal{V}(P_\mathbf{I})}{\|\mathcal{V}(P_\mathbf{I})\|_2}, \tag{3.2.6}$$

with $\mathcal{V}(P_\mathbf{I}) = [K_1 \otimes P^g_\mathbf{I}, K_2 \otimes P^g_\mathbf{I}, \ldots, K_{12} \otimes P^g_\mathbf{I}]^T \in \mathbb{R}^{12}$, where $\otimes$ denotes the convolution operator. In patches $P^g_\mathbf{I}$, the central pixel (whose grey-level value is multiplied by 3) can be seen as the origin of a star shaped structure from which grey-level variations are computed along 12 directions. These grey-level variations encode the shape and sharpness of the local texture or intensity variations. With or without illumination changes between images, two descriptors vectors $\mathbf{D}\left(P_{\mathbf{I}_s}(\mathbf{x})\right)$ and $\mathbf{D}\left(P_{\mathbf{I}_t}(\mathbf{x}+\mathbf{u}_\mathbf{x})\right)$ should have the same component values. In [TD19] it was shown that a descriptor $\mathbf{D}$ is invariant to illumination changes when:

$$\mathbf{D}(P_\mathbf{I}) = \mathbf{D}(a_\mathbf{x} P_\mathbf{I} + b_\mathbf{x}), \forall a_\mathbf{x} \in \mathbb{R}_{>0}, \forall b_\mathbf{x} \in \mathbb{R}. \tag{3.2.7}$$

As seen in Fig. 3.2, the sum of the coefficients is null in each convolution kernel K_d. It follows that the effect of additive term $b_\mathbf{x}$ is compensated since

$$K_d \otimes (a_\mathbf{x} P^g_\mathbf{I} + b_\mathbf{x}) = a_\mathbf{x}(K_d \otimes P^g_\mathbf{I}), \forall d = 1, 2, \ldots, 12. \tag{3.2.8}$$

This leads to $\mathcal{V}(a_\mathbf{x} P_\mathbf{I} + b_\mathbf{x}) = a_\mathbf{x}\mathcal{V}(P_\mathbf{I})$ and $\|\mathcal{V}(a_\mathbf{x} P_\mathbf{I} + b_\mathbf{x})\|_2 = a_\mathbf{x}\|\mathcal{V}(P_\mathbf{I})\|_2$. Therefore, the effect of multiplicative term $a_\mathbf{x}$ is also compensated:

$$\frac{\mathcal{V}(a_\mathbf{x} P_\mathbf{I} + b_\mathbf{x})}{\|\mathcal{V}(a_\mathbf{x} P_\mathbf{I} + b_\mathbf{x})\|_2} = \frac{a_\mathbf{x}\mathcal{V}(P_\mathbf{I})}{a_\mathbf{x}\|\mathcal{V}(P_\mathbf{I})\|_2} = \frac{\mathcal{V}(P_\mathbf{I})}{\|\mathcal{V}(P_\mathbf{I})\|_2} \tag{3.2.9}$$

$$\Leftrightarrow \mathbf{D}(a_\mathbf{x} P_\mathbf{I} + b_\mathbf{x}) = \mathbf{D}(P_\mathbf{I}), \forall a_\mathbf{x} \in \mathbb{R}_{>0}, \forall b_\mathbf{x} \in \mathbb{R}. \tag{3.2.10}$$

Thus, vector $\mathbf{D}$, as defined in Eq. (3.2.6), is an illumination-invariant descriptor.

In this work, the optimization problem defined by Eqs. (3.2.1), (3.2.3), and (3.2.4) is solved using the projection-proximal point algorithm [CP11, DN13]. Moreover, the well-known coarse-to-fine multiscale warping strategy is also used to deal with large displacements. The optimal parameter values used in the OF computation are given in Subsection 3.5.1.

A visualization of DOF field between two gastroscopic images obtained by the proposed method is showed in Fig. 3.6(c). The effectiveness of the proposed OF method for determining homologous points is demonstrated in Figs. 3.15(b), 3.15(d) and 3.15(f). As visually perceptible for the same gastroscopic image pair as in Figs. 3.15(a), 3.15(c) and 3.15(e) (SIFT feature matching), the use of OF can robustly determine homologous point-pairs.

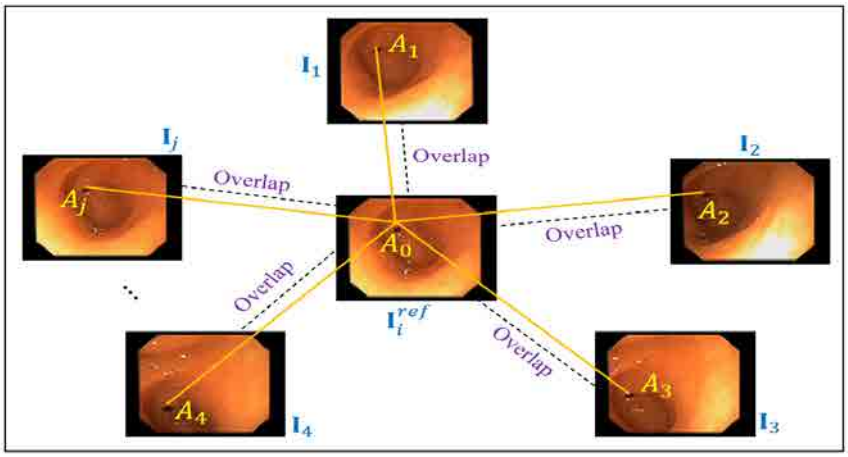


Figure 3.3: HP-group example. Pairs $(\mathbf{I}_i^{ref}, \mathbf{I}_j)$ consist both of consecutive and non-consecutive images in video-sequence S.

3.3 Determination of groups of images with common scene parts

Suppose that the input of the SfM algorithm is video-sequence $S = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N\}$ of N temporally numbered images having a size of $H \times W$ pixels. Let $Z = \{1, 2, \ldots, N\}$ be the index set of S. The proposed matching method is based on the fact that if $\{A_k\}_{k \in Z_i \subset Z}$ is a group of 2D homologous points on images in sequence S, then $A_k \in \bigcap_{j \in Z_i} \mathbf{I}_j, \forall k \in Z_i$. Therefore, to find homologous point groups, our idea is to firstly determine reference images, referred to as $\mathbf{I}_i^{ref}$, which have an overlap with a maximum of other images.

The set of images overlapping $\mathbf{I}_i^{ref}$ is denoted by $S_i = \{\mathbf{I}_k\}_{k \in Z_i} \subset S$. Then, if A_0^{ref} is a point in $\mathbf{I}_i^{ref}$ and A_k $(k \in Z_i)$ are corresponding points of A_0^{ref} in images $\mathbf{I}_k$ with $k \in Z_i$ and $i \neq k$, then set $\{A_0^{ref}, A_k\}_{k \in Z_i}$ is defined as a group of homologous points or, in abbreviated form, as a *HP-group* (see Fig. 3.3).

The next two sub-sections successively present the method for determining overlapping image pairs in sequence S and the algorithm which determines the reference images $\mathbf{I}_i^{ref}$, as well as their corresponding sets S_i.

3.3.1 Overlap estimation

Definition 1 *Two images $\mathbf{I}_i$ and $\mathbf{I}_j$ are called τ-overlapped when their common area $\mathbf{I}_i \cap \mathbf{I}_j$ is greater than a given threshold τ in pixels.*

When the acquisition distance is small (e.g., as in gastroscopy or cystoscopy where the endoscope's distal tip is close to the tissue), the FoV is limited and the displacement field between consecutive images mainly consists of almost parallel translation vectors. For this reason, simple translations can be used to represent the displacement between common scene parts approximated by rectangular sub-regions in the images.

As sketched in Fig. 3.4, the translation vector between two images $\mathbf{I}_i$ and $\mathbf{I}_j$ in S is denoted by $\mathbf{v}_{i,j}(v^1_{i,j}, v^2_{i,j})$, where $v^1_{i,j}$ and $v^2_{i,j}$ are the vector components. Vector $\mathbf{v}_{i,j}(v^1_{i,j}, v^2_{i,j})$ is extracted from the DOF fields $\mathbf{F}_{t,t+1}$ between the consecutive images $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$ of sequence $\mathbf{I}_i, \mathbf{I}_{i+1}, \ldots, \mathbf{I}_{j-1}, \mathbf{I}_j$. The motion vector at the central pixel $(W/2, H/2)$ of image $\mathbf{I}_t$ to image $\mathbf{I}_{t+1}$ is denoted by $\mathbf{c}_{t,t+1}(c^1_{t,t+1}, c^2_{t,t+1})$ with:

$$\mathbf{c}_{t,t+1}(c^1_{t,t+1}, c^2_{t,t+1}) = \mathbf{F}_{t,t+1}\left(\frac{W}{2}, \frac{H}{2}\right). \tag{3.3.1}$$

If two images $\mathbf{I}_i$ and $\mathbf{I}_j$ are consecutive (i.e., $|i-j|=1$), then the translation vector between images pair $(\mathbf{I}_i, \mathbf{I}_j)$ is defined by:

$$\mathbf{v}_{i,j}(v^1_{i,j}, v^2_{i,j}) = \mathbf{c}_{i_0,i_0+1}(c^1_{i_0,i_0+1}, c^2_{i_0,i_0+1}), \tag{3.3.2}$$

with image index $i_0 = \min(i,j)$ making Eq. (3.3.2) valid for two cases: $j = i-1$ and $j = i+1$. For two non-consecutive images $\mathbf{I}_i$ and $\mathbf{I}_j$ (i.e., $|i-j|>1$), two image indexes are considered : $i_0 = \min(i,j)$ and $j_0 = \max(i,j)$. In this case, the translation between $\mathbf{I}_i$ and $\mathbf{I}_j$ is defined (both for $i>j$ and $i<j$) by the sum of the translation vectors between consecutive images from $\mathbf{I}_{i_0}$ to $\mathbf{I}_{j_0}$:

$$\mathbf{v}_{i,j}(v^1_{i,j}, v^2_{i,j}) = \sum_{t=i_0}^{j_0-1} \mathbf{c}_{t,t+1}(c^1_{t,t+1}, c^2_{t,t+1}). \tag{3.3.3}$$

Two images $\mathbf{I}_i$ and $\mathbf{I}_j$ with translation vector $\mathbf{v}_{i,j}$ are τ-overlapped when the following condition is fulfilled:

$$\begin{cases} -W < v^1_{i,j} < W \\ -H < v^2_{i,j} < H \\ Area_{i,j} = (W - |v^1_{i,j}|)(H - |v^2_{i,j}|) \geq \tau, \end{cases} \tag{3.3.4}$$

where W and H are the width and height of the images, $Area_{i,j}$ is the overlap area in pixels, and τ $(0 < \tau \leq WH)$ is a threshold parameter. The two first equations in Eq. (3.3.4) ensure that $\mathbf{I}_i \cap \mathbf{I}_j \neq \emptyset$, whereas the third equation defines the area of the overlap region $\mathbf{I}_i \cap \mathbf{I}_j$.

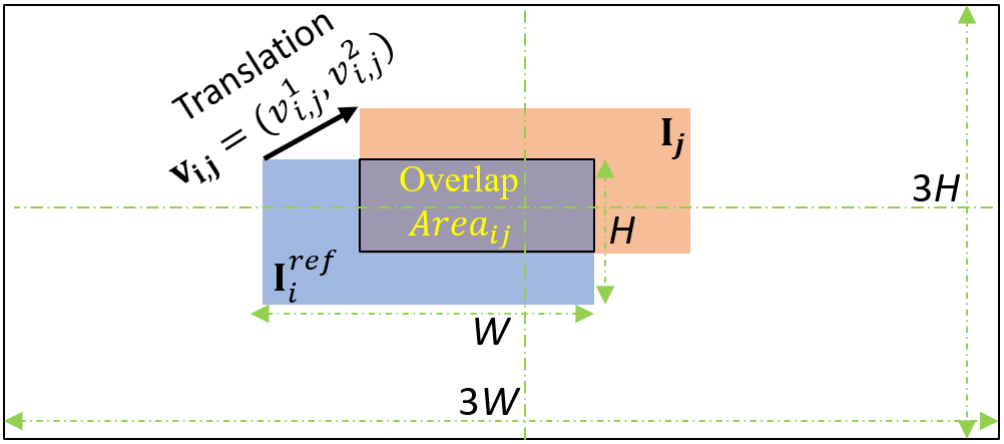


Figure 3.4: Rectangular overlap area of $\mathbf{I}_j$ and $\mathbf{I}_i^{ref}$, and vector $\mathbf{v}_{i,j}$ between their centres.

3.3.2 Reference images

The proposed algorithm determines the reference images $\mathbf{I}_i^{ref}$ which favor groups consisting of numerous homologous points. From the practical point of view, such $\mathbf{I}_i^{ref}$ images are those including scene parts which are geometrically surrounded by other scene parts seen in numerous other images acquired from different viewpoints and having common areas with $\mathbf{I}_i^{ref}$. References $\mathbf{I}_i^{ref}$ are images in S that simultaneously fulfill two conditions: *(i)* a reference image must be τ-overlapped with as much as possible of other images, and *(ii)* two reference images cannot be τ-overlapped. The first condition ensures that HP-groups involve numerous images, whereas the second condition favours the distribution of the 3D points over the complete surface.

Fig. 3.5 illustrates the importance of an appropriate choice of reference images $\mathbf{I}_i^{ref}$ in terms of 3D point reconstruction accuracy (SfM methods require numerous homologous points acquired from different viewpoints to determine accurate 3D point positions) and in terms of gapless surfaces. In this example, only two reference images enable already to reconstruct a surface part representing 104 images (the whole surface was reconstructed with seven reference images).

The proposed algorithm (see Algorithm 2) for the determination of the reference images consists of two parts.

Part 1: Determination of the τ-overlapped image sets. A set S_i (with $i = 1, 2, \ldots, N$) of τ-overlapped images consists of all images $\mathbf{I}_j$ of S which are τ-overlapped with $\mathbf{I}_i$, and of $\mathbf{I}_i$ itself. At the beginning of part 1, $S_i = \{\mathbf{I}_i\}$ for all i. For all image pairs $(\mathbf{I}_i, \mathbf{I}_j)$ with $j \neq i$, translation vector $\mathbf{v}_{i,j}$ is computed differently depending on whether $\mathbf{I}_i$ and $\mathbf{I}_j$ are consecutive images or not. When $|i-j| = 1$ (consecutive images), translation $\mathbf{v}_{i,j}$ is computed with Eq. (3.3.2). If $|i-j| > 1$ (non-consecutive images), vector $\mathbf{v}_{i,j}$ is obtained with Eq. (3.3.3). Set S_i is updated with image $\mathbf{I}_j$ only when the τ-overlap condition given in Eq. (3.3.4) is fulfilled for image pair $(\mathbf{I}_i, \mathbf{I}_j)$. This algorithm part leads to set G gathering all sets S_i: $G = \{S_1, S_2, \ldots, S_N\}$.

Part 2: Reference image determination. Let $\mathbf{I}_i^{ref}$ be a reference image and let Ω^{ref} ($\Omega^{ref} \subset S$) be the set of reference images maximizing the number $|S_i|$ of τ-overlapped images of set S_i associated with $\mathbf{I}_i^{ref}$. At each iteration of part 2, the algorithm searches for set S_i in G with the highest image number $|S_i|$. Image $\mathbf{I}_i^{ref}$ is added to $\Omega^{ref} \subset S$ and becomes a reference image. Before the next iteration, all image sets S_j corresponding to an image $\mathbf{I}_j \in S_i$ are removed from set G. The iterative process ends when set G is empty. After the last iteration, all reference images are gathered in set Ω^{ref} and image group S_i is known for each $\mathbf{I}_i^{ref}$.

The reference images and their overlapping images is then used to generate HP-groups based on the DOF fields detailed in Section 3.4.

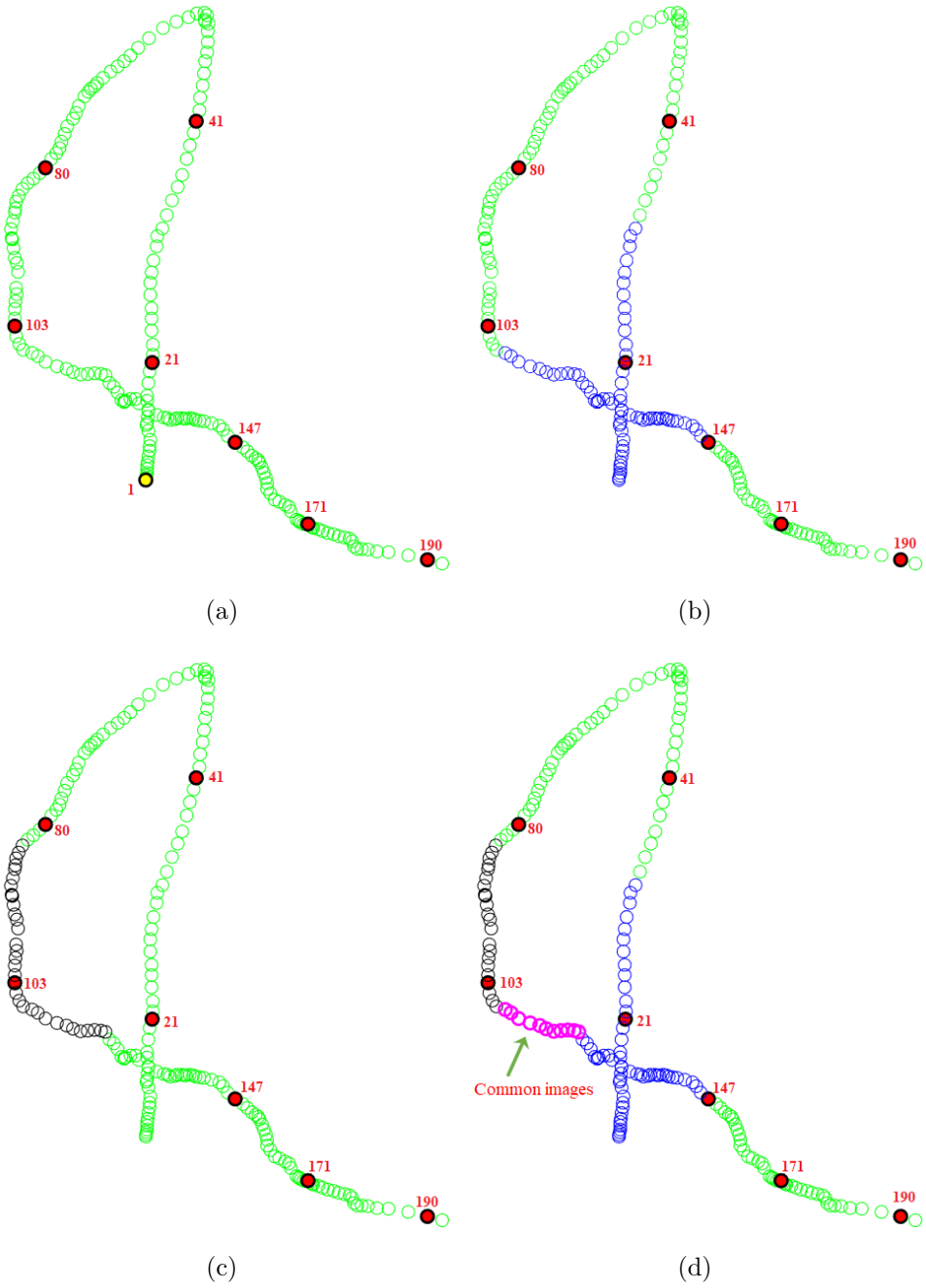


Figure 3.5: Reference images and their overlapping image set S_i determined for a gastroscopic video-sequence consisting of 191 images. (a) Starting from the central pixel of first image $\mathbf{I}_1$ represented by the yellow disc, the central pixels of all images (green discs) are placed in a 2D image trajectory plane. The positions of the green discs are estimated with the motion vectors between the consecutive images (see Eqs. (3.3.1)-(3.3.3)). The red discs represent the reference images that are determined by Algorithm 2 which is described in this section. (b) Reference image $\mathbf{I}_{21}^{ref}$ and the associated set S_{21} of 71 $\tau-$overlapped images represented by blue discs. Such a large image set ensures a robust and accurate 3D point reconstruction. (c) Reference image $\mathbf{I}_{103}^{ref}$ and set S_{103} of 33 $\tau-$overlapped images corresponding to the black discs. (d) The common images (pink discs) of sets S_{21} and S_{103} ensure that locally the surface is without a gap of 3D points.

Algorithm 2 Reference image determination

Input: Set S of N consecutive images $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N$, area threshold τ, and central flow field vectors $\mathbf{c}_{1,2}$, $\mathbf{c}_{2,3}, \ldots$, $\mathbf{c}_{N-1,N}$ given by Eq. (3.3.1).
Initialization: $\Omega^{ref} = \emptyset$.

/* ***Part 1***: *Determination of sets* S_i: */
for $i = 1$ to N **do**
  $S_i = \{\mathbf{I}_i\}$
  for $j = 1$ to N **do**
    if $(|i - j| = 1)$ **then**
      Compute $\mathbf{v}_{i,j}$ using Eq. (3.3.2).
    end if
    if $(|i - j| > 1)$ **then**
      Compute $\mathbf{v}_{i,j}$ using Eq. (3.3.3).
    end if
    if $(j \neq i)$ and ($\mathbf{I}_j$ is τ-overlapped with $\mathbf{I}_i$) **then**
      $S_i \leftarrow S_i \cup \mathbf{I}_j$
    end if
  end for
end for
$G = \{S_1, S_2, \ldots, S_N\}$.

/* ***Part 2***: *Reference image determination:* */
while $G \neq \emptyset$ **do**

- $\Omega^{ref} \leftarrow \Omega^{ref} \cup \mathbf{I}_i$, where i satisfies $|S_i| \geq |S_k|$, for all $S_{k \neq i} \in G$.
- For all images $\mathbf{I}_j \in S_i$, removing corresponding set S_j from G:

$$G \leftarrow G \setminus \bigcup_{j:\mathbf{I}_j \in S_i} S_j. \tag{3.3.5}$$

end while
Output: Set Ω^{ref} of images $\mathbf{I}_i^{ref}$ and their groups S_i.

3.4 Homologous point set determination for SfM

After obtaining the set of reference images $\Omega^{ref} = \{\mathbf{I}_i^{ref}\}_{i \in \hat{Z} \subset Z}$ (where $\hat{Z}$ denotes the index set of the reference images) and the sets $S_i = \{\mathbf{I}_j\}_{j \in Z_i \subset Z}$ with $i \in \hat{Z}$, HP-groups can be easily established based on the DOF fields between images $\mathbf{I}_i^{ref}$ and their τ-overlapped images belonging to the sets S_i. Suppose the DOF fields $\mathbf{F}_{i,j}$ between $\mathbf{I}_i^{ref}$ and images $\mathbf{I}_j$ in S_i as determined. For every reference image $\mathbf{I}_i^{ref}$ in Ω^{ref}, one first considers the set Ξ_i^{ref} of regularly distributed 2D points $A_{xy}^{i,ref}$ on $\mathbf{I}_i^{ref}$ given by

$$\Xi_i^{ref} = \left\{ A_{xy}^{i,ref}(xh, yh) \mid x, y \in \mathbb{N}, x \leq \frac{W}{h}, y \leq \frac{H}{h} \right\}, \tag{3.4.1}$$

where parameter h represents the distance between neighbor points of a grid visible in Fig. 3.6(i).

Then, for each point $A_{xy}^{i,ref} \in \Xi_i^{ref}$ in $\mathbf{I}_i^{ref}$ which is not indicated as a specular reflection pixel in mask M_{SR} defined by Eq. (3.2.2), one computes the corresponding points in images $\mathbf{I}_j \in S_i$ using the DOF fields $\mathbf{F}_{i,j}$:

$$A_{xy}^{j} = A_{xy}^{i,ref} + \mathbf{F}_{i,j}(A_{xy}^{i,ref}), \forall j \in Z_i. \tag{3.4.2}$$

In the proposed SfM pipeline, not only pixels in SR regions (mask M_{SR}), but also those in occluded regions are excluded from the homologous point determination. The term "occluded" refers classically to scene parts visible only in one image. For non-occluded pixels, the forward OF from the first image should be the opposite of the backward OF at the corresponding pixels in the second image. Thus, for two images $\mathbf{I}_i^{ref}$ and $\mathbf{I}_j$, every pixels $A_{xy}^{i,ref}$ in $\mathbf{I}_i$ violating at least one of the three constraints:

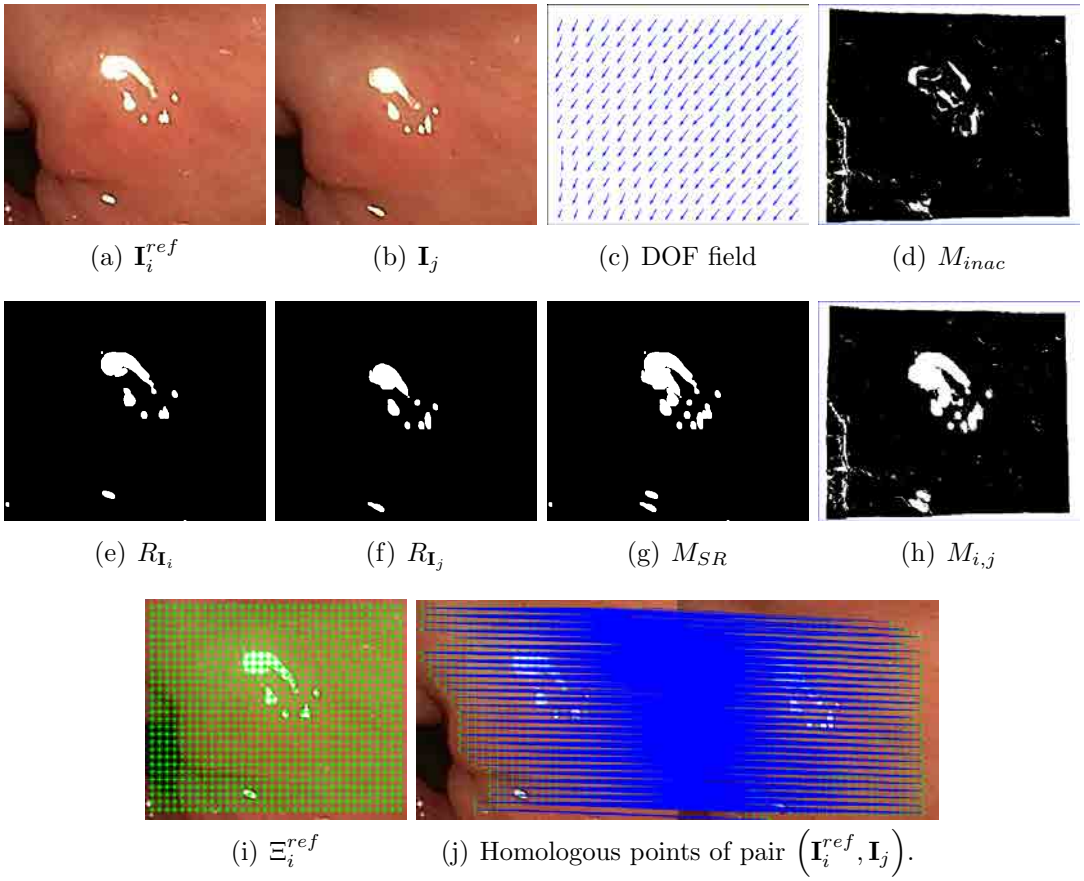
$$\begin{cases} A_{xy}^{j} = A_{xy}^{i,ref} + \mathbf{F}_{i,j}\left(A_{xy}^{i,ref}\right) \\ A'^{i,ref}_{xy} = A_{xy}^{j} + \mathbf{F}_{j,i}\left(A_{xy}^{j}\right) \\ \|A_{xy}^{i,ref} - A'^{i,ref}_{xy}\|_2 \leq \epsilon \end{cases} \tag{3.4.3}$$

is marked as having an inaccurate flow field vector, where a weak ϵ threshold parameter value ensures an accurate pixel correspondence. Both occluded pixels and pixels with too inaccurate OF vectors are encoded in binary image M_{inac}, where $M_{inac}(\mathbf{x}) = 1$ refers either to an occluded pixel (also detected with Eq. (3.4.3)) or a pixel without a very accurate OF vector (such a pixel is not necessarily associated with a wrong OF vector, but it simply not leads to a correspondence with a high accuracy). Binary mask $M_{i,j}$ defined as $M_{i,j} = M_{SR} \cup M_{inac}$ is used to mark pixels which will be excluded from the homologous point determination of two images $\mathbf{I}_i^{ref}$ and $\mathbf{I}_j$. An example of mask $M_{i,j}$ can bee seen in Fig. 3.6(h). The flow field obtained using the proposed variational OF method for the textureless image pair $(\mathbf{I}_i^{ref}, \mathbf{I}_j)$ in Figs. 3.6(a)-3.6(b) is illustrated in Fig. 3.6(c). Only the OF vectors corresponding to black pixels which verify $M_{i,j} = 0$ in Fig. 3.6(h) are used to determine the homologous point sets.

Finally, a HP-group is defined by a point $A_{xy}^{i,ref}$ in $\mathbf{I}_i^{ref}$ and all its homologous points in images $\mathbf{I}_j \in S_i$. It is noticeable that with the proposed method numerous HP-groups consisting each of a large amount of accurately matched points can be established. The number of HP-groups depends on the value of parameter h and of the overlap parameter τ. The important point of the algorithm is that HP-groups can be robustly and accurately determined because no optical flow errors accumulate due to a point tracking along a image sequence. Besides that, instead of using a field of null vectors as starting point of the variational OF algorithm described in Subsection 3.2.2, the OF computation is initialized with a field of constant vectors $\mathbf{v}_{i,j}$ corresponding to the translation computed in Eq. (3.3.2) for consecutive images ($|i - j| = 1$) or in Eq. (3.3.3) for non-consecutive images ($|i - j| > 1$).

3.5 Parameter value adjustment for the HP-group determination

The aim of this section is to estimate a set of parameter values that enable a robust



(a) $\mathbf{I}_i^{ref}$ (b) $\mathbf{I}_j$ (c) DOF field (d) M_{inac}

(e) $R_{\mathbf{I}_i}$ (f) $R_{\mathbf{I}_j}$ (g) M_{SR} (h) $M_{i,j}$

(i) Ξ_i^{ref} (j) Homologous points of pair $\left(\mathbf{I}_i^{ref}, \mathbf{I}_j\right)$.

Figure 3.6: Determination of valid homologous points between image pair $(\mathbf{I}_i^{ref}, \mathbf{I}_j)$. (a) Source image $\mathbf{I}_i^{ref}$. (b) Target image $\mathbf{I}_j$. (c) DOF field from $\mathbf{I}_i^{ref}$ to $\mathbf{I}_j$. For the sake of visibility, not all vectors of the dense field are represented. (d) White pixels in mask M_{inac} indicate inaccurate OF vectors. (e) SR regions segmented in source image $\mathbf{I}_i^{ref}$. (f) SR regions segmented in target image $\mathbf{I}_j$. (g) Specular reflection mask M_{SR} computed with Eq. (3.2.2) by using jointly the segmented source (e) and target (f) images. (h) Mask $M_{i,j}$ taking into account all error sources: specular reflections, occlusions, and inaccurate correspondences. (i) Grid of 2D points in reference image $\mathbf{I}_i^{ref}$. (j) Determination of homologous points between image pair $(\mathbf{I}_i^{ref}, \mathbf{I}_j)$.

surface construction so that the proposed SfM approach can deal with various scenes (e.g., for endoscopic scenes of different image modalities, and endoscopic and non-endoscopic scenes). The main and most crucial parameters in terms of reconstruction robustness can be classified into two groups:

- The accuracy and the robustness of the OF scheme described in Section 3.2.2 mainly depend on the appropriate setting of four parameters, namely the weights γ_1 and γ_2 of the regularization coefficient in Eq. (3.2.5), the λ coefficient which controls the tradeoff between the data-term and the regularization term in Eq. (3.2.1), and the pyramid scale Py_s in the coarse-to-fine strategy.
- Concerning the homologous point determination method proposed in Section 3.4, three parameters mainly influence the algorithm. These parameters are the optical flow error ϵ in Eq. (3.4.3), the overlap parameter τ in Eq. (3.3.4),

and the grid cell size h in Eq. (3.4.1).

3.5.1 OF computation parameters

The work in [TD19] proposed two general forms of illumination-invariant descriptors, as well as several descriptors which are designed based on zero-sum kernels. However, as showed in [TD19] the performance of OF models depends not only on the design of the kernels and their related descriptors, but also on the parameters of the OF scheme which uses these descriptors in their data-term.

In order to evaluate the descriptor vector proposed in this thesis, and to adjust the parameters of the OF algorithm, numerous experiments were performed on well-known benchmarks including images of different scenes seen under various illumination conditions. The Middleburry[2] and KITTI[3] datasets are with both small and strong illumination changes, and both small and large displacements. Different combinations of the values of the important OF scheme parameters were tested:

- $\lambda \in \{1, 2, 3, \ldots, 100\}$,
- γ_1 and $\gamma_2 \in \{1, 2, \ldots, 7\}$,
- and scale of the Pyramid: $Py_s \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$

to evaluate the performance of the proposed descriptor, and to choose the best values in terms of minimizing OF errors. The criteria used to adjust the parameters are those classically used by the OF community to evaluate the accuracy of their algorithms. These criteria are the average end-point-error (AEPE, [BSL+11]) and the average angle errors (AAE, [BFB94]). Ground truth flow fields are available for the images of the Middleburry and KITTI datasets. Let $\mathbf{u}_c$ be the ground truth flow. Then the AAE and AEEP criteria are defined as follows:

$$AAE := \frac{1}{|\Omega|} \int_\Omega \arccos\left(\frac{\mathbf{u}^T \mathbf{u}_c}{|\mathbf{u}|\,|\mathbf{u}_c|}\right) d\mathbf{x}, \tag{3.5.1}$$

$$\tag{3.5.2}$$

$$AEPE := \frac{1}{|\Omega|} \int_\Omega |\mathbf{u} - \mathbf{u}_c|\, d\mathbf{x}, \tag{3.5.3}$$

where $\mathbf{u} = (u, v, 1)^T$, $|\mathbf{u}| = \sqrt{u^2 + v^2 + 1}$ is the Euclidean norm of the flow vector, and $|\Omega|$ represents the total number of pixels $\mathbf{x}$. The values of criteria AEPE and AAE are ideally null.

All images of the Middleburry and KITTI datasets were used to systematically test all parameter combinations whose intervals were given above. Among these numerous experiments, this section focuses on representative results obtained for two image pairs (see Fig. 3.7), namely, the original RubberWhale image pair (small displacement and without illumination changes) in the Middlebury benchmark, and

[2] http://vision.middlebury.edu/flow/data/
[3] http://www.cvlibs.net/datasets/kitti/eval_flow.php

(a) Image pair without illumination changes

(b) Image pair with strong illumination gradients

Figure 3.7: Images pairs used to illustrate the adjustability of the λ and Py_s scale parameters for different illumination conditions. (a) Image pair without illumination changes between the two viewpoints. (b) Image pair simulating a strong illumination change. The upper part of the image on the left includes a strong (bright) light gradient, while another bright gradient affects the bottom of the image on the right.

the image pair with strong illumination changes which are simulated from the original images. Parameters γ_1 and γ_2 were set to 3 and 5 (as recommanded by [TD19]), respectively.

Fig. 3.8 shows the experimental results obtained in terms of AEPE and AAE for the two image pairs in Fig. 3.7 according to the values of parameters λ and Py_s. As seen on this figure, for strongly changing illumination conditions (right column), the optimal values of the pyramid scale parameter (Py_s) are in $[0.6, 0.9]$, while the optimal values of parameter λ are in $[3, 15]$. In these intervals, the weak AEPE and AAE values indicate a very low OF error. These two parameter intervals are even larger for weak illumination changes (except for $Py_s = 0.5$, the possible λ values are spread over two decades, from 1 to almost 100). These large interval values with low errors show that it is easy to adjust the parameters to obtain in a robust way an accurate OF, even for strong illumination changes. Since the computation time increases with the pyramid scale, Py_s is set to 0.7 for the experiments presented in the next section. For these coming experiments, the value of parameter λ was set to 9, since choosing the middle of the interval $[3, 15]$ which ensures high OF accuracy leads to the best safety margin in terms of robustness.

The size of structure element *se* in Eq. (3.2.2) is empirically set to 7×7 (following our experimental results, the choice of a structuring element size does not strongly affect the OF results, and it seems to be independent of image size), whereas the size of neighborhood $\mathcal{N}_x$ in Eq. (3.2.4) is 5×5. The experimental approach described

Figure 3.8: Experimental results: visualization of the impact of the values of parameter λ and of the pyramid scale Py_s on the accuracy of the proposed OF algorithm. The red ellipses in the right column indicate the intervals with constantly weak errrors.

in [TD19] was adopted to search the optimal values of parameters λ in Eq. (3.2.1), γ_1 and γ_2 in Eq. (3.2.5), as well as the pyramid scale Py_s in the coarse-to-fine strategy. Consequently, λ, γ_1, γ_2, and Py_s are experimentally set to 0.7, 3, 5 and 9, respectively. The values of these parameters are constant for all tests performed in Section 3.5.2 to adjust the point grouping parameters and in Chapter 5 for the experiments with patient data.

3.5.2 Adjustment of the point grouping parameters

The aim of this section is to adjust the parameters relating to the homologous point group determination. The important parameters are the grid cell size h determining the density of the flow field, the ϵ parameter corresponding to the maximal error allowed for each OF vector, and the τ image overlap threshold.

The two phantoms used in Section 5.1 of Chapter 5 to assess the inherent accuracy of the proposed optical-flow based SfM method are also used in this section to adjust the optimal values of parameter triplet (h, ϵ, τ). As sketched in Fig 3.9.(a), each phantom surface consists of a half cylinder carrying a sphere. The diameters of the cylinder D_{gt} and of the sphere d_{gt} are exactly known by construction and act as ground truth. Stomach and skin images were printed on paper sheets and glued on the cylinder surfaces to simulate epithelial tissue textures. The exact dimensions of the stomach and skin phantom construction are detailed in Chapter 5.

The accuracy of the 3D surface reconstruction depends strongly on the precision

and size of the homologous point groups (HP-groups) corresponding each to a 3D point. For this reason, the criteria used to adjust the triplet (h, ϵ, τ) directly relate to the robustness and accuracy of the 3D point/surface reconstruction of the optical flow based SfM method. The data used to adjust the parameter values are simply a video-sequence acquired for each of the two phantoms. The images are used to reconstruct the point cloud including both the half cylinder and the sphere. A cylinder is then fitted to the 3D points corresponding to this surface part, and a sphere is fitted to the remaining points. This procedure is described in Chapter 5. The cylinder and sphere computed in this way have a diameter D and d, respectively. Five criteria were defined to quantify the reconstruction accuracy:

- *Criterion 1.* Shape criterion. Even with a SfM method which reconstruct surfaces at an arbitrary scale, diameter ratio D_{gt}/d_{gt} (ground truth) should ideally be equal to the ratio D/d of the computed diameters. Thus, percentage p defined in Eq. (3.5.4) should ideally be equal to 100%. The value of p decreases when the shape of the surface becomes less accurate.

$$p = \left(1 - \frac{|D_{gt}/d_{gt} - D/d|}{D_{gt}/d_{gt}}\right) \times 100. \tag{3.5.4}$$

- *Criterion 2.* Outlier Rate. A 3D point is considered as an outlier when its distance to the estimated phantom surface is greater than $0.005 \times D$ (i.e., 0.5% of the cylinder diameter). The outlier rate (percentage of the number of outliers with respect to the number of reconstructed points) should be as low as possible.

- *Criterion 3.* Mean outlier error. This error corresponds to the mean distance between outlier points and the fitted phantom surface.

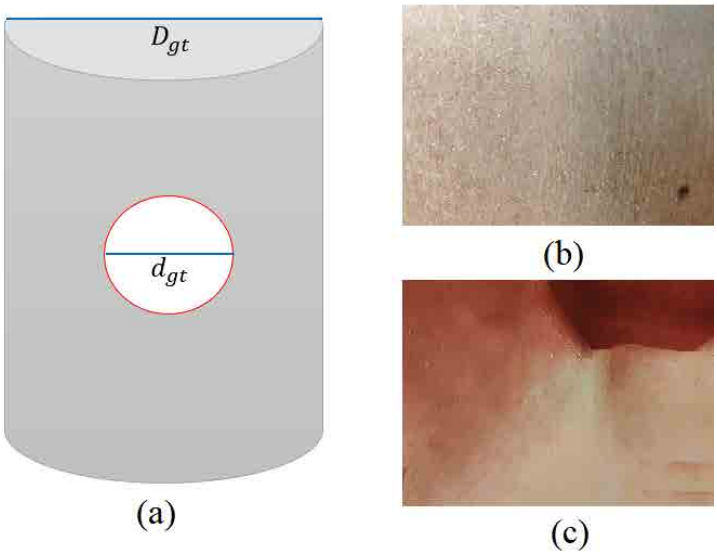


Figure 3.9: Description of the two phantoms used to adjust the point grouping parameters. (a) Geometry and dimensions of the phantoms. (b) Image of skin textures. (c) Image of stomach textures.

- *Criterion 4.* Amount of 3D points. The number of reconstructed 3D points should be as large as possible to ensure a precise surface representation.
- *Criterion 5.* Computation time. This value corresponds to the time required by the point grouping method and the SfM method to build a 3D surface.

Test with different textures (stomach and skin) favours the search of optimal parameters for the reconstruction of different scenes or organs. Numerous tests were conducted for a set of triplets (h, ϵ, τ) chosen among following values:

- $h = \{5, 10, 15, 20, 25\}$,
- $\epsilon = \{0.06, 0.08, 0.1, 0.3, 0.5, 0.7, 0.9\}$, and
- $\tau = \left\{\frac{WH}{4}, \frac{11WH}{24}, \frac{2WH}{3}, \frac{7WH}{8}\right\}$, with W and H the image width and height.

The aim of these experiments is to find values of the triplet (h, ϵ, τ) with the best compromise in terms of surface shape accuracy, the number of reconstructed 3D points (ideally it should be as high as possible), the outlier rate (as low as possible), the mean outlier error, and computation time (as low as possible).

3.5.2.1 Adjustment of the grid size h

The effect of the grid size parameter on the reconstruction results was tested with $h = \{5, 10, 15, 20, 25\}$, $\epsilon = 0.1$, and $\tau = \frac{2WH}{3}$. The later two parameters were kept constant to perceive the effect of the grid size in a situation that favours a robust 3D reconstrusction: the ϵ value is small enough to ensure the selection of very accurate OF vectors and the chosen τ value ensures large common image regions. As visible on Fig. 3.10 and in Table 3.1, the changes of the 3D shape accuracy (percentages p), of the outlier rates, and of the mean outlier errors are small for both phantom

Table 3.1: Experimental results with $\epsilon = 0.1$, $\tau = \frac{2WH}{3}$, and $h = \{5, 10, 15, 20, 25\}$.

Phantom type	Grid size h	3D phantom shape accuracy (p in %)	Outlier rate (%)	Mean outlier error (mm)	Number of 3D points	Computation time (Minutes)
Skin surface phantom	$h = 5$	99.32	3.58	7.51	558397	147.81
	$h = 10$	99.32	4.05	7.38	140204	60.94
	$h = 15$	99.20	4.38	7.3	62331	44.45
	$h = 20$	98.99	4.78	7.28	35061	36.77
	$h = 25$	97.12	5.01	7.25	22339	24.77
Stomach surface phantom	$h = 5$	98.76	5.15	5.65	297632	88.45
	$h = 10$	98.76	5.89	5.61	76808	47.88
	$h = 15$	98.66	6.37	5.59	34181	35.01
	$h = 20$	98.02	6.45	5.58	19227	20.45
	$h = 25$	96.69	7.55	5.56	12205	16.79

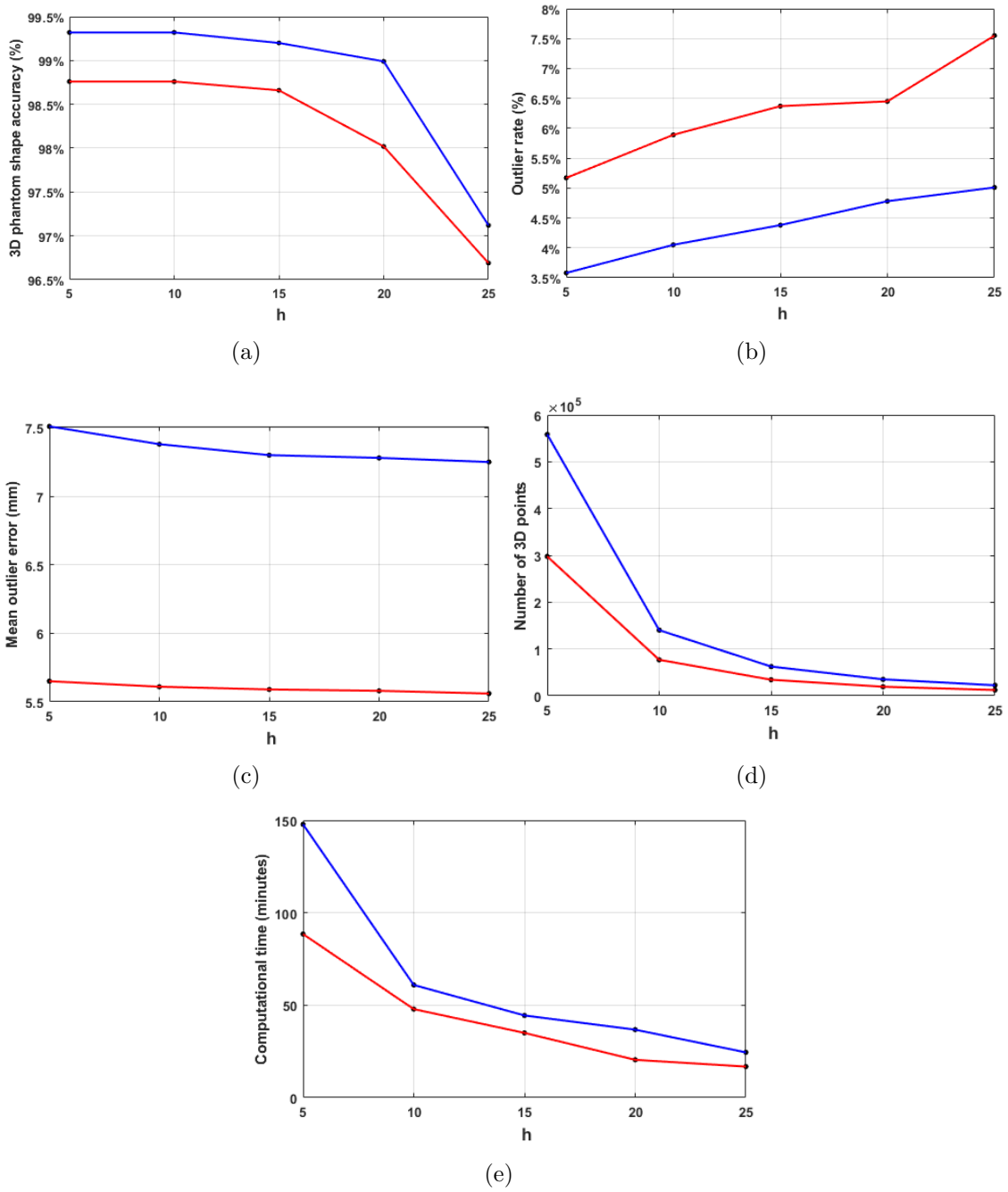


Figure 3.10: Impact of the grid cell parameter values on the reconstruction results. Blue and red curves are given for the skin and stomach phantoms, respectively. (a) Relationship between the value of h and shape value p. (b) Outlier rate according to the value of h. (c) Mean outlier error with respect to h. (d) Number of 3D points according to h. (e) Effect of the value of h on the computation time.

types. This observation indicates that the accuracy of the proposed method is quite stable with respect to h. The number of reconstructed 3D points quickly decreases with increasing values of h. The computation time quickly decreases from $h = 5$ to $h = 10$, while the decrease of this criterion is more slow from $h = 10$ to $h = 25$. Small values of the grid cell size (e.g., $h = [5, 15]$) lead to dense 3D point clouds so

that the proposed SfM incremental pipeline does not require a MVS step.

3.5.2.2 Impact of parameter ϵ on the reconstruction results

Table 3.2 and Fig. 3.11 show experimental results with $h = 10$, $\tau = \frac{2WH}{3}$, and for ϵ which varies in $\{0.06, 0.08, 0.1, 0.3, 0.5, 0.7, 0.9\}$. The first two parameters were kept constant to see the impact of the OF threshold parameter (which adjusts the homologous point accuracy) on the 3D reconstruction accuracy. Cell grid size h was set to 10 to have a compromise between the amount of 3D points and the computation time (see Table 3.1), while the chosen τ-image overlap threshold still ensures the selection of image pairs with a significant overlap. The curves in Figs. 3.11(a), 3.11(b), and 3.11(c) are nearly horizontal when $\epsilon \geq 0.1$. This means that the accuracy of the flow fields estimated by the proposed OF scheme is constantly high for $\epsilon \geq 0.1$. Notably, the shape value (percentage p) remains close to 100% for this ϵ values, while the difference in terms of outlier rates and mean errors between the two phantoms is probably due to the texture differences (see Figs. 3.9(b)-(c)). However, the 3D reconstruction remains globally accurate for all tested $\epsilon \geq 0.1$. For $\epsilon \leq 0.1$ (very high OF accuracy is enforced), the number of 3D point amount and the shape accuracy value decrease, while the outlier rate becomes higher compared to those with $\epsilon \geq 0.1$. For $\epsilon \geq 0.1$, the changes of the number of 3D points and computation times are negligible. Therefore, we can choose ϵ values in $[0.1, 0.9]$. That is why in our experiments, ϵ is set to 0.1.

Table 3.2: Experimental results with $\epsilon = \{0.06, 0.08, 0.1, 0.3, 0.5, 0.7, 0.9\}$, $\tau = \frac{2WH}{3}$, and $h = 10$.

Phantom type	ϵ	3D phantom shape accuracy (p in %)	Outlier rate (%)	Mean outlier error (mm)	Number of 3D points	Computation time (Minutes)
Skin surface phantom	$\epsilon = 0.06$	97.24	4.94	7.25	79819	52.17
	$\epsilon = 0.08$	99.24	4.01	7.31	110204	57.34
	$\epsilon = 0.1$	99.32	4.05	7.38	140204	60.94
	$\epsilon = 0.3$	99.32	4.06	7.38	148601	61.29
	$\epsilon = 0.5$	99.32	4.07	7.39	154219	62
	$\epsilon = 0.7$	99.32	4.07	7.40	159289	62.27
	$\epsilon = 0.9$	99.32	4.08	7.41	162864	63.01
Stomach surface phantom	$\epsilon = 0.06$	96.83	6.82	5.50	41707	39.07
	$\epsilon = 0.08$	98.61	5.86	5.57	60121	44.29
	$\epsilon = 0.1$	98.76	5.89	5.61	76808	47.88
	$\epsilon = 0.3$	98.76	5.89	5.62	77917	48.65
	$\epsilon = 0.5$	98.76	5.90	5.63	78707	48.79
	$\epsilon = 0.7$	98.76	5.90	5.64	79878	49.11
	$\epsilon = 0.9$	98.76	9.91	5.64	81432	49.83

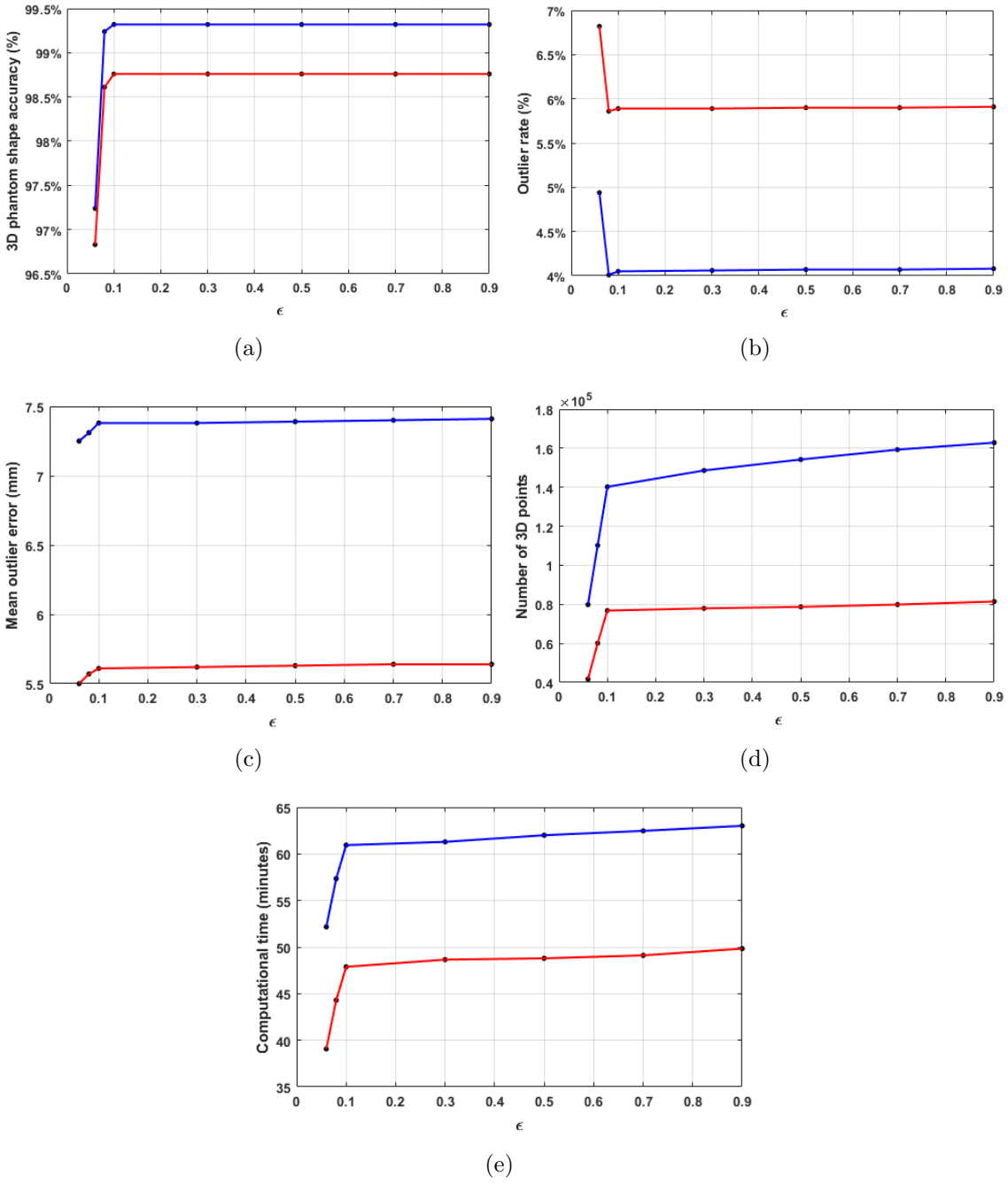


Figure 3.11: Impact of the OF threshold accuracy values on the reconstruction results. Blue and red curves are given for the skin and stomach phantoms, respectively. (a) Relationship between the value of ϵ and shape value p. (b) Outlier rate according to the value of ϵ. (c) Mean outlier error with respect to ϵ. (d) Number of 3D points according to ϵ. (e) Effect of the value of ϵ on the computation time.

3.5.2.3 Impact of parameter τ on the reconstruction results

Experimental results were computed for $\tau = \frac{WH}{4}$, $\frac{11WH}{24}$, $\frac{2WH}{3}$, and $\frac{7WH}{8}$ to see the effects of the image overlap threshold on the reconstruction results. $\epsilon = 0.1$ and $h = 10$ ensure accurate OF flow fields and a compromise between the amount of 3D points and computation time, respectively. Fig. 3.12 provides plots showing the

Table 3.3: Experimental results with $\tau = \left\{\frac{WH}{4}, \frac{11WH}{24}, \frac{2WH}{3}, \frac{7WH}{8}\right\}$, $\epsilon = 0.1$, and $h = 10$.

Phantom type	τ	3D phantom shape accuracy (p in %)	Outlier rate (%)	Mean outlier error (mm)	Number of 3D points	Computation time (Minutes)
Skin surface phantom	$\tau = (1/4)\,WH$	93.78	4.72	7.34	31568	33.91
	$\tau = (11/24)\,WH$	95.20	4.90	7.36	63236	45.09
	$\tau = (2/3)\,WH$	99.32	4.05	7.38	140204	60.94
	$\tau = (7/8)\,WH$	91.17	3.82	7.64	440204	136.19
Stomach surface phantom	$\tau = (1/4)\,WH$	92.42	5.86	5.52	21620	22.67
	$\tau = (11/24)\,WH$	93.18	6.03	5.57	39716	37.08
	$\tau = (2/3)\,WH$	98.76	5.89	5.61	76808	47.88
	$\tau = (7/8)\,WH$	90.07	4.55	6.69	327632	98.79

variation of the quality criteria according to τ. Fig. 3.12(a) clearly shows that the highest shape accuracy is obtained with $\tau = \frac{2WH}{3}$. The variations of the outlier rate and mean outlier rate are not significant when the overlap value changes (see Table 3.3 and Figs. 3.12(b)-3.12(c)). The number of 3D points and computation time increase with an increasing τ. In order to find a trade-off between the accuracy, the number of reconstructed 3D points and computation time, τ was set to $\frac{2WH}{3}$ in all our experiments presented in this work.

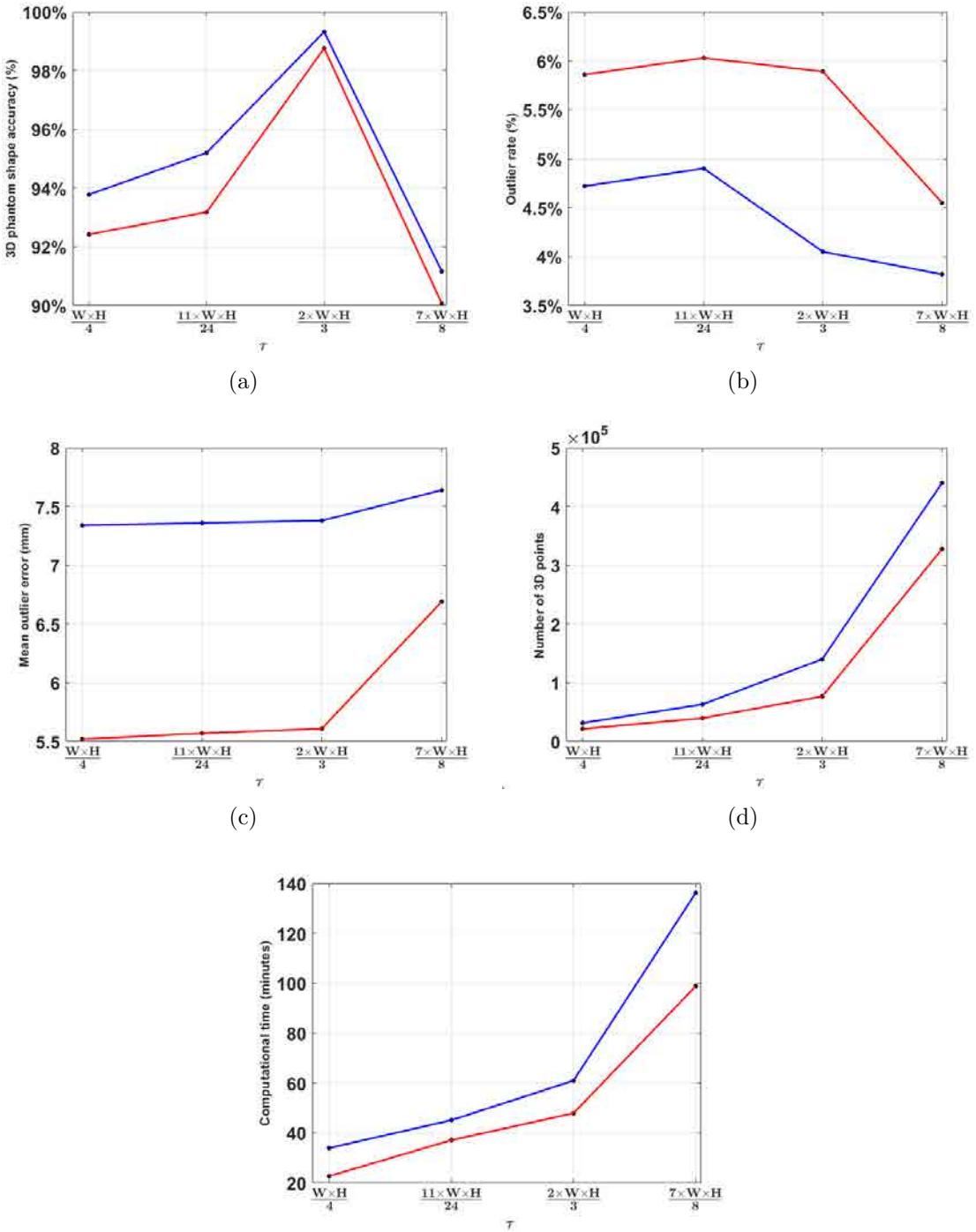
By considering globally all results of these experiments one can propose a set of values for triplet (h, ϵ, τ) that represent a compromise between the amount of reconstructed 3D points, 3D point accuracy and computation time:

$$\{h, \epsilon, \tau\} = \left\{10, 0.1, \frac{2WH}{3}\right\}.$$

The optimal values of all parameters (including the OF parameters in Subsection 3.5.1) are summarized in Table 3.4. HP-groups obtained by the proposed point grouping method (Section 3.4) with the parameters setting of Table 3.4 are the input of the second stage of SfM which is presented in Chapter 2 and which produces both 3D point clouds and camera poses. The SfM pipeline in Fig. 3.13 and the parameters values in Table 3.4 were systematically used in this thesis for the presented results.

	Optical flow				Point grouping		
Parameter	λ	Py_s	γ_1	γ_2	τ	h	ϵ
Adjusted value	9	0.7	3	5	$\frac{2WH}{3}$	10	0.1

Table 3.4: Constant parameter values used for all experiments with the proposed DOF-based SfM method.



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

Figure 3.12: Impact of the image overlap threshold values on the reconstruction results. Blue and red curves are given for the skin and stomach phantoms, respectively. (a) Relationship between the value of τ and shape value p. (b) Outlier rate according to the value of τ. (c) Mean outlier error with respect to τ. (d) Number of 3D points according to τ. (e) Effect of the value of τ on the computation time.

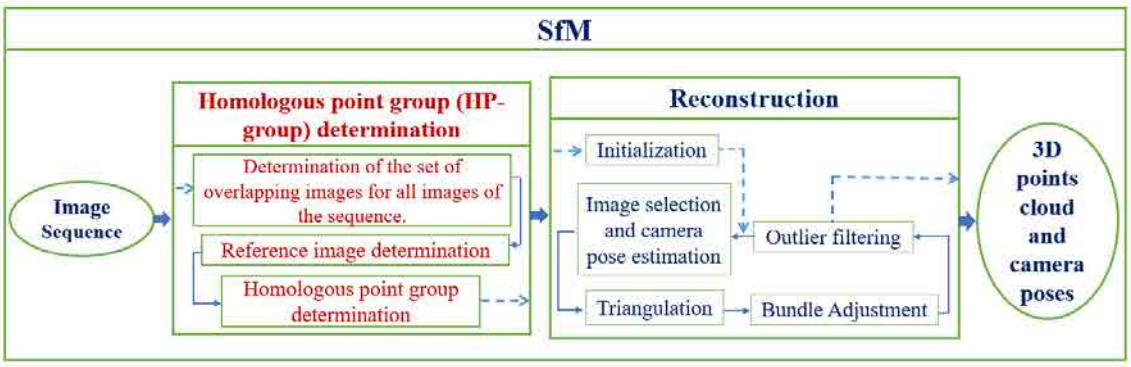
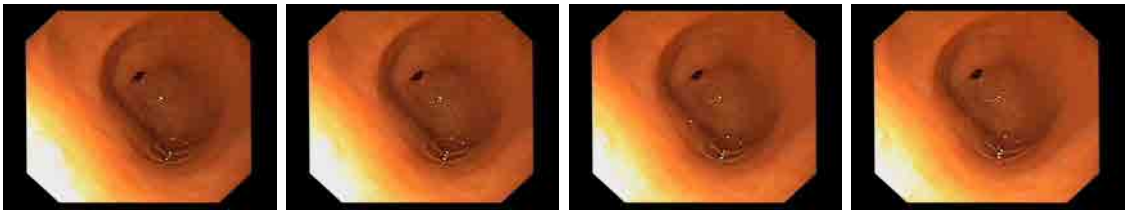


Figure 3.13: DOF-based SfM pipeline used in this thesis.

3.6 Robustness of the DOF based SfM scheme

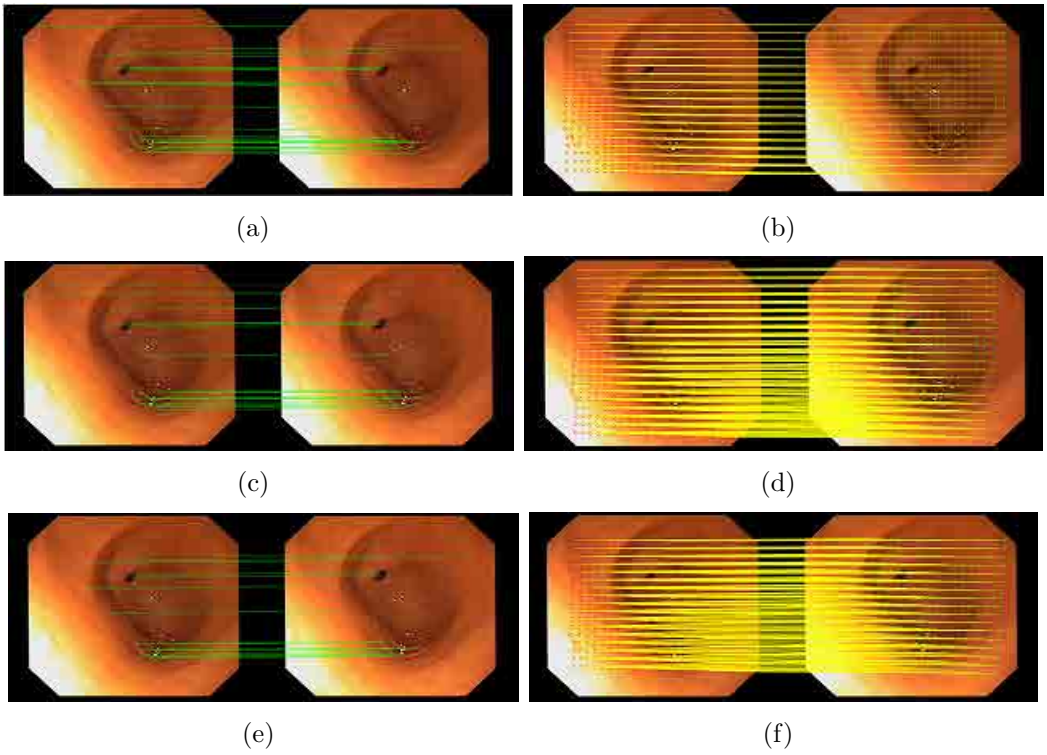
Four gastroscopic images are used to illustrate the importance of a robust homologous point group determination in the frame of SfM applied to medical scenes. The small displacement between the four gastroscopic images in Fig. 3.14 facilitates both the matching step in feature based approaches and the point correspondence determination in OF schemes. Thus, the difficulty in finding homologous points in these images mainly relates to the quality of the iconic information available for finding corresponding points.

In our experiments, the second image was manually and arbitrary chosen as reference image: reference $\mathbf{I}_2^{ref}$ is associated to set $S_2 = \{\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \mathbf{I}_4\}$. However, since the images are close in terms of positions in the scene, choosing another image as reference will not change the results given below. The DOF and SIFT matches were determined for images pairs ($\mathbf{I}_2^{ref}$, $\mathbf{I}_1$), ($\mathbf{I}_2^{ref}$, $\mathbf{I}_1$) and ($\mathbf{I}_3^{ref}$, $\mathbf{I}_4$). It can be seen in Table 3.5, as well as in Figs. 3.15(a), 3.15(c), and 3.15(e) that only few tens of homologous pixels were found for each image pair with a SIFT based approach [VF10] as used in most SfM methods as [SF16, NSR06, SSH+15]. Table 3.5 also shows that more than 400 point correspondences were established for each image pair using the DOF approach. To facilitate the visualization of homologous points, Fig. 3.15 only visualize the OF matches for a grid cell size of $h \times h = 20 \times 20$ (instead showing all matches of a grid $h \times h = 10 \times 10$). This grid size, together with the size of



(a) Image $\mathbf{I}_1$ (b) Image $\mathbf{I}_2^{ref}$ (c) Image $\mathbf{I}_3$ (d) Image $\mathbf{I}_4$

Figure 3.14: Gastroscopic images acquired from four viewpoints and used to show the impact of the homologous point grouping efficiency on the robustness of a SfM method. The image size $H \times W$ is 640×482 pixels.



(a) (b) (c) (d) (e) (f)

Figure 3.15: Homologous points determined for three image pairs with the SIFT features (left column) and the DOF vectors (right column). The first, second and third rows correspond to pairs $(\mathbf{I}_2^{ref}, \mathbf{I}_1)$, $(\mathbf{I}_2^{ref}, \mathbf{I}_3)$ and $(\mathbf{I}_2^{ref}, \mathbf{I}_4)$, respectively

the images (640×482 pixels), can maximally lead to 440 correspondences when all OF vectors are precisely determined. With $\epsilon = 0.1$ in Eq. (3.4.3), more than 400 correspondences were considered as very accurate, even for the endoscopic images including few textures and structures (see Figs. 3.15(b), 3.15(d), and 3.15(f)). Thus, the rate (at least 400 matches among 440 possible vectors) of accurate correspondences highlights the robustness of the proposed DOF determination.

Due to their systematic link with reference image $\mathbf{I}_2^{ref}$, the corresponding points between the three images pairs in Table 3.5 can be easily used to group homologous points in triplets (homologous points simultaneously viewed in three images) and in quadruplets (homologous points appearing in four images). The number of these homologous point groups are given in Table 3.6.

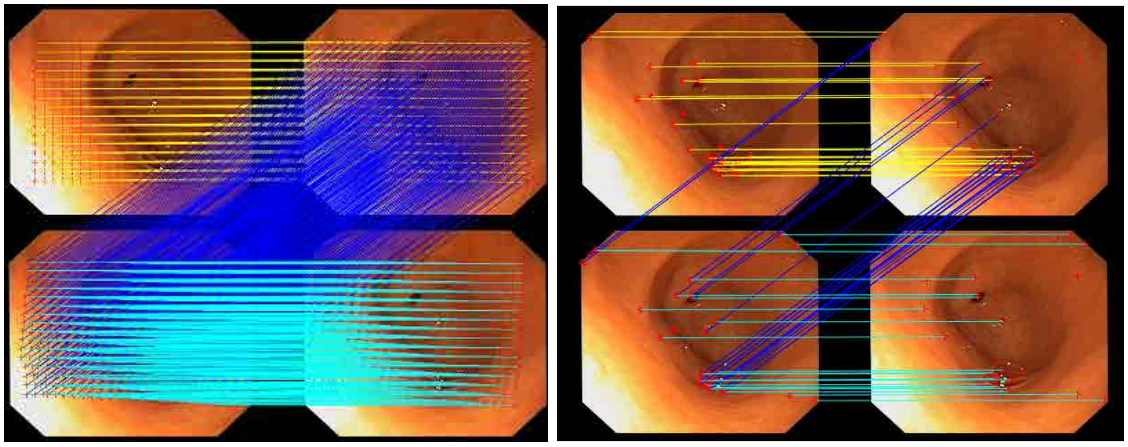
As seen in Table 3.6 and Fig. 3.16(a), numerous point triplets and 368 point quadruplets were obtained with the DOF fields. On the contrary, a few homologous point triplets and only 7 quadruplets were obtained with SIFT features. This en-

Table 3.5: Number of correspondences determined between each image pair using SIFT features and DOF.

Image pair	$(\mathbf{I}_2^{ref}, \mathbf{I}_1)$	$(\mathbf{I}_2^{ref}, \mathbf{I}_3)$	$(\mathbf{I}_2^{ref}, \mathbf{I}_4)$
SIFT matches	33	23	18
OF correspondences	410	406	404

Table 3.6: Number of homologous point triplets and quadruplets obtained with the SIFT features and DOF matches in Table 3.5.

Point groups	$\mathbf{I}_1, \mathbf{I}_2^{ref}, \mathbf{I}_3$	$\mathbf{I}_2^{ref}, \mathbf{I}_3, \mathbf{I}_4$	$\mathbf{I}_2^{ref}, \mathbf{I}_3, \mathbf{I}_4$	$\mathbf{I}_1, \mathbf{I}_2^{ref}, \mathbf{I}_3, \mathbf{I}_4$
OF	400 triplets	389 triplets	372 triplets	368 quadruplets
SIFT	12 triplets	10 triplets	14 triplets	7 quadruplets



(a) (b)

Figure 3.16: Visualization of the HP-groups obtained with the SIFT and DOF methods. (a) 368 quadruplets of homologous points achieved by the DOF method are linked by yellow, dark blue and light blue line segments while (b) only 7 quadruplets of homologous points were obtained by the SIFT approach.

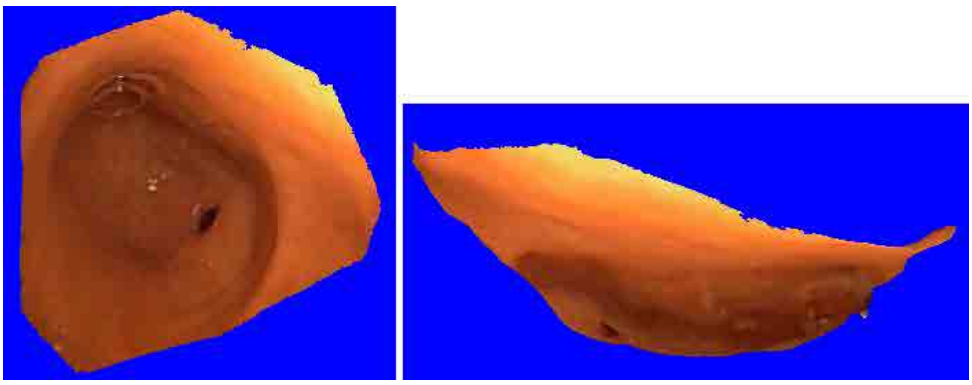


Figure 3.17: Two viewpoints of the pyloric antrum region surface reconstructed with the DOF-based SfM method. This surface was obtained with the four gastroscopic images in Fig. 3.14.

doscopic scene example explains why state-of-art methods (and software) based on features extraction [SF16, SSH+15, MMMO, Wu13, NSR06] are often inoperative in very extreme acquisition and scene conditions (scenes including almost no textures and structures and acquired under strongly changing illumination).

Both the proposed DOF and the COLMAP SfM methods were applied to the four gastroscopic images in Fig. 3.14. It can be seen in Fig. 3.17 that the DOF-

based approach led to a realistic pyloric antrum shape, while COLMAP was unable to reconstruct a surface due to a too limited number of point correspondences. The potential of the described HP-grouping method for different scenes is discussed on numerous 3D reconstruction results given in Chapter 5.

3.7 Main contributions and conclusion

The efficiency of a SfM method strongly depends on the quality of the homologous image point determination. A robust optical flow method giving a dense point correspondence between image pairs and an original strategy for finding numerous images showing common scene parts lead to a new SfM approach for scenes with few textures and images acquired under strong illumination changes. The main contributions of this chapter can be enumerated as follows:

1. High accuracy optical flow estimation using a new illumination-invariant descriptor and determination of large HP-groups using DOF fields.
2. The DOF field is only computed between two τ-overlapped images such that accumulated errors along images sequences (as arising in classical 2D mosaicing of the literature [ADGB16, TDBL18]) are avoided.
3. Novel optical flow-based SfM for scenes with few textures and dense 3D point cloud generation without any MVS step.
4. An overview video of the proposed algorithm and the MATLAB code for homologous point grouping can be seen at:

   https://github.com/CRAN-BioSiS-Imaging/PR2020.

List of publications

International journal

- T.-B. Phan, D.-H.Trinh, D. Wolf, C. Daul. Optical flow-based structure-from-motion for the reconstruction of epithelial surfaces. *Pattern Recognition,* Volume 105, 107391, September 2020. [PTWD20]

International conferences

- T.-B. Phan, D.-H. Trinh, D. Lamarque, D. Wolf, and C. Daul. Dense optical flow for the reconstruction of weakly textured and structured surfaces: Application to endoscopy. IEEE International Conference on Image Processing *(ICIP)*, pages 310-314, Taipei, Taiwan, 2019. [PTL+19b]

The next chapter will present the determination of HP-groups based on DOF combined to the feature matching method.

Chapter 4

SfM based on the feature matching method combined to DOF

Contents

This chapter proposes an approach which combines a feature matching method and the DOF matching technique for generating groups of numerous homologous points. Contrary to Chapter 3 which describes a point grouping method designed for scenes with almost no textures, this chapter shows how a feature based grouping method can be improved using DOF fields when textures are partially available (i.e. when texture or structure information are missing in some images or images regions).

4.1 Introduction

Since the beginning of SfM, and for more than a decade, feature detection and matching methods were overwhelmingly integrated in numerous SfM-schemes [SF16, MMMO, GFF10, Wu13]. These methods gained impressive results in various application fields as geoscience [JR12], the reconstruction of large urban scenes [COSH13], consumer photography [NSR06], and monument modeling [GFF10]. Feature points detected by those methods are located with sub-pixel accuracy, while their feature descriptors can be invariant to scale, rotation and intensity changes. Homologous image points are classically obtained by matching feature points using their descriptor vectors. Outliers are rejected by the RANSAC method [FB81] which takes a

homography as geometric transformation model between image pairs. An introduction to feature matching methods was given in Chapter 2.

The works in [SPS12, LAZ+17] show the potential of SfM based feature matching methods (SIFT) in cystoscopy. However, even in white light cystoscopy where textures are available in numerous images (mainly due to the blood vessels into the bladder epithelium), images can also be with few contrasted textures (e.g., due to motion blur or defocussing of the endoscope's camera) or significant image region parts can be without textures (e.g., due to scars after a surgical intervention, or radiotherapy). Important viewpoint changes can also modify the aspect of the textures and make feature-based matching more difficult. Moreover, even in the presence of contrasted textures, the available information can lead to a limited number of homologous point matches. In all these situations DOF fields can be used to increase the number of correspondences in the homologous point groups.

Besides the increase of the homologous point number, the objective of the joint use of the two matching methods is to take advantage of the strengths of both methods. On the one hand, an OF approach is able to provide a dense correspondence in textureless image regions and, on the other hand, the accuracy of feature points is exploited whenever textures are available. The proposed SfM algorithm is referred to as FMDOF-SfM algorithm.

Among the different feature detection and matching methods (e.g., corner detection [HS88, ST94], SURF [BETV08], and KAZE [ABD12]), SIFT [Low04] is probably the method which was the most often integrated in SfM pipelines. In conjunction with SfM, SIFT has been proven as an effective method for the detection and matching of feature points in numerous scenes. Therefore, SIFT matches were combined with DOF-correspondences in this thesis. However, other feature matching methods can replace the SIFT approach in the algorithm described in this chapter.

4.2 Determination of homologous point groups

A point group is defined from at least three images $\mathbf{I}_i$, $\mathbf{I}_j$ and $\mathbf{I}_k$ taken all from different viewpoints and with $1 \leq i, j, k \leq N$ (the video-sequence consists of N images). If $\left(\mathbf{p}_i^{b^1}, \mathbf{p}_j^{b^2}\right)$ and $\left(\mathbf{p}_j^{b^2}, \mathbf{p}_k^{b^3}\right)$ are homologous point pairs in image pairs $(\mathbf{I}_i, \mathbf{I}_j)$ and $(\mathbf{I}_j, \mathbf{I}_k)$ respectively, then $\mathbf{p}_i^{b^1}, \mathbf{p}_j^{b^2}$ and $\mathbf{p}_k^{b^3}$ belong to group b with a minimal size of three points.

The main idea of proposed method is to search scene regions seen in as numerous images as possible and to use the DOF and/or feature matching methods to determine the homologous points between image pairs. SfM is accurate when the 3D points are reconstructed from numerous viewpoints. In scene regions with a large number of image overlaps, the point groups have the highest probability to be large. Since in these common scene regions the corresponding images have not to be registered (homologous points have only to be determined), a simple rectangle can be used to delineate the common parts of image pairs geometrically linked by a translation vector (vector $\mathbf{v}_{i,i+1}$ in Fig. 3.4 of Chapter 3). Next sections detail the

Algorithm 3 HP-groups determination based on the combination of DOF and feature matching methods

Input: Set S of N consecutive images $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N$, number $|K_i|$ of feature points of each image i and number $|M_{i,j}|$ of point matches of all image pair $(\mathbf{I}_i, \mathbf{I}_j)$, where $1 \leq i, j \leq N$. α and β are the thresholds for the number of feature points and matching points between two images, respectively.

/* ***Part 1***: *Determination of image translations:* */
for each $\mathbf{I}_i$ with $i \in [1, N-1]$ **do**
  if $|K_i| > \alpha$, $|K_{i+1}| > \alpha$ and $|M_{i,i+1}| > \beta$ **then**
    $\mathbf{v}_{i,i+1} = \left(t_x^{i,i+1}, t_y^{i,i+1}\right)$ in equation Eq. (4.2.1).
  else
    $\mathbf{v}_{i,i+1} = \mathbf{F}_{i,i+1}\left(O_{\mathbf{I}_i}\right)$, where $O_{\mathbf{I}_i}$ is the central pixel of image $\mathbf{I}_i$.
  end if
end for
for each pair $(\mathbf{I}_i, \mathbf{I}_j)$ with $j \neq i \pm 1$ **do**
  $\mathbf{v}_{i,j}(v_{i,j}^1, v_{i,j}^2) = \sum_{t=i}^{j-1} \mathbf{v}_{t,t+1}(v_{t,t+1}^1, v_{t,t+1}^2)$.
end for

/* ***Part 2***: *Determination of reference images:* */
Set Ω^{ref} of images $\mathbf{I}_i^{ref}$ and their groups S_i.
(Determined by the Algorithm 2 in Section 3.3 of Chapter 3)

/* ***Part 3***: *Point groups determination:* */
for each $\mathbf{I}_i^{ref}$ **do**
  for each $\mathbf{I}_j \in S_i$ **do**
    if $|K_i| > \alpha$, $|K_j| > \alpha$ and $|M_{i,j}| > \beta$ **then**
      $M_{i,j}$ is the set of matched feature points between $\mathbf{I}_i^{ref}$ and $\mathbf{I}_j$.
    else if $|K_i| > \alpha$ and $|K_j| \leq \alpha$ **then**
      for each $\mathbf{p}_a^{i,ref} \in K_i$ **do**
        $\mathbf{p}_a^j = \mathbf{p}_a^{i,ref} + \mathbf{F}_{i,j}\left(\mathbf{p}_a^{i,ref}\right)$
        if $\mathbf{p}_a^j$ fulfill the SR and accuracy criteria (see Section 3.4) **then**
          Keep $\left(\mathbf{p}_a^{i,ref}, \mathbf{p}_a^j\right)$ as a homologous point pair.
        else
          Reject pair $\left(\mathbf{p}_a^{i,ref}, \mathbf{p}_a^j\right)$.
        end if
      end for
    else
      Homologous points are determined by the DOF fields as in Section 3.4.
    end if
  end for
end for
Output: Set of HP-groups.

three steps of the method given in Algorithm 3.

1. ***Step 1:*** *Determination of image translations.* Two matching methods are considered to find the translation vectors $\mathbf{v}_{i,i+1}$ between all consecutive images $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$ of a sequence. Since the feature-based methods have the highest accuracy and the shortest computation time, it is first checked whether the SIFT algorithm can be used or not to find the translation between $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$. When this attempt with SIFT fails, a DOF method is used to determine $\mathbf{v}_{i,i+1}$ in a robust way.

   Let K_i ($i \in [1, \ldots, N]$) be the set of $|K_i|$ feature points detected in $\mathbf{I}_i$ by the SIFT algorithm [Low04]. The feature points of sets K_i and K_{i+1} are matched using their descriptor vectors, and by rejecting the outliers with the RANSAC method [FB81] taking homography $\mathbf{H}_{i,i+1}$ as transformation model between images $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$. A homography geometrically links the homologous pixels of two images under the assumption that the small FoV images visualize quasi-planar surfaces. The homography matrix between two consecutive images $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$ can be written as follows:

$$\mathbf{H}_{i,i+1} = \begin{pmatrix} f\cos\phi & -s_x \sin\phi & t_x^{i,i+1} \\ s_y \sin\phi & f\cos\phi & t_y^{i,i+1} \\ h_1 & h_2 & 1 \end{pmatrix}, \tag{4.2.1}$$

   where f, ϕ, (s_x, s_y), $\left(t_x^{i,i+1}, t_y^{i,i+1}\right)$ and (h_1, h_2) denote the scale factor, the in-plane rotation, the shearing factors, the 2D translation, and the perspective changes, respectively [ABD13].

   $M^{i,i+1}$ corresponds to the set of $|M^{i,i+1}|$ point pairs which were successfully matched. The feature-based matching is considered as valid under two conditions: (i) the number of detected features must be above a threshold α for images $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$ (i.e., $|K_i|$ and $|K_{i+1}| > \alpha$) and (ii) the number of matches $|M^{i,i+1}|$ must be larger than threshold β. If these two conditions are fulfilled, the components $(v_{i,i+1}^1, v_{i,i+1}^2)$ of vector $\mathbf{v}_{i,i+1}$ take the value of the translation parameters located in the last column of the homography matrix taken as model in RANSAC; in other words $\mathbf{v}_{i,i+1} = \left(t_x^{i,i+1}, t_y^{i,i+1}\right)$.

   The DOF from $\mathbf{I}_i$ to $\mathbf{I}_{i+1}$ is computed when at least one of the two previous conditions is not fulfilled. This vector field between consecutive images (denoted by $\mathbf{F}_{i,i+1}$) is computed with a robust variational method developed for scenes with few textures and affected by strong illumination changes (see Subsection 3.2.2 in Chapter 3). The central vector of flow field $\mathbf{F}_{i,i+1}$ of $\mathbf{I}_i$ is taken as translation $\mathbf{v}_{i,i+1}$ which is estimated according to Eq. (3.3.1) in Section 3.3.1.

   The translation vectors between two non-consecutive images $\mathbf{I}_i$ and $\mathbf{I}_j$ (with $j \neq i \pm 1$) are defined by the sum of the vectors between the consecutive images from i to j as described in Eq. (3.3.3).

2. ***Step 2:*** *Determination of reference images favouring large point groups.* Reference images are images that are $\tau-$overlapped with as much as possible other images. The algorithm used to determine these reference images is detailed in Section 3.3 of Chapter 3. The reference images $\mathbf{I}_i^{ref}$ and their sets S_i of

$\tau-$overlapped images are used in step 3 to determine HP-groups (homologous point groups).

3. ***Step 3:*** *Point group determination.* Point groups are computed for each reference image $\mathbf{I}_i^{ref}$ by determining the homologous points for all pairs $(\mathbf{I}_i^{ref}, \mathbf{I}_j)$, with $\mathbf{I}_j \in S_i$. According to the SIFT algorithm efficiency criteria defined in step 1, one among three methods is used to optimize the accuracy and robustness of the homologous point determination between $\mathbf{I}_i^{ref}$ and $\mathbf{I}_j$:

   - If enough SIFT points are detected in both images ($|K_i|$ and $|K_j| > \alpha$) and successfully matched ($|M^{i,j}| > \beta$), then the homologous points are computed with SIFT and RANSAC. In this case, the homologous points of image pair $(\mathbf{I}_i, \mathbf{I}_j)$ are only determined using feature information.
   - If enough SIFT points are detected in the reference $\mathbf{I}_i^{ref}$, but not enough SIFT points were found in $\mathbf{I}_j$ ($|K_j| \leq \alpha$) or the matching failed ($|M^{i,j}| \leq \beta$), then for each feature point $\mathbf{p}_a^{i,ref} \in K_i$, the point $\mathbf{p}_a^j \in \mathbf{I}_j$ defined by $\mathbf{p}_a^j = \mathbf{p}_a^{i,ref} + \mathbf{F}_{i,j}\left(\mathbf{p}_a^{i,ref}\right)$, is the homomologous of $\mathbf{p}_a^{i,ref}$ when it is preserved by specular reflections and occlusions in $\mathbf{I}_j$ (see homologous points determination in Section 3.4 of Chapter 3). In this case, the homologous points of image pair $(\mathbf{I}_i, \mathbf{I}_j)$ are determined by jointly using feature points and optical flow field $\mathbf{F}_{i,j}$.
   - When not enough SIFT points can be found in $\mathbf{I}_i^{ref}$, the homologous point search is completely based on the flow field $\mathbf{F}_{i,j}$ from $\mathbf{I}_i^{ref}$ to $\mathbf{I}_j$. A grid $\mathbf{C}_i^{ref}$ of 2D points in $\mathbf{I}_i^{ref}$ is created, $h \times h$ being the square cell grid size:

     $$\mathbf{C}_i^{ref} = \{\mathbf{p}_{xy}^{i,ref}(xh, yh) \mid x, y \in \mathbb{N}, x \leq \frac{W}{h}, y \leq \frac{H}{h}\}. \tag{4.2.2}$$

     Each $\mathbf{p}_{xy}^j \in \mathbf{I}_j$, defined by $\mathbf{p}_{xy}^j = \mathbf{p}_{xy}^{i,ref} + \mathbf{F}_{i,j}\left(\mathbf{p}_{xy}^{i,ref}\right)$, is a homologous point of $\mathbf{p}_{xy}^{i,ref}$ in $\mathbf{I}_i^{ref}$ if it satisfies the specular reflection, occlusion, and accuracy correspondence conditions which are given in Section 3.4 (see Eq. (3.4.3)).

It is noticeable that Algorithm 3 becomes the DOF algorithm proposed in Chapter 3 when all reference images have $|K_i| \leq \alpha$.

4.3 Parameter values for the FMDOF HP-group computation

It is a challenging task to find the best (constant) values of pair (α, β) since the optimal threshold settings depend on the features which can effectively be extracted from different scene types. The feature extraction depends on the scene content (on the organ type and its epithelial tissue), on the image modality (white light, narrow band imaging or fluorescence endoscopic modality), and on the data quality (motion blur, reflections, etc.). The aim is to determine robustly as much as possible feature matches to take advantage of the sub-pixel accuracy of such point correspondences,

while optical flow matches should provide complementary information ensuring a dense correspondence to avoid the use of an additional MVS step.

As done in Subsection 3.5.2 for the parameter adjustment of the homologous point grouping based only on DOF fields, the efficiency of the FMDOF-based HP-group determination is assessed with criteria relating to the quality of the surface construction algorithm which exploits the correspondences. Thus, the FMDOF-based SfM scheme is tested with different combinations of α and β values using the two cylindrical phantoms with known dimensions as seen in Fig. 4.1 (these cylinder phantoms have the same geometry as those given in Fig. 3.9 of Subsection 3.5.2). The cylinders are covered with paper sheets on which skin or bladder images were printed. In these tests, the SIFT algorithm can find a significant number of homologous point correspondences since, due to the printing process, the epithelial textures are much more contrasted than those visible directly in the images. A video-sequence of 621 images and another of 293 images were acquired for the skin and bladder phantom, respectively. All images have a size of $H \times W = 780 \times 580$ pixels.

Phantom reconstruction results are given for numerous combinations of α and β threshold values and the performance of the 3D point reconstruction is assessed with the five quality criteria already used in Subsection 3.5.2:

- 3D phantom shape accuracy (percentage p, see Eq. (3.5.4)),
- the outlier rate (in %),
- the mean outlier error (in mm),
- the number of 3D points, and
- the computation time.

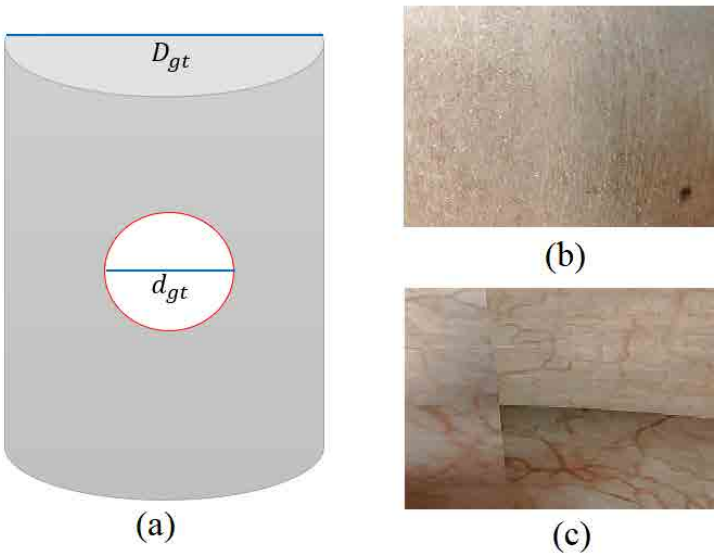


Figure 4.1: Description of the two phantoms used to adjust the point grouping parameters of the FMDOF method. (a) Geometry and dimensions of the phantoms. (b) Image of skin textures. (c) Image of bladder textures in the white light modality.

However, an additional criterion to be taken in account in these tests is the number of images for which SIFT information was used (i.e., the number of images for which SIFT and DOF information was jointly used added to the number of images for which SIFT extraction and matching alone led to a dense correspondence). The aim is to choose the parameter values leading to a high 3D phantom shape accuracy and for which the number of 3D points is as high as possible, while the outlier rate, mean outlier error, and computation time remain as low as possible.

The optimal values of the HP-group parameters τ, ϵ and h, as well as the DOF field parameter $(\lambda, Py_s, \gamma_1, \gamma_2)$ values used in the FMDOF algorithm remain those tuned in Chapter 3 (see Table 3.4 of Subsection 3.5.2).

In the first phantom reconstruction tests, a set of feature point threshold values was considered ($\alpha = \left\{\frac{WH}{3000}, \frac{WH}{1500}, \frac{WH}{1000}, \frac{WH}{750}, \frac{WH}{600}\right\}$) while β was kept constant. For each value of an α-threshold, it is assumed that a minimal number of feature points should be detected in average in a given image area. For instance, for $\alpha = \frac{WH}{1000}$, the amount of features is considered as sufficient then at least one feature point is detected in average in each square of approximativelly 32×32 pixels ($32 \times 32 \approx 31,62 \times 31,62 = 1000$). The chosen α-values cover a large interval of image size/square area surface ratios and allow to understand the impact of this parameter on the reconstruction performances.

The main issue making the feature matching methods often inoperative for endoscopic scenes relates either to the high number of wrong matches or to the weak number of accurately matched point pairs. Often too few point pairs can be accurately matched, even if numerous feature points were detected in two images. Therefore, the value of the threshold for the number of matched feature point pairs (β-value) must be defined with respect to the threshold of the detected feature point number (α-value). The successful matching of about the half of the detected feature points represents a compromise ensuring both a high number of accurately matched points for a selected image pair, and a reasonable chance to exploit features in endoscopic examinations like cystoscopy. It means that the value of β should be close to $\frac{\alpha}{2}$. For that reason, for the values of α used in the experiment shown in Table 4.1 and Fig 4.2, the β-value must be approximately equal to the half of the average value of the α values.

Let α_{avg} be the average value of the α-sequence given below:

$$\alpha = \left\{\frac{WH}{3000}, \frac{WH}{1500}, \frac{WH}{1000}, \frac{WH}{750}, \frac{WH}{600}\right\}.$$

The value of α_{avg} given by:

$$\alpha_{avg} = \frac{\left(\frac{WH}{3000} + \frac{WH}{1500} + \frac{WH}{1000} + \frac{WH}{750} + \frac{WH}{600}\right)}{5} = \frac{WH}{1000}. \tag{4.3.1}$$

The value of β to be chosen in this experiment should be close to $\frac{\alpha_{avg}}{2} = \frac{WH}{2000}$. In fact, the value of β was set to $\frac{WH}{1750}$. With this setting, the value of β is that which is the closest in average to all $\frac{\alpha}{2}$ values. Such (α, β) pair values lead to the most constant trade-off for the number of detected feature points in the images and the successfully matched point pairs.

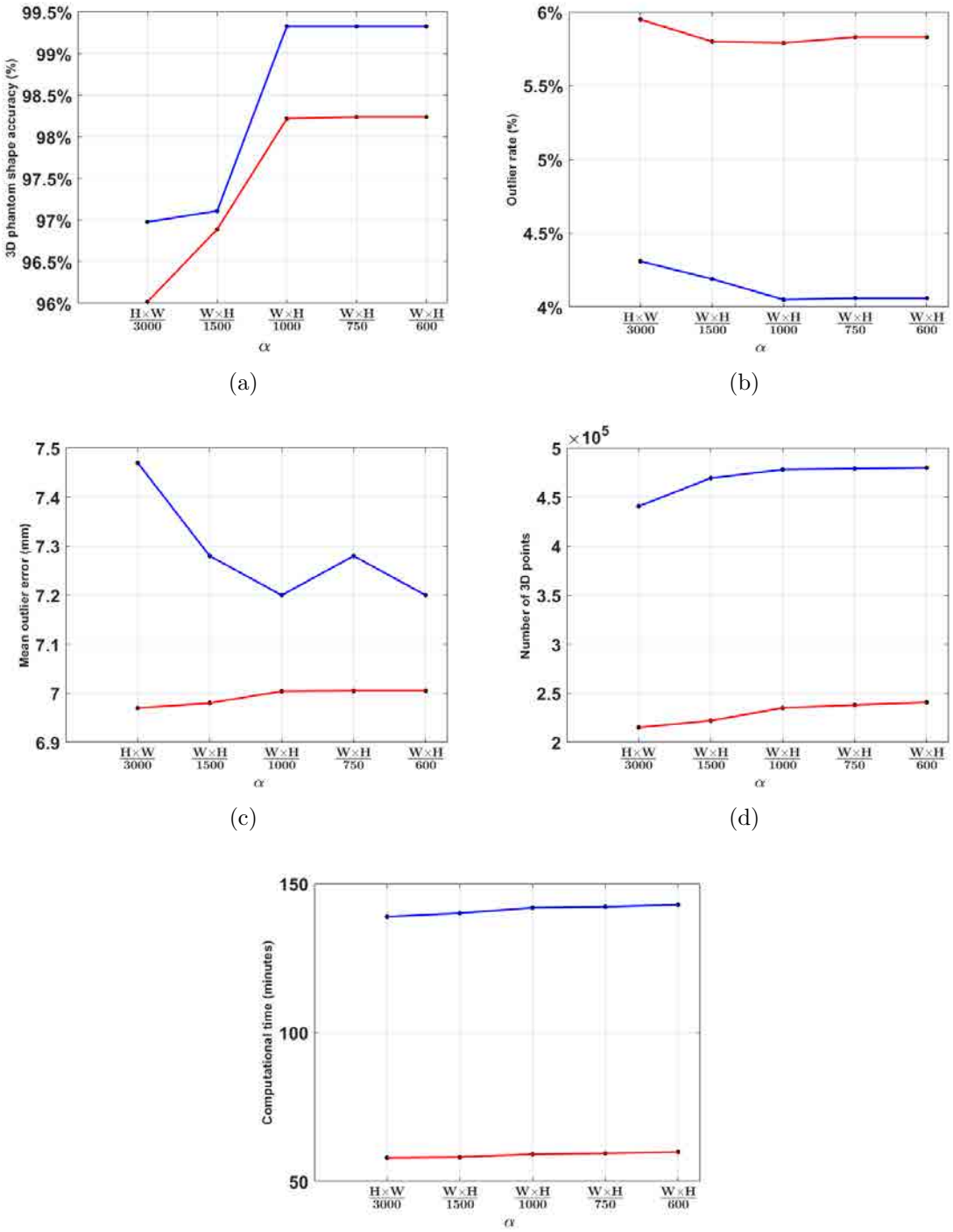
Phantom type	Fixed β, changing α thresholds	3D phantom shape accuracy (p in %)	Outlier rate (%)	Mean outlier error (mm)	Number of 3D points	Computation time (Minutes)	Number of images using SIFT
Skin surface phantom	$WH/3000$	96.98	4.31	7.47	440667	138.86	521
	$WH/1500$	97.11	4.19	7.28	469345	140.08	498
	$WH/1000$	99.33	4.05	7.2	478101	141.9	468
	$WH/750$	99.33	4.06	7.28	479061	142.21	413
	$WH/600$	99.33	4.06	7.2	479997	142.89	385
Bladder surface phantom	$WH/3000$	96.02	5.95	6.97	215031	57.92	284
	$WH/1500$	96.89	5.8	6.98	221987	58.13	241
	$WH/1000$	98.22	5.79	7.004	235117	59.08	200
	$WH/750$	98.24	5.83	7.005	238067	59.39	167
	$WH/600$	98.24	5.83	7.005	240813	59.87	141

Table 4.1: Experimental results obtained for the 621 images of the skin phantom and the 293 images of the bladder phantom. These results were determined with $\beta = \frac{WH}{1750}$ and $\alpha = \left\{\frac{WH}{3000}, \frac{WH}{1500}, \frac{WH}{1000}, \frac{WH}{750}, \frac{WH}{600}\right\}$.

An analysis of the results given in Table 4.1 and Fig. 4.2 shows that the global performances of the reconstruction are quite constant when α is greater or equal to $\frac{WH}{1000}$. Indeed, starting from this value, and for increasing α values, the number of 3D points, the outlier rate, the outlier mean value, the number of 3D points and the computation time are all almost constant for both the skin and the bladder phantoms. When α is lower than $\frac{WH}{1000}$, the global performances decrease (see the values of percentage p, the 3D phantom shape accuracy criterion). For $\alpha = \frac{WH}{1000}$ almost two thirds of the images (468 versus 621 for the skin phantom and 200 versus 293 for the bladder, see Table 4.1) use the SIFT method for the point matching. Therefore, in the proposed FMDOF algorithm, an α-value of $\frac{WH}{1000}$ can be chosen. With more matches based on feature values, the method is neither faster nor more precise. It is noticeable that, other threshold values of α which are greater than $\frac{WH}{1000}$ can be also selected to ensure similar performances.

Table 4.2 and Fig. 4.3 illustrate the influence of the matching threshold parameter (β) with the α-threshold set to $\frac{WH}{1000}$ and $\beta = \left\{\frac{WH}{2000}, \frac{WH}{1750}, \frac{9WH}{14000}, \frac{WH}{1400}\right\}$. The tested values of β are all greater than or equal to $\frac{\alpha}{2}$ in these experiments. With varying β-thesholds, the values of the outlier rate, the mean outlier error, the number of 3D points, and the computation time do not significantly change while the 3D shape accuracy (percentage p) increases a little bit when passing from $\beta = \frac{WH}{2000}$ to $\beta = \frac{WH}{1750}$. To preserve this high accuracy, and to obtain as much as possible images with SIFT matches, a $\beta-$value of $\frac{WH}{1750}$ can represent an appropriate choice for the matching threshold value.

These tests (and other tests made in this thesis for the FMDOF-SfM method) were done for $WH \geq 160000$ pixels (images with a minimal square area of 400×400 pixels). For this minimal image size, and based on the tests presented in this section,



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

Figure 4.2: 3D reconstruction performances according to the feature number threshold α. β was set at $\frac{WH}{1750}$. The blue and red curves give the results for the skin and the bladder phantoms, respectively. (a) Accuracy of the 3D shape. (b) Outlier rate. (c) Mean outlier error. (d) Number of recovered 3D points. (e) Computation time.

the threshold values were set to $\frac{WH}{1000}$ for the feature number α-parameter and to $\frac{WH}{1750}$ for the matched feature number. However, as seen in Tables 4.1 and 4.2, these parameter values are not critical since a constantly high performance can be obtained

with other pairs of the tested threshold values.

The results are given in Table. 4.3 confirms the efficiency of the proposed FMDOF-

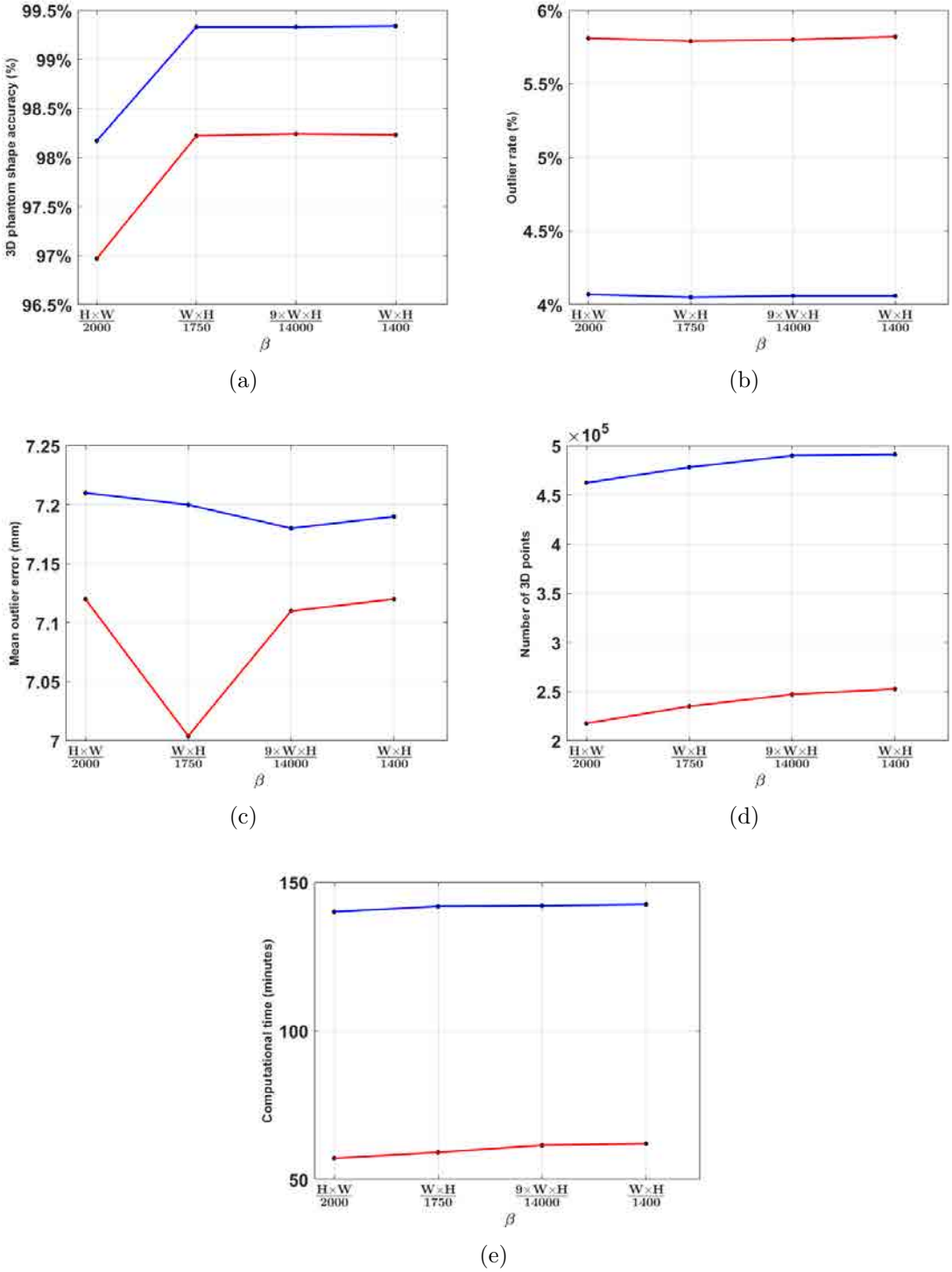


Figure 4.3: 3D reconstruction performances according to the matched feature number threshold β. Parameter α was set to $\frac{WH}{1750}$. The blue and red curves give the results for the skin and the bladder phantom surface, respectively. (a) Accuracy of the 3D shape. (b) Outlier rate. (c) Mean outlier error. (d) Number of recovered 3D points. (e) Computation time.

Phantom type	Fixed α, changing β thresholds	3D phantom shape accuracy (p in %)	Outlier rate (%)	Mean outlier error (mm)	Number of 3D points	Computation time (Minutes)	Number of images using SIFT
Skin surface phantom	$WH/2000$	98.17	4.07	7.21	462378	140.04	511
	$WH/1750$	99.33	4.05	7.2	478101	141.9	468
	$9WH/14000$	99.33	4.06	7.18	489985	142.08	427
	$WH/1400$	99.34	4.06	7.19	491204	142.55	381
Bladder surface phantom	$WH/2000$	96.97	5.81	7.12	217620	57.11	235
	$WH/1750$	98.22	5.79	7.004	235117	59.08	200
	$9WH/14000$	98.24	5.8	7.11	247051	61.47	185
	$WH/1400$	98.23	5.82	7.12	252698	62.02	151

Table 4.2: Reconstruction results for the 621 and 293 image sequences acquired for the skin and bladder phantoms, respectively. These experiments were conducted with $\alpha = \frac{WH}{1000}$, and $\beta = \left\{\frac{WH}{2000}, \frac{WH}{1750}, \frac{9WH}{14000}, \frac{WH}{1400}\right\}$.

	3D phantom shape accuracy (p in %)	Outlier rate (%)	Mean outlier error (mm)	Number of 3D points	Computation time (Minutes)
COLMAP+MVS	99.35	4.29	6.92	591126	150.5
DOF	99.32	3.58	7.5	558397	147.8
FMDOF	99.33	4.05	7.2	478101	141.9

Table 4.3: Comparison of the skin surface phantom reconstruction accuracy obtained for the FMDOF-based SfM (matches with both feature and OF information), the DOF-based SfM (only OF matches), and the COLMAP method (only feature matches) including a MVS step.

based SfM through a comparison with other SfM methods. The proposed method reconstructs the phantom surfaces with an accuracy that is similar both to that of the state-of-the-art COLMAP approach and to that of the DOF-based SfM method described in Chapter 3. This is a first important step towards the design of a robust SfM method able to deal with the complex conditions of endoscopic scenes.

4.4 Effectiveness of the FMDOF method

The results in Fig. 4.4, Table 4.4, and Table 4.5 illustrate the interest of combining two matching methods in the white light bladder endoscopy. For the image pair $\left(\mathbf{I}_2^{ref}, \mathbf{I}_1\right)$, the SIFT approach was able to match 267 point pairs (see Table 4.5) using the 1007 and 919 feature points detected in images $\mathbf{I}_1$ and $\mathbf{I}_2^{ref}$, respectively (see Table 4.4). By observing the cyan lines in Fig. 4.4(b), one can note that in $\mathbf{I}_1$ and $\mathbf{I}_2^{ref}$, both the feature points and the matches are well spread over the

Table 4.4: Number of feature points obtained by the SIFT method. This table shows that the amount of detectable feature points can strongly vary, even between consecutive images.

Image	$\mathbf{I}_1$	$\mathbf{I}_2^{ref}$	$\mathbf{I}_3$
Feature points	1007	919	805

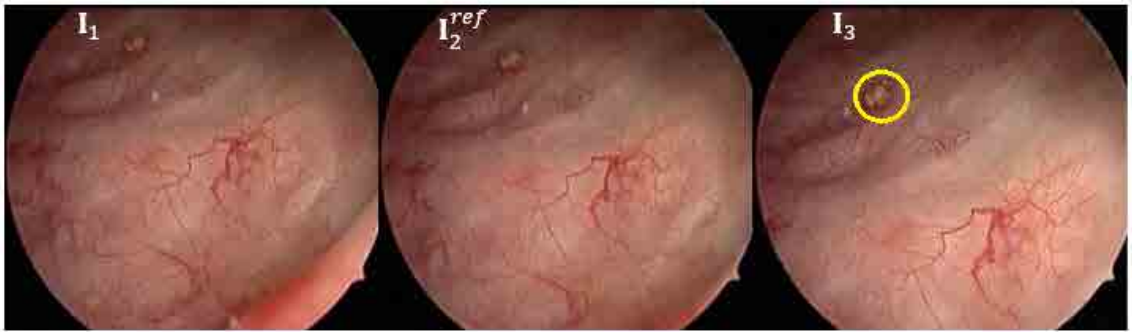
images. However, the green lines in Fig. 4.4(b) show that feature matches do not cover all image regions in $\mathbf{I}_2^{ref}$ and $\mathbf{I}_3$. The number of matches is rather small (122 correspondences, see Table 4.5) even if 805 features were detected in images $\mathbf{I}_3$ (see Table 4.4). This limited number of matches can be explained by the few textures available in the regions delineated in $\mathbf{I}_3$ by yellow rectangles. In Fig. 4.4(c), the link of the features points detected in reference image $\mathbf{I}_2^{ref}$ with their homologous points in $\mathbf{I}_3$ was established with the DOF field. Significantly more matches were obtained in all image regions and especially in the lower yellow rectangle. It is recalled that numerous homologous points should be spread over the images for an accurate and robust surface reconstruction with a SfM approach.

The polyp positions in the upper regions of the three images in Fig. 4.4(a) show that in urology large displacements can occur between consecutive images. Displacements of some tens of pixels (e.g., 40 or 50 pixels) often separate homologous pixels in cystoscopy or gastroscopy. It is well known that when features are available, SIFT can deal with such displacement magnitudes. However, the results shown here indicate also that the DOF field can be accurately and robustly determined with the method described in Section 3.4, even for large displacements and missing textures. Combining both matching methods thus favors the establishment of both numerous and accurate correspondences.

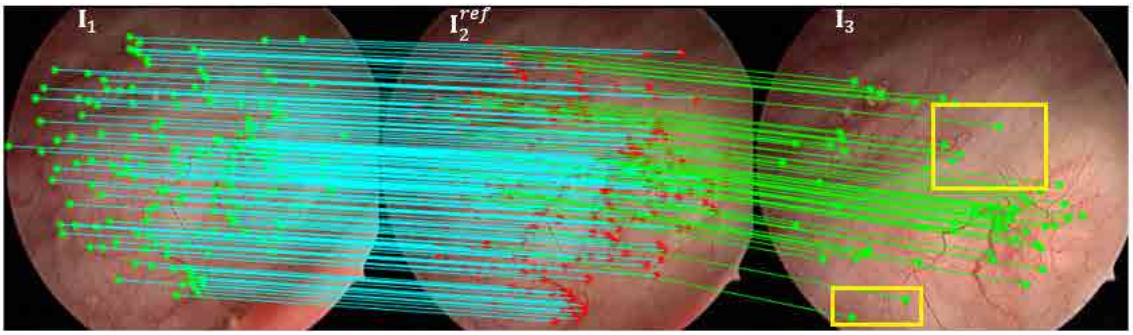
A cystoscopic video-sequence of 2468 images was used to reconstruct a large epithelial surface part of an inner bladder wall. The three images of Fig. 4.4(a) belong to this image sequence of about 100 seconds. In the proposed FMDOF-SfM method, the SIFT algorithm was able to determine enough matches in 1695 images among the 2468 images of the whole sequence. A surface reconstruction as in Fig. 4.5 allows for a second diagnosis (after the endoscopy) by zooming on regions of interest (polyp) of the archived map. A comparison of the surface reconstruction efficiency of the FMDOF-based SfM with that of other SfM methods, such as COLMAP and the proposed DOF-SfM scheme described in Chapter 3, is given in Subsection 5.2.2 of Chapter 5.

Table 4.5: Number of homologous points between the images of Fig. 4.4.

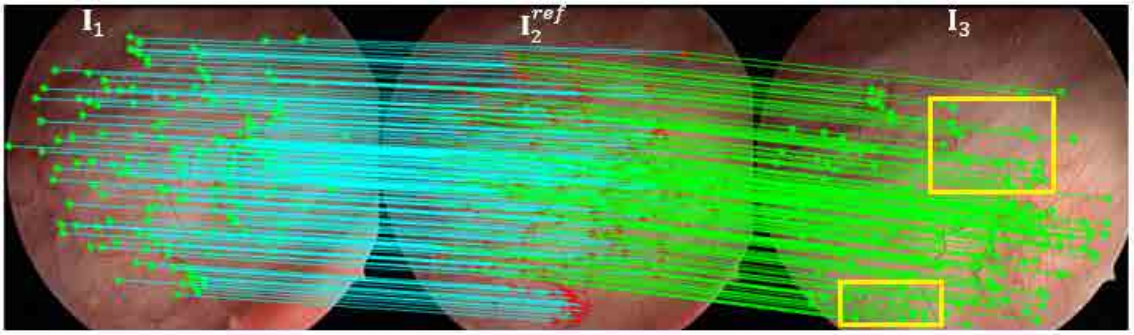
Image pair or triplet	$\left(\mathbf{I}_1, \mathbf{I}_2^{ref}\right)$	$\left(\mathbf{I}_2^{ref}, \mathbf{I}_3\right)$	$\left(\mathbf{I}_1, \mathbf{I}_2^{ref}, \mathbf{I}_3\right)$
SIFT method	267	122	89
SIFT+DOF method	267	218	218



(a)



(b)



(c)

Figure 4.4: Illustration of the complementary contributions of SIFT features and DOF fields in the search of homologous points in real bladder images. (a) Three cystoscopic images including a polyp delineated by the yellow circle in image $\mathbf{I}_3$. The image size $H \times W$ is 661×576 pixels. (b) SIFT matches obtained for image pairs $\left(\mathbf{I}_1, \mathbf{I}_2^{ref}\right)$ and $\left(\mathbf{I}_2^{ref}, \mathbf{I}_3\right)$. Only 89 triplets of homologous points were found mainly due to the few SIFT correspondences between images $\mathbf{I}_2^{ref}$ and $\mathbf{I}_3$. (c) The joint use of the two matching techniques leads to 218 triplets of homologous points. For image pair $\left(\mathbf{I}_2^{ref}, \mathbf{I}_1\right)$ the SIFT method alone provided all matches, while for the matches of pair $\left(\mathbf{I}_2^{ref}, \mathbf{I}_3\right)$, the features points detected in $\mathbf{I}_2^{ref}$ were matched with their homologous points in $\mathbf{I}_3$ using the DOF-field.

4.5 Main contributions and conclusion

Tables 4.4 and 4.5 show that, despite the numerous feature points obtained by SIFT,

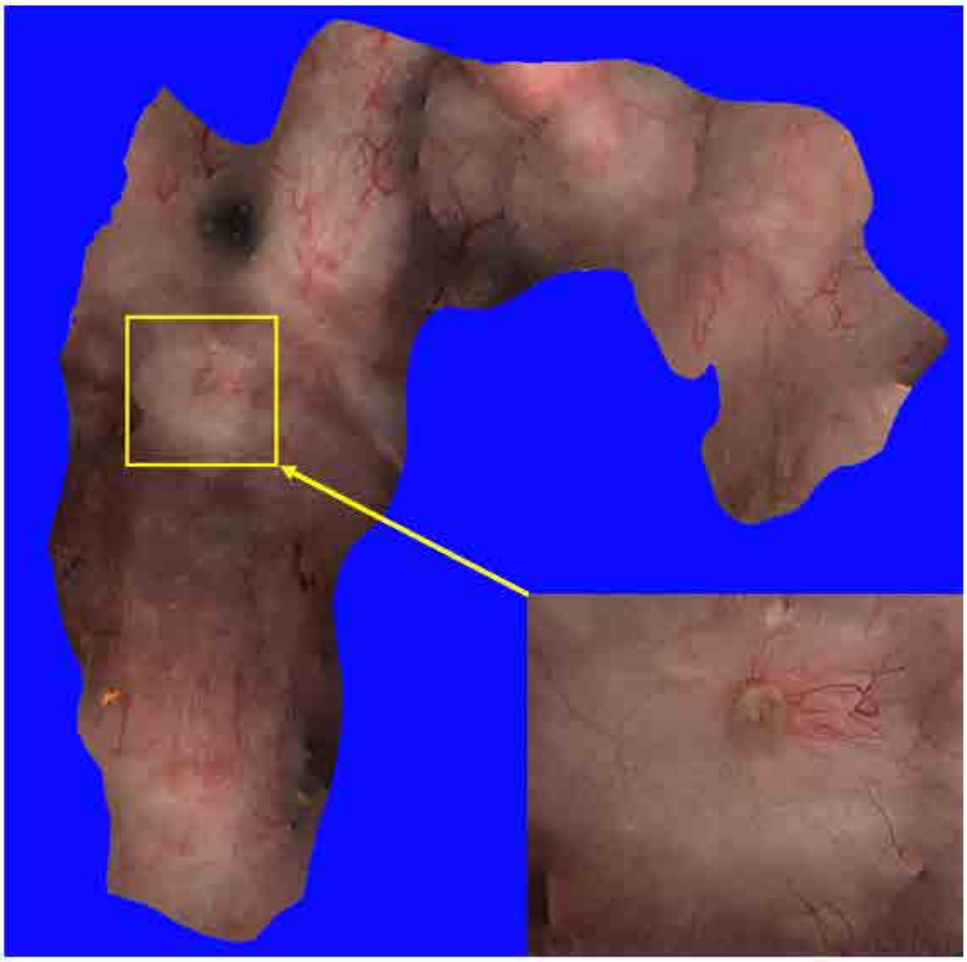


Figure 4.5: Bladder surface reconstruction with a polyp obtained by the proposed FMDOF-based SfM method.

only a limited number of homologous point groups can be established between some images. That is the reason why most of current SfM methods such as [SF16, Wu13, SSH+15] are not optimal to reconstruct the 3D surfaces in cystoscopy or fail in gastroscopy when they are used alone.

This chapter proposes a new solution for the HP-groups determination. The main contribution lies in the joint use of DOF and feature matches for generating large 2D point groups. This SfM step is crucial when complex scenes have to be reconstructed.

The method described in this chapter is dedicated to scenes in which features are present, but are lacking in some images or image regions. In video sequences with numerous features in all images the algorithm works in a similar way as the numerous contributions described in the literature [SF16]. On the contrary, in scenes without textures, the method is equivalent to the SfM pipeline described in Chapter 3. In the case of scenes like that occurring in cystoscopy (for example in Figs. 4.4(a) and 4.5), the combination of two matching methods is robust because, when possible, it takes advantage of the accuracy of classical feature based methods, and exploits the robustness of DOF fields when features are lacking or with too few contrast.

This chapter has shown that, rather than improving the accuracy, the combination of the two matching methods contributes to the robustness of the SfM method. As illustrated in Chapter 5, this matching method combination can also lead to surfaces with a larger extent in some situations. The work described in this chapter was presented at a conference.

National conference

- T.-B. Phan, D.-H. Trinh, D. Lamarque, D. Wolf, C. Daul, 3D surface reconstruction using dense optical flow combined to feature matching: Ap- plication to endoscopy, in: Colloque GRETSI, Lille, France, 2019. [PTL+19a]

The next chapter gives an overview of the 3D reconstruction results obtained with the methods given in Chapters 3 and 4. Both objective and subjective evaluations will be presented to validate the accuracy and robustness of the proposed methods.

Chapter 5

Epithelial surface reconstruction results

Contents

This chapter quantifies the performance of the DOF-based and FMDOF-based SfM schemes which are described in Chapters 3 and 4, respectively. The experiments done in Section 5.1 on phantoms with known ground truths give an idea on the inherent accuracy of the proposed SfM methods, while Section 5.2 highlights the robustness of the SfM algorithms which can deal with very different scenes and acquisition conditions. Besides that, Section 5.3 shows the potential of the proposed methods in non-medical scenes by reconstructing a small kitchen room. The DOF-based SfM detailed in Chapter 3 was successfully applied on various medical scenes and a non-medical scene, while the FMDOF-based SfM method can be used for scenes with partially lacking textures as arising in cystoscopy.

The performance of the proposed methods was compared to that of state-of-the-art methods as COLMAP ([SF16]) and VisualSFM ([Wu13]) which are well-established incremental SfM pipelines. Even published in 2013 and 2016 respectively, VisualSFM and COLMAP remain among the reference methods in terms of accuracy and robustness. More recent publications, instead of improving strongly the accuracy and robustness, adapted the SfM principle to improve its suitability for

different scenes or acquisition devices. It is notably the case for the contribution detailed in [NLB19]. The novelty of this recent publication is that the SfM approach is usable for very different camera systems (e.g., light-field cameras, camera-rigs, stereo cameras, etc.).

The proposed SfM methods were designed for endoscopic scenes. However, they are also applicable to other medical scenes and non-medical scenes which are with more or less textures (such scenes occur in dermatology for instance). This chapter demonstrates the robustness of the proposed methods on skin textures associated with phantom data for an accuracy evaluation, and with patient data for a robustness assessment.

For all tests with the DOF-based SfM scheme, the overlap threshold τ, error threshold ϵ, parameter h, and the DOF field parameters (λ, Py_s, γ_1, and γ_2) take the constant values given in Table 3.4 of Chapter 3. In the same way, the constant threshold values for the number of feature points (α threshold) and matched points (β threshold) used in the FMDOF-based SfM algorithm are $\frac{WH}{1000}$ and $\frac{WH}{1750}$, respectively, where W and H stand for the width and the height of the images.

5.1 Accuracy of the proposed SfM schemes

An objective evaluation is impossible on endoscopic data since no ground truth is available for patients. For this reason, the objective evaluation is performed on phantoms with known dimensions and carrying epithelial textures (see Fig. 5.1). We assess the accuracy of the proposed algorithms through the evaluation of the reconstructed phantom surfaces.

A cylindrical shape has been chosen for the phantoms since the accuracy of this parametrical surface can be quantified using a single value: its diameter. Moreover, gluing epithelial textures on the internal or external half-cylinder surfaces allows to simulate the acquisition of inner hollow organ walls or of the external skin surface. The two hollow organ surface phantoms (one for the bladder, the other for the stomach) are shown in Fig. 5.1.

5.1.1 Phantom description and data acquisition

The four phantoms used for the surface recovery tests consist of half-cylinders with precisely known internal and external diameters and carrying each an orange sphere whose diameter d_{gt} equals 40.14 mm (see Fig. 5.1). Epithelial tissue images were printed on paper sheets and glued onto the cylinders. Due to the printing of bladder and stomach images, the textures are by far more pronounced and numerous on the paper sheets than directly in the video-sequence images. In these experiments, two state-of-the-art SfM methods, namely COLMAP [SF16] and VisualSFM [Wu13] which use SIFT features [Low04], are placed in ideal conditions to find numerous correspondences. Thus, the accuracy of the proposed SfM methods (DOF-based SfM and FMDOF-based SfM) can be evaluated through a comparison with the results obtained with COLMAP and VisualSFM. A camera equipped by a 12 mm focal

length objective is used to acquire sequences of images with a size of 780 × 580 pixels. As in medical scenes where the acquisition is done close to the epithelial tissue, the camera/phantom surface distance was short so that each image only visualize a small object region.

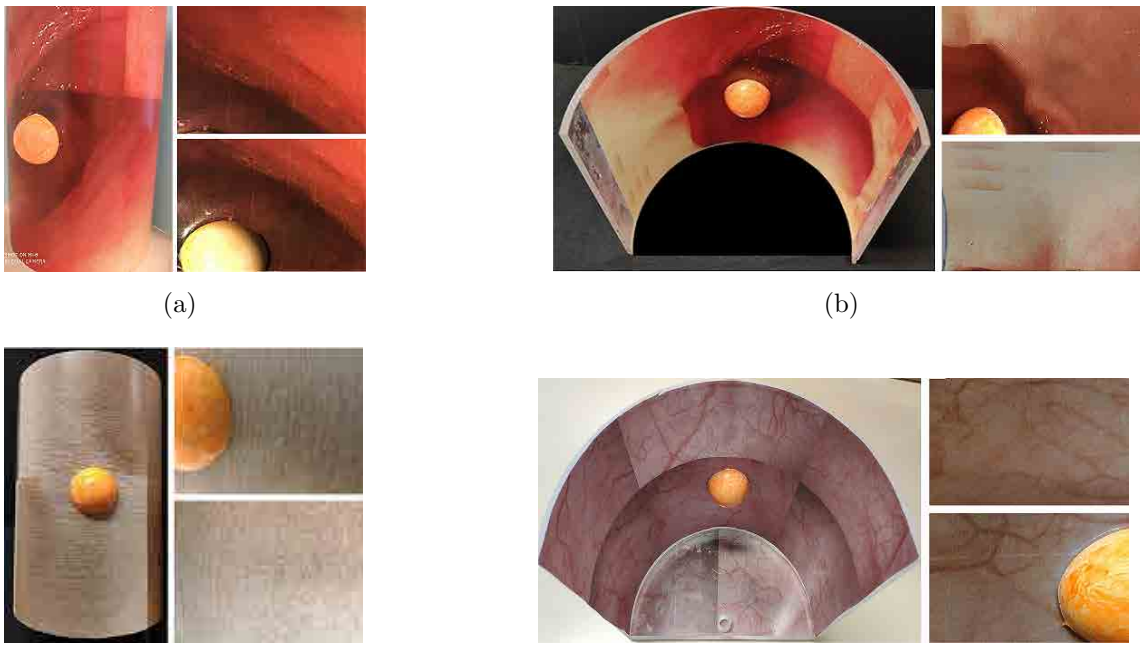
Internal stomach surface phantom. The inner surface of this phantom (see Fig. 5.1(b)) is covered with paper sheet printings of stomach images acquired during gastroscopies. The internal diameter D_{gt} of the cylinder equals 191.8 mm. A sequence of 265 small FoV images was acquired for this phantom.

External stomach surface phantom. The external cylinder surface with diameter $D_{gt} = 159.45mm$ is lined by gastroscopic images (see the left image in Fig. 5.1(a)). Gastroscopic images are printed on a paper sheet that was glued onto the cylinder. A sequence of 111 small FoV images was acquired for this phantom.

Since the textures of the internal and external stomach phantoms in Figs. 5.1(a) and 5.1(b) are quite similar, it is possible to assess the impact of the surface shape (positive or negative curvatures) on the reconstruction results.

External skin surface phantom. The size of this phantom is exactly the same as that of the external stomach surface phantom in Fig. 5.1(a). This phantom roughly simulates the shape of arm or leg parts. As in dermatology, the epithelium is on an external body surface and the camera which is close to the simulated tissues acquired a sequence of 621 images (two of the latter are shown on the right in Fig. 5.1(c)).

Internal bladder surface phantom. The phantom in Fig. 5.1(d) includes a cylinder with known internal diameter ($D_{gt} = 191.8$ mm). Cystoscopic images were printed



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

Figure 5.1: Snapshot of the complete phantom surface (on the left) and two small field of view images randomly chosen in the video-sequence. (a) External stomach surface phantom. (b) Internal stomach surface phantom. (c) External skin surface phantom. (d) Internal bladder surface phantom.

on a paper sheet that were glued onto the cylinder surface and a sequence of 293 small images was acquired for this phantom.

In our tests, the intrinsic parameters of the camera were not calibrated to be close to clinical conditions where such a calibration should be avoided for practical reasons. However, an additional experiment on the internal stomach phantom is presented in this section to assess the impact of the knowledge of the intrinsic camera parameters on the surface reconstruction.

The ratio of the cylinder and sphere diameters being scale independent, the ratio of the ground truth diameters (known by construction) should ideally be equal to the ratio of the diameters determined using the reconstructed point cloud. This explains why a sphere is carried by the half-cylinders.

5.1.2 Evaluation criteria

For the result evaluation, the 3D point clouds delivered by the SfM methods are first separated in two different surface parts, namely the cylinder part and the sphere part. Then, a fitting technique is separately applied to each part to obtain the equations of the reconstructed cylinder and sphere surfaces. For each reconstruction, both the diameters of the cylinder and sphere surfaces (denoted by D and d, respectively), and information relating to inlier and outlier points are calculated. A point of a reconstructed cloud is considered as an outlier when its distance to the estimated phantom surface is greater than $0.005.D$ (i.e., 0.5% of the cylinder diameter).

Five criteria are used to evaluate the accuracy of the reconstruction methods:

1. The **outlier rate** (in %) corresponds to the ratio of the outlier number over the whole 3D point number of the cloud.

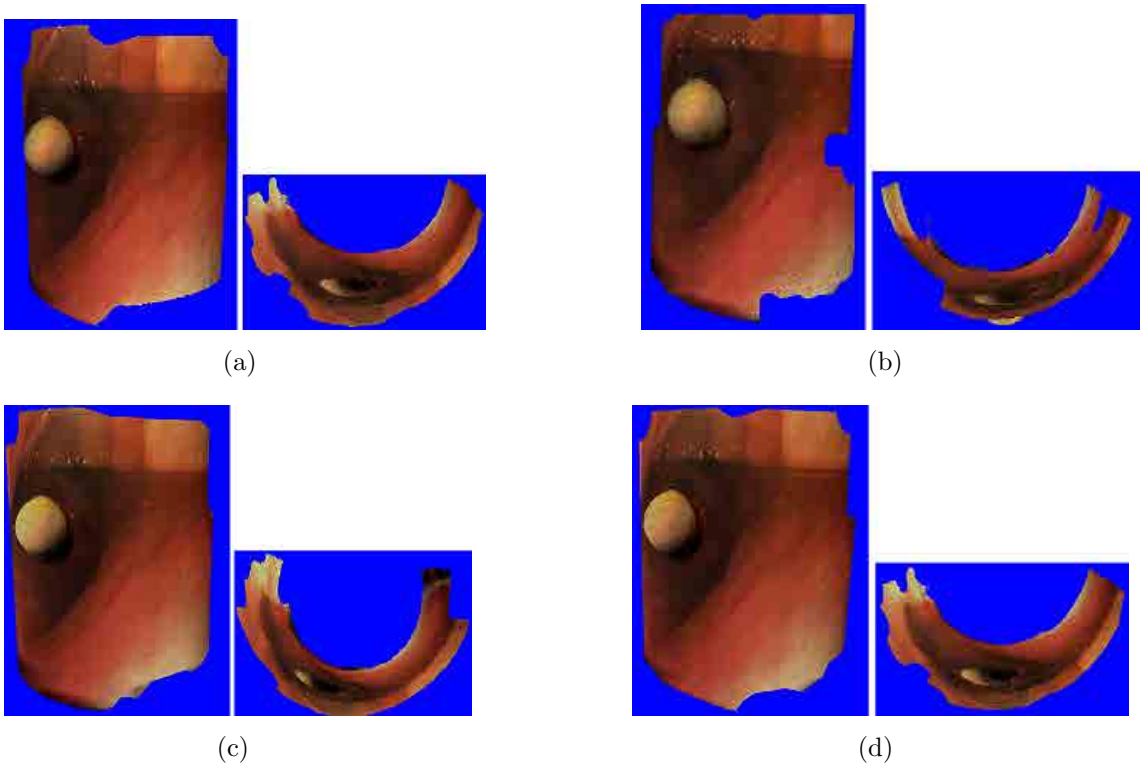
2. The **mean outlier error** (in mm) gives the mean distance between outlier points and the fitted phantom surface.

3. The **3D phantom shape accuracy** is assessed by comparing the diameter ratio D/d of the reconstructed cylinder and sphere surfaces with their ground truth D_{gt}/d_{gt}. This criterion is defined by Eq. (5.1.1) in which $p = 100\%$ and $p = 0\%$ indicate a perfect and a completely wrong shape, respectively.

$$p = \left(1 - \frac{|D_{gt}/d_{gt} - D/d|}{D_{gt}/d_{gt}}\right) \times 100. \tag{5.1.1}$$

4. The **number of 3D points**. In the proposed methods, this number is directly obtained after the SfM step which provides a dense point cloud. In COLMAP and VisualSFM, the 3D point number is obtained after a SfM step leading to a sparse point cloud and which is classically followed by a MVS step giving the final dense point cloud under interest.

5. For COLMAP and VisualSFM, the **computation time** criterion includes the total time of the SfM and MVS parts (C++), while for the proposed methods



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

Figure 5.2: Textured surface of the external stomach phantom (under two viewpoints) obtained with (a) COLMAP, (b) VisualSFM, (c) DOF, and (d) FMDOF. The surfaces constructed with the four methods are visually similar.

it corresponds to the SfM part (MATLAB) and the optical flow computation (C++). Experiments were performed on a HP Pavillion laptop with an Intel Core i5 1.60GHz and 16GB RAM and NVIDIA GeForce 940MX GPU.

5.1.3 Phantom reconstruction results

5.1.3.1 Evaluation on the external and internal stomach surface phantoms

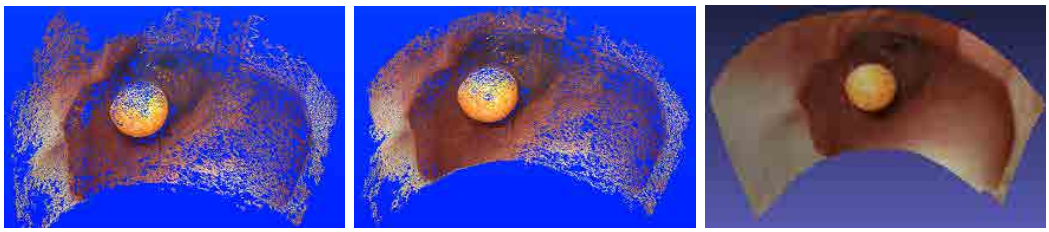
As seen in Tables 5.1 and 5.2, the results obtained for the external and internal stomach surface phantoms show the superiority of the DOF-based SfM over the three

	3D phantom shape accuracy (p in %)	Outlier rate (%)	Mean outlier error (mm)	Number of 3D points	Computation time (Minutes)
COLMAP+MVS	98.18	8.29	7.47	398096	79.16
VisualSFM+MVS	98.19	9.14	7.95	364428	75.52
DOF	98.23	7.98	6.12	478306	86.91
FMDOF	98.11	8.18	7.37	355901	81.07

Table 5.1: Reconstruction accuracy for the external stomach surface phantom.

	3D phantom shape accuracy (p in %)	Outlier rate (%)	Mean outlier error (mm)	Number of 3D points	Computation time (Minutes)
COLMAP+MVS	98.29	6.84	8.298	173262	44.1
VisualSFM+MVS	97.21	7.14	8.017	159971	43.4
DOF	98.76	5.17	5.65	233639	79.15
FMDOF	97.13	6.71	6.998	191148	55.7

Table 5.2: Reconstruction accuracy for the internal stomach surface phantom.



(a) COLMAP (SfM+MVS) (b) DOF-based SfM (c) Textured surface

Figure 5.3: Internal phantom reconstruction results (same viewpoint as in Fig. 5.1(b)). (a) Point cloud given by COLMAP [SF16]. (b) Point cloud obtained with the proposed SfM method. (c) Textured triangle mesh obtained with the point cloud in (b).

other SfM methods (COLMAP, VisualSFM, and FMDOF) in terms of accuracy, outlier rate, mean outlier error and number of 3D points. The DOF method provides the most dense point cloud, while leading to the lowest outlier rate. However, the computation time of the DOF-based SfM method is higher than that of the three other methods, especially in comparison to COLMAP and VisualSFM.

Although the performances of the FMDOF-SfM method are lower than those of the DOF-SfM approach, the joint use of features and DOF leads to a reconstruction with a rather low outlier rate and mean outlier error in comparison to the corresponding criterion values of COLMAP and VisualSFM.

The number of 3D points in the proposed methods depend on the initialization of the 2D points on the reference images (i.e. on the grid cell size). One can easily increase the number of 3D points through an adjustment of the HP-group parameters. For instance, decreasing the grid size h while keeping ϵ and τ constant during the point group computation, or increasing the value of α and β in the FMDOF algorithm to get more 2D points on each reference image allow to gain more 3D points. However, in this work, one only need a sufficient amount of 3D points so that the use of the MVS step can be avoided. In addition, too many 3D points will lead to a significant increase of the computation time and may cause more outlier points.

The surface reconstructions obtained with the two proposed methods, COLMAP and VisualSFM for the external stomach phantom are given in Fig. 5.2. The DOF-based SfM scheme with both the highest shape accuracy and the greatest number of

	3D phantom shape accuracy (p in %)	Outlier rate (%)	Mean outlier error (mm)	Number of 3D points	Computation time (Minutes)
COLMAP+MVS	99.35	4.29	6.92	591126	150.5
VisualSFM+MVS	99.33	4.26	7.0	572101	144.9
DOF	99.32	3.58	7.5	558397	147.8
FMDOF	99.33	4.05	7.2	478101	141.9

Table 5.3: Reconstruction accuracy for the external skin surface phantom.

3D points (see Tables 5.1) led to a very realistic shape of the half-cylinder. Fig. 5.3 proposes a visualization of the two most accurate methods for the internal stomach phantom reconstruction. The number of 3D points obtained by the DOF-based SfM (233639) is much higher than the number of 3D points (173262) given by COLMAP, even if the latter uses a MVS step.

It is noticeable that all methods have slightly better global performances for the external stomach phantom than for the internal phantom. But the differences in performances are weak and are not necessarily due to the curvature signs (positive of negative according to the internal or external cylinder phantoms) but can also depend on the number of images required to scan whole surfaces (111 images for the external phantom cylinder with a diameter of 159.5 mm and 265 images for the internal phantom cylinder with a diameter of 191.8 mm). For both phantoms, the relative performances of all methods are the same (i.e, the ranking of the methods according to their performances is the same).

5.1.3.2 Evaluation on the external skin surface and internal bladder surface phantoms

Tables 5.3 and 5.4 give the reconstruction results for the external skin surface and internal bladder surface phantoms. The SfM methods based on SIFT (COLMAP and VisualSFM) led to more dense 3D point clouds than the proposed methods since the reconstructed phantoms exhibit more contrasted textures than the two stomach phantoms in Figs. 5.1(a) and 5.1(b). The shape accuracy criteria have all very similar values for the four SfM methods. This is an indication that the DOF based SfM method can lead to a similar reconstruction accuracy than a feature based SfM method, even if an optical flow vector localizes homologous points with a pixel accuracy and the SIFT point localization has a subpixel accuracy.

The DOF-based SfM method still obtained the lowest outlier rate among the four methods. However, for the skin and bladder phantoms, the FMDOF-based SfM method outperforms the DOF-based SfM in terms of 3D shape accuracy and computation time. This result highlights the interest of the joint use of features and DOF fields when giving textures are available for the 3D reconstruction of homologous points.

A visual comparison of the two reconstructions in Fig. 5.4 shows that the point cloud obtained with the proposed DOF method covers a greater cylinder surface

(a) COLMAP (SfM + MVS) (b) Proposed DOF-based SfM

Figure 5.4: SfM results obtained for the skin phantom shown in Fig. 5.1(c). (a) 3D point cloud obtained with COLMAP [SF16] and zoom on the sphere region delineated by the green rectangle. (b) 3D point cloud obtained with the proposed DOF method and zoom on the sphere.

than that provided by COLMAP (see the cylinder borders). Fig. 5.5 allows for a comparison of the textures obtained with the FMDOF-based SfM method and those given by Visual SFM. The surface obtained by visualSFM is with missing parts in the upper phantom region. These parts are present in the surface reconstructed with the FMDOF-based SfM method. Despite VisualSFM provides more 3D points than FMDOF, the distribution of the 3D points reconstructed by VisualSFM doesn't spread out over the entire surface. On the contrary, for the FMDOF-based method, the 3D points cover the whole phantom surface.

5.1.3.3 Surface construction using an uncalibrated and calibrated camera

Tests were performed with the internal stomach phantom to assess the impact of the knowledge of the intrinsic camera parameter values on the 3D surface reconstruction accuracy.

The intrinsic camera parameters consist of the focal length, the principle point coordinates, the scale factor, and the distortion coefficients. There are various techniques for estimating the internal camera parameters: self-calibration [HZ04], the

	3D phantom shape accuracy (p in %)	Outlier rate (%)	Mean outlier error (mm)	Number of 3D points	Computation time (Minutes)
COLMAP+MVS	98.79	5.77	7.107	243987	55.02
VisualSFM+MVS	98.10	5.83	7.009	239655	54.26
DOF	98.14	5.72	6.98	238989	73.97
FMDOF	98.22	5.79	7.004	235117	59.08

Table 5.4: Reconstruction accuracy for the internal bladder phantom.

(a) (b) (c)

Figure 5.5: Internal bladder surface phantom reconstruction results (same viewpoint as in 5.1(d)). (a) SfM result obtained with the FMDOF method. (b) Textured surface obtained by VisualSFM [Wu13]. (c) Textured (and meshed) surface obtained with the point cloud in (a).

calibration with patterns (e.g., checkerboards [Zha00]), or simply using the standard camera parameters given by the Exchangeable image file format (Exif tags [EXI]). The introduction to these approaches was given in Chapter 2.

Three situations were considered to evaluate the influence of the knowledge of the intrinsic camera parameters on the reconstruction of 3D surfaces.

1. The intrinsic camera parameters are calibrated and kept constant during the SfM process.

2. The intrinsic camera parameters are calibrated and optimized by the bundle adjustment algorithm of the incremental SfM scheme.

3. The camera is uncalibrated. In this situation, the initial values of the internal camera parameters are given by the Exif tags file and optimized by the bundle adjustment steps of the incremental SfM method.

For these tests, the Basler camera, equipped with a 12 mm focal length and used in Subsection 5.1.1 with unknown intrinsic parameters, was calibrated using the state-of-the-art algorithm of Zhang [Zha00]. This algorithm delivers accurate intrinsic parameters when numerous images of a pattern (standardly a checkerboard) are acquired from different viewpoints (different angles and distances).

The DOF-based SfM, the FMDOF-based SfM, the VisualSFM (together with a MVS algorithm), and the COLMAP (also together with a MVS algorithm) methods are used in three situations enumerated above to assess the impact of the knowledge of the camera calibration parameters on the reconstruction of the internal stomach phantom. Table 5.5 gathers the values of the shape accuracy criterion (percentage p in Eq. (5.1.1)), the outlier rate (OR in %) and the mean outlier error (MOE in mm), these results being given for each of the four SfM methods placed in the three different calibration situations. The number of reconstructed 3D points and the computation time are not given since their values are the same in all three camera calibration scenarios.

Table 5.5 shows that the shape accuracy criterion, the outlier rate, and the mean outlier error have similar values for each SfM method taken individually, regardless of the calibration situations.

	Internal stomach phantom								
	Camera calibration (fixed parameters)			**Camera calibration with BA**			**Uncalibrated camera with BA**		
	p (%)	OR (%)	MOE (mm)	p (%)	OR (%)	MOE (mm)	p (%)	OR (%)	MOE (mm)
COLMAP+MVS	97.67	6.82	8.33	98.3	6.79	8.18	98.29	6.84	8.29
VisualSFM+MVS	96.02	7.2	8	97.2	7.15	7.87	97.21	7.14	7.9
DOF	97.12	5.19	5.69	98.63	5.12	5.61	98.76	5.17	5.65
FMDOF	96.11	6.76	7.1	97.11	6.7	6.9	97.13	6.71	6.99

Table 5.5: Accuracy criterion values according to the SfM method and the calibration scenario. The terms OR and MOE refer to as the outlier rate and the mean outlier error, respectively.

Intuitively, it would not have been surprising that a calibration of the intrinsic parameters with the precise method of Zhang would improve the accuracy of the reconstruction since fewer parameters need to be optimized by the incremental SfM method. However, the results obtained without any calibration are the best for all four SfM methods. Even the second best scenario with initially calibrated cameras whose intrinsic parameters are adjusted by the incremental SfM scheme lead to a higher accuracy than the scenario with calibrated and fixed camera parameters. This is an indication that obtaining very precise surfaces is subject to the simultaneous adjustment of all the geometrical parameters, i.e. not only the camera pose but also the perspective projection and distortion parameters.

These results show that it is not required to conceive a simple and flexible endoscope calibration method usable in clinical situations. An incremental SfM scheme based on data acquired with uncalibrated cameras delivers the most accurate results.

5.1.3.4 Global discussion

Globally, the results obtained with the four phantoms highlight the accuracy of the two proposed methods since the values of their quality criteria were quite similar to those of COLMAP and VisualSFM which have a high precision in the presence of contrasted textures (the diameter ratios obtained by the four SfM methods are notably very close). It was important to objectively assess the inherent accuracy of the proposed methods before applying them to medical scenes since the real shape of hollow organs are unknown and can only be evaluated in a subjective way. The high inherent reconstruction accuracy measured on phantoms with the proposed methods indicate that the DOF and FMDOF-based SfM approaches can potentially reconstruct organs which are coherent in terms of 3D shape.

The DOF-based SfM outperformed the three remaining SfM methods when applying it to the scenes with rather few textures as the external and internal stomach phantoms while the FMDOF-based SfM is effective when it is applied to scenes with more textures and structures as the skin and bladder phantoms. The results confirm that, according to the scene content, one has to choose the DOF or the FMDOF-

based SfM method. For scenes with almost no textures the DOF-based method is the unique solution, and for scenes with few textures the FMDOF method can lead to optimal results.

The tests with the three camera calibration scenarios have also shown that an accurate and initial calibration of the intrinsic parameters does not improve the SfM algorithms. Moreover, an endoscope calibration being not done in practice in clinical conditions, there is no interest to modify the standard protocols of endoscopic examinations when surfaces have to be reconstructed. The tests performed in the next section on clinical data were performed on video-sequences acquired during standard examination procedures.

5.2 Tests on various medical scenes

The epithelium is visualized by cameras in various medical examinations. In dermatology, the skin is observed by colour cameras with the aim to distinguish between moles and cancers (carcinomas and melanomas). In gastroenterology, gastroscopes (i.e., endoscopes acquiring colour or narrow band image sequences) are used for the diagnosis of inflammations in the stomach. Chronic epithelium inflammations in the pyloric antrum or cardia regions increase the risk of stomach cancer. In urology, endoscopes (cystoscopes) enable the detection and follow-up of various bladder lesions (polyps, multi-focal cancers, etc.). The cystoscopic video-sequences are classically acquired in the white light modality or sometimes in the fluorescence modality. This section will give the results for all of the modalities mentioned above.

5.2.1 3D mosaicing of the pyloric antrum in gastroscopy

5.2.1.1 Medical motivations

The aim of this section is to highlight the ability of the DOF-based SfM method to deal with real datasets which are taken from very complicated scenes. In such scenes the methods based on SIFT are almost inoperative.

The pyloric antrum region is a privileged site of chronic inflammations that can lead to stomach cancers. In this context, the aim of a collaborative research project (EMMIE project sponsored by the French Research Agency, ANR) including a gastroenterologist from the Ambroise Paré Hospital in Boulogne Billancourt (Prof. D. Lamarque) and the University of Burgundy (Prof. F. Marzani) is to build a multi-spectral gastroscope which provides, in addition to the standard white light colour videos, multi-spectral images of the pyloric antrum region. These images at different wavelengths are acquired using a fiber that is brought into the stomach through the operator channel of the endoscope.

A work currently in progress at the University of Burgundy aims to classify the data extracted from the multi-spectral images and to detect different grades of inflammations [KCT+18]. Inflammations are not visible in the white light modality at an early stage, but after their detection in the multi-spectral data it would be

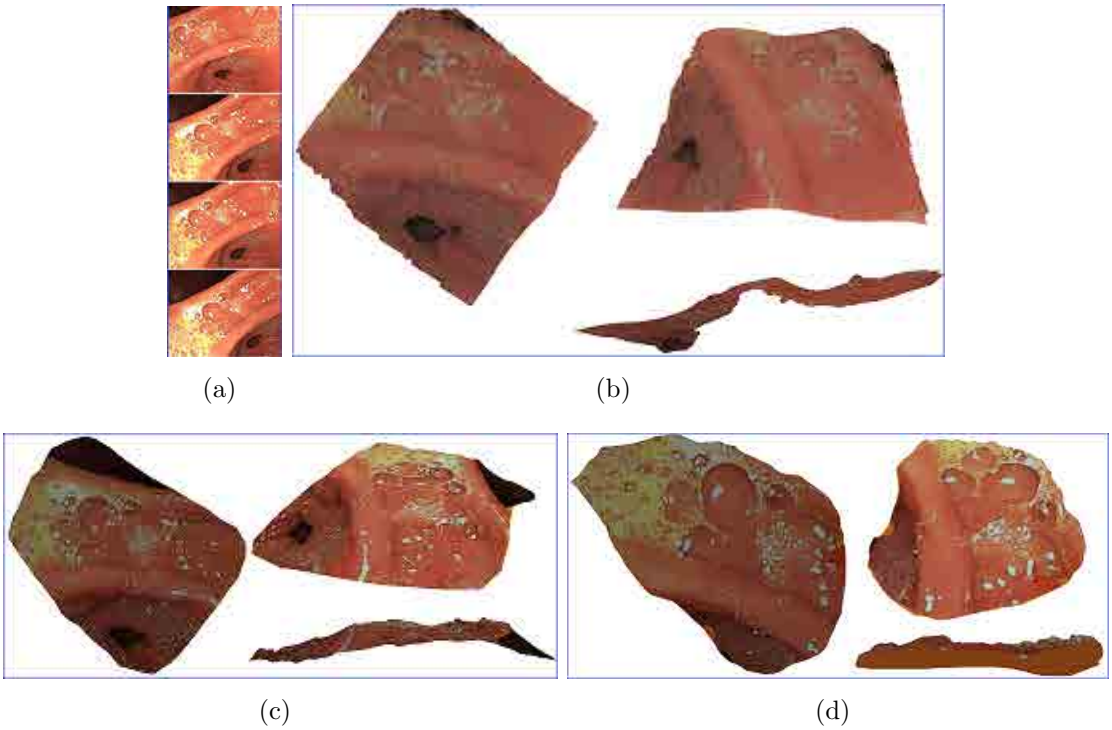


Figure 5.6: Surface construction comparison. (a) Four viewpoints (small FoV images) of the pyloric antrum region. (b) Realistic surface obtained with the proposed DOF method. (c) and (d) Too planar surfaces obtained with COLMAP and VisualSfM, respectively.

possible to mark their position (and grade) on the appropriate place onto a 3D mosaic build with the standard colour images of the stomach wall.

Constructing extended field of view (FoV) surfaces of the pyloric antrum region has several advantages:

(i) it will lead to an original and medically interesting way to document a gastroscopic examination,

(ii) 3D stomach regions marked with multi-spectral information will be a new information exchange media between gastroenterologists and other specialists,

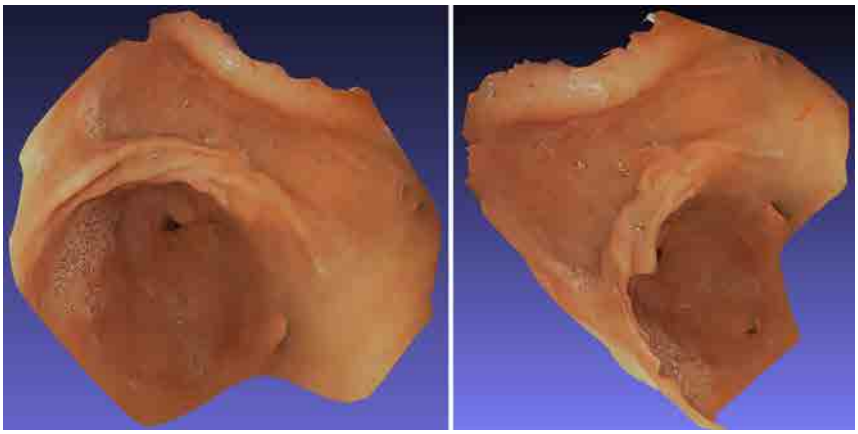
(iii) and the comparison of two marked 3D mosaics built with two video-sequences acquired at some weeks or month intervals will allow for a inflammation follow-up.

Although the classification algorithm and the placement of the inflammation information onto the colour images are still in progress, 3D mosaics of the pyloric antrum region can already be constructed using the DOF based SfM method proposed in this thesis.

5.2.1.2 White light gastroscopy

The four gastroscopic images in Fig. 5.6(a) were used to reconstruct a part of the pyloric antrum region connecting the stomach to the duodenum (intestine beginning,

(a)



(b)

Figure 5.7: Reconstruction of the pyloric antrum region. (a) Three different viewpoints among 101 images. (b) Reconstructed surface under two different viewpoints.

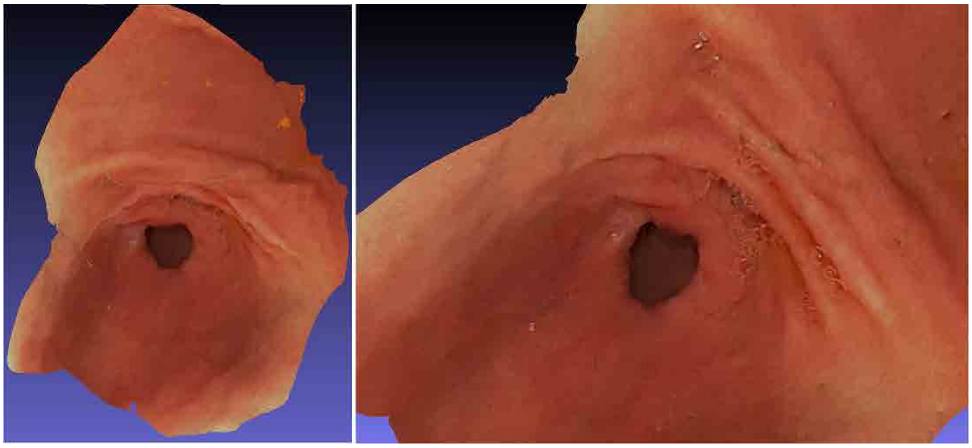
see the "black hole"). Around the duodenum, the stomach wall is characterized by strong curvatures which can be seen in the surface obtained by the proposed DOF method (see lower bottom region in Fig. 5.6(b)). On the contrary, as seen in Figs. 5.6(c) and 5.6(d), with such few images, COLMAP and VisualSFM give an unrealistic planar shape due to the lack of feature points, the strong illumination changes and the liquid leading to false SIFT correspondences.

A sequence of 101 endoscopic images was used to reconstruct the extended FoV of the stomach wall (see Fig. 5.7(b)). The three images in Fig. 5.7(a) show the complexity of the scene almost without textures. While COLMAP and VisualSFM completely failed in the reconstruction of this scene[1], the proposed DOF algorithm was able to retrieve the "funnel shaped" surface corresponding to the pyloric region (see the right image in Fig. 5.7(a)) and the more "planar" surface of the antrum (see the central image in Fig. 5.7(a)).

Fig. 5.8(a) shows four images among 191 frames of a second white light video-sequence acquired in a stomach. No textures and only few structures are visible in these images which are classically affected by reflections. The lack of textures would impede SIFT based approached to establish numerous correspondances, while the refections would lead to wrong matches. The pyloric antrum region surface constructed with the DOF-based SfM approach in presented in Fig. 5.8(b) under two viewpoints (i.e., at different orientations and scales). With such surfaces, gastroen-

[1]Refer to Section 3.6 of Chapter 3 which compares the SIFT matching and the DOF based HP-grouping for such scenes. This section explains why COLMAP fails for such scenes.

(a)



(b)

Figure 5.8: Pyloric antrum region surface obtained with the DOF-based SfM method in the white light modality. (a) Four colour images of the sequence of 191 frames. (b) Pyloric antrum region under to viewpoints.

(a) (b)

Figure 5.9: Pyloric antrum region surface obtained with the DOF based SfM method in the NBI modality. (a) Four colour images of the sequence of 214 frames. (b) Constructed stomach surface.

terologists are able to virtually navigate into the stomach after the examination. It is also noticeable that most of the reflections do not appear in the 3D mosaics since these illumination effects are not systematically present for all viewpoints on a same 3D point (viewpoints without reflections can be chosen during the surface texturing).

5.2.1.3 Narrow-band imaging

Fig. 5.9(a) gives four images among 214 frames of a narrow band imaging (NBI, green-blue light source) video-sequence acquired in a stomach[2]. Liquid, motion blur and reflections appear standardly in such images. For these medical scenes, SIFT based correspondances are also difficult to establish. Inflammations at an advanced stage can visually be earlier detected in the NBI modality than in the white light modality. For this reason NBI gastroscopy is often used as a complementary modality. However, placing multi-spectral information on NBI 3D mosaics would also clearly facilitate an early diagnosis and medical data archiving. A greatly extended 3D FoV of a pyloric antrum region is shown in Fig. 5.9(b).

5.2.2 3D bladder wall mosaicing in cystoscopy

5.2.2.1 Medical context

In cystoscopy, urologists or surgeons scan the internal bladder wall with the endoscope's distal tip maintained close to the epithelial tissue. Urologists have to mentally reconstruct bladder parts in order to be able to know the position and orientation of the cystoscope into the organ. To do so, the endoscopist comes regularly back to landmarks like either the meatus of the urethra and the ureters, or the air bubble that locates the top of the bladder which is filled with a saline isotonic liquid. Locating the exact position of the instrument in the bladder in this way is only possible during the cystoscopic examination. This is one reason why cystoscopic video sequences are usually not archived to document an examination. Building 2D or 3D mosaics of large bladder parts is a solution to archive important information for the examination traceability and lesion evolution assessment. Moreover, controlling exactly the trajectory of the cystoscope is very difficult. Consequently, it is never obvious whether a region of interest (e.g., with potential lesions) was completely scanned or not. Constructing large FoV mosaics is also a solution for this issue since non scanned regions are visible when a cartography of organ parts can be robustly done.

5.2.2.2 White light cystoscopy

Fig. 5.10(a) shows four small FoV bladder images of a white light video sequence consisting of 1101 frames. The 2D mosaic in Fig. 5.10(b), even constructed using a robust image registration technique [ADG+16], has several drawbacks. On the one hand, all images are placed in a 2D mosaicing plane whose coordinate system

[2]See the video available at https://github.com/CRAN-BioSiS-Imaging/PR2020

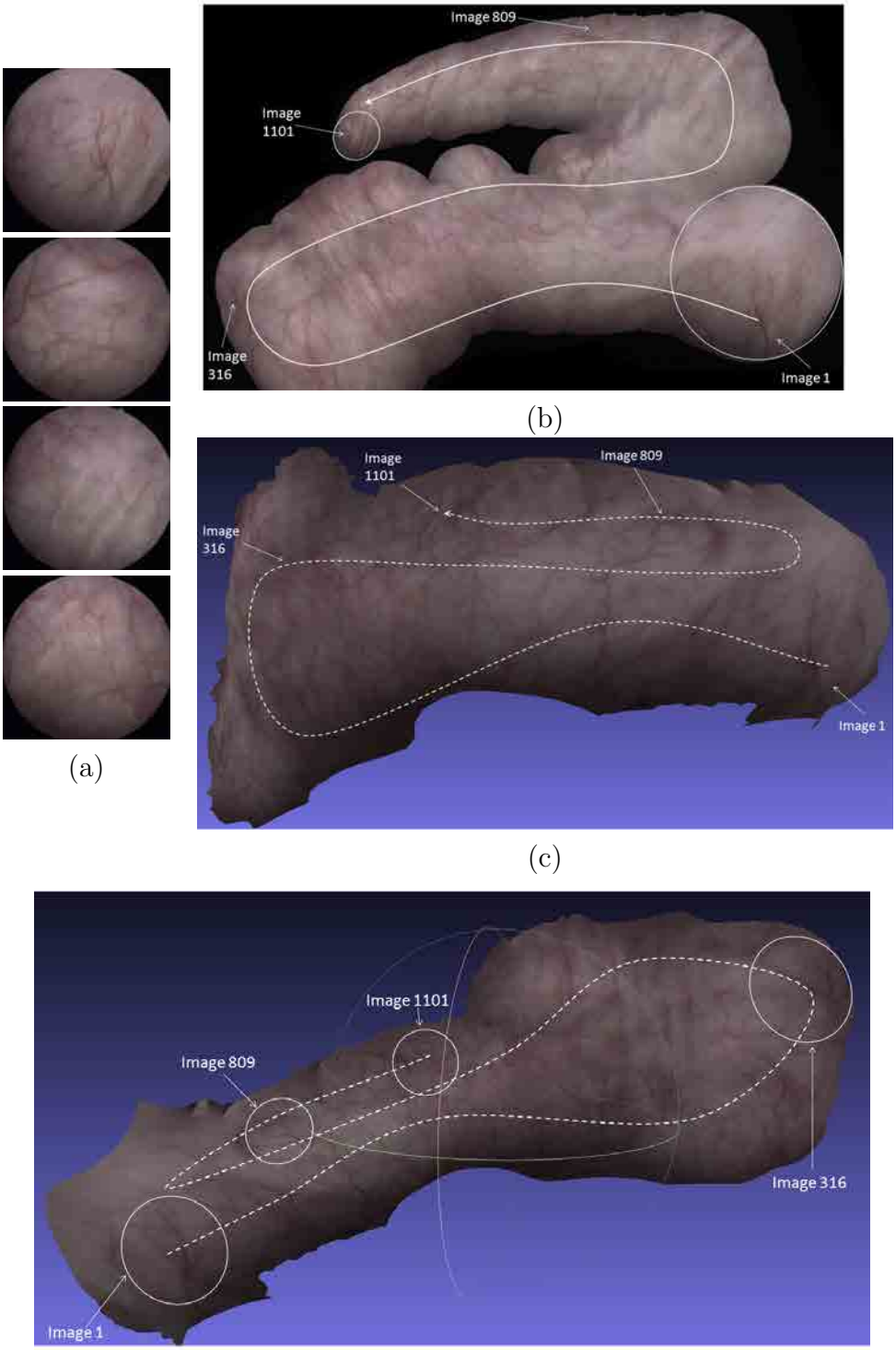


Figure 5.10: Bladder surface part represented by 2D and 3D mosaics. (a) Four bladder images of a sequence of 1101 frames (from the top to the bottom: images 1, 316, 809 and 1101). (b) 2D mosaic (c) 3D mosaic: bladder wall under the viewpoint of the 2D mosaic in (b). (d) 3D mosaic seen from a viewpoint located into the bladder.

is defined by that of image 1 acting as reference. The mosaic path is materialized by the white curve (solid line), the pixels of image I_i being added to the current mosaic built with information of images I_1 to I_{i-1}. The arrow of this curve indicates the last image of the sequence (I_{1101}). As visible in Fig. 5.10(b), the first and last images correspond to ellipses with very different areas in pixels. The resolution of an image in the mosaic plane depends strongly on the viewpoint changes of the endoscope between the images, leading to both strong image distorsions and significant resolution losses (only image I_1 is without resolution loss). On the other hand, due to accumulating registration errors, images which should be overlapped are in different (non-overlapping) places in the 2D mosaic plane. Thus, images I_{809} to I_{1101} should partly overlap the previous images of the sequence. The gaps without bladder texture in the 2D map are not due to tissue areas which were not scanned by the endoscope, but to an accumulating registration error that grows during the map construction. These registration errors are difficult to be globally corrected, even with sophisticated techniques as described in [WDW+12]. In 2D mosaics it is not possible to distinguish between mosaicing errors and tissues that were not scanned by the endoscope.

Fig. 5.10(c) shows the bladder surface reconstructed using the proposed DOF-based SfM method applied to the 1101 image sequence. Even if this surface is curved, its orientation was chosen so that the 3D mosaic content can be visually compared to that of the 2D mosaic in Fig. 5.10(b).

It can be seen that the surface is without gaps and that the organ was locally completely scanned by the endoscope. The dashed white line approximatively represents the endoscope trajectory position computed for the 2D mosaic (this trajectory is not computed in the proposed SfM method). In Fig. 5.10(b) it is visible that, when moving along the endoscope trajectories, the scales of the individual images in the 2D mosaic are strongly changing even if an urologist try to keep a more or less constant distance between the epithelium and the endoscope distal tip, at least for a part of the wall. In Fig. 5.10(c), this distance is approximatively constant since globally the blood vessel appear visually with a same size. The global surface shape and proportions are quite different in the 2D and 3D mosaics.

Due to the comparison with the 2D mosaic, Fig. 5.10(c) shows the surface as if the endoscope could be outside the bladder and as if the epithelial textures could by seen by transparency. Fig. 5.10(d) proposes a real surface viewpoint located inside the bladder. Such a surface allows for a virtual navigation after the cystoscopic examination with the possibility to observe each part from freely chosen viewpoints. Such a 3D mosaic can be archived and used as information media exchange and patient follow-up.

A second cystoscopic video-sequence of 2468 images was used to reconstruct a large part of an internal bladder wall surface (see three images of the sequence in Fig. 5.11(a)). The DOF-based SfM, the FMDOF-based SfM, and COLMAP methods were applied to this data and were able to reconstruct 1140654, 961598, and 819923 3D points, respectively. As shown in Fig. 5.11, among the three surface construction methods, the DOF-based SfM approach led to the surface with the largest extend (see Fig. 5.11(c)), notably because this method determined the greatest num-

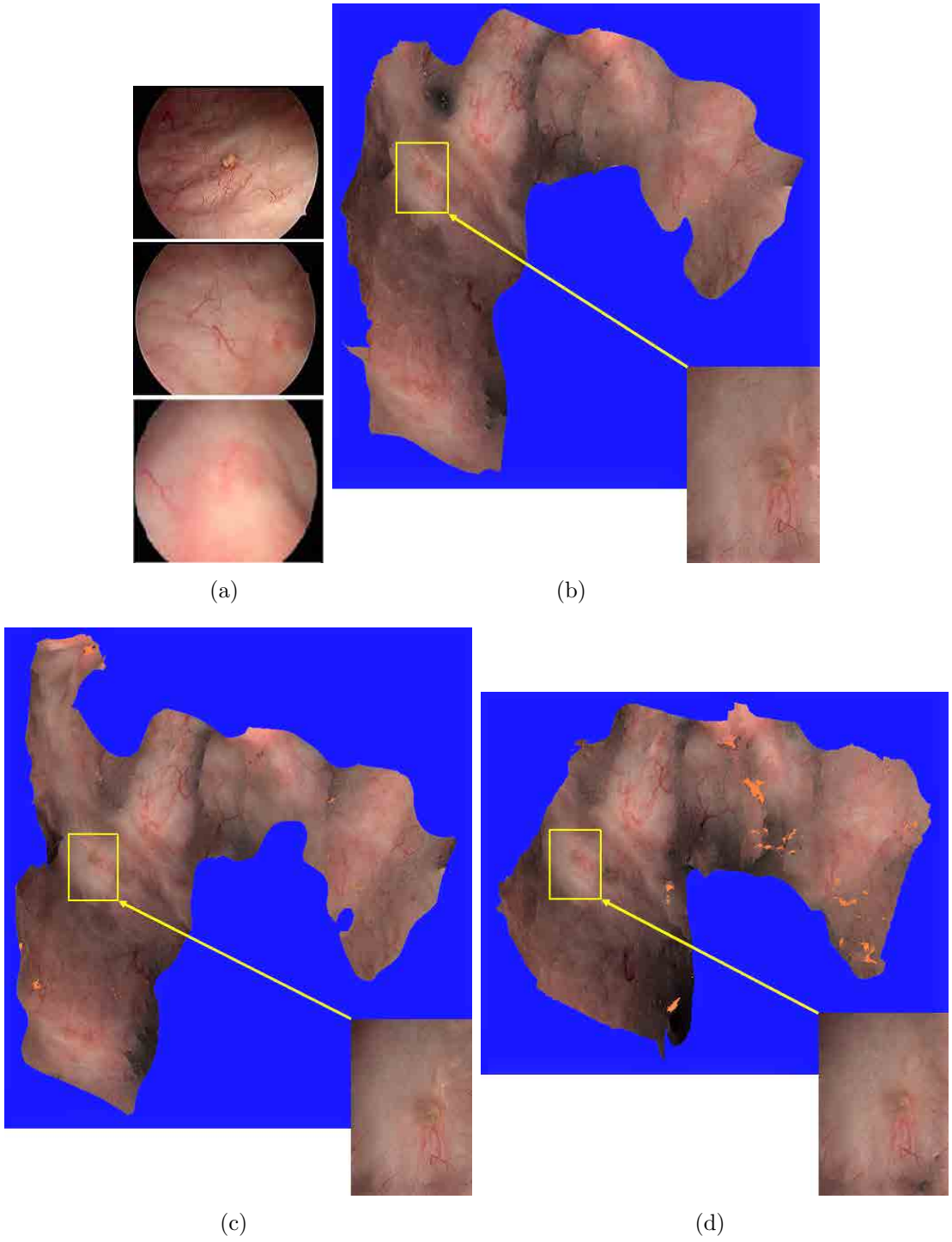


Figure 5.11: Second bladder surface reconstruction example. The yellow rectangles delineate an area with a polyp. (a) Three small FoV images of the video-sequence of 2468 frames. (b) Surfaces obtained with the FMDOF-based SfM method (the image in the lower right subfigure part is a zoom on the polyp included in the yellow rectangle). (c) 3D mosaic obtained with the DOF-based SfM method which led to the surface with the largest extend. (d) Surface reconstructed with the COLMAP method.

ber of 3D points (1140654 points). The 3D surface in Fig. 5.11(b) was constructed with the proposed FMDOF-based SfM method. Homologous points were provided by the DOF fields for the images with few textures (see the image in the bottom of Fig. 5.11(a)) and by SIFT matches for the remaining images (see the two top images in Fig. 5.11(a)). The SIFT algorithm was able to determine enough matches in 1695 images among the 2468 images of the whole sequence. The surface reconstructed by COLMAP has mainly two drawbacks: one the one hand it includes several gaps ("holes" without textures, see the orange areas in Fig. 5.11(d)) and, on the other hand, the surface is with less extend than the surfaces obtained with the FMDOF and DOF SfM methods. A surface reconstruction as in Fig. 5.11 allows for a second diagnosis (after the endoscopy) by zooming on regions of interest (e.g., on the region with the polyp in Fig. 5.11) of the archived map.

5.2.2.3 Fluorescence cystoscopy

Fig. 5.12(a) presents three cystoscopic images of a video-sequence of 84 frames acquired in the fluorescence modality[3]. While the images in fig. 5.12(a) include textures (blood vessels), numerous images of the sequence are without significant textures (or at least with large regions without textures) and all images are affected by motion blur due to the displacement speed of the endoscope. Fig. 5.12(b) shows that the reconstructed surface part is without gaps and the textures are visually coherent on the 3D surface defined everywhere with a high resolution.

[3]See the video available at https://github.com/CRAN-BioSiS-Imaging/PR2020

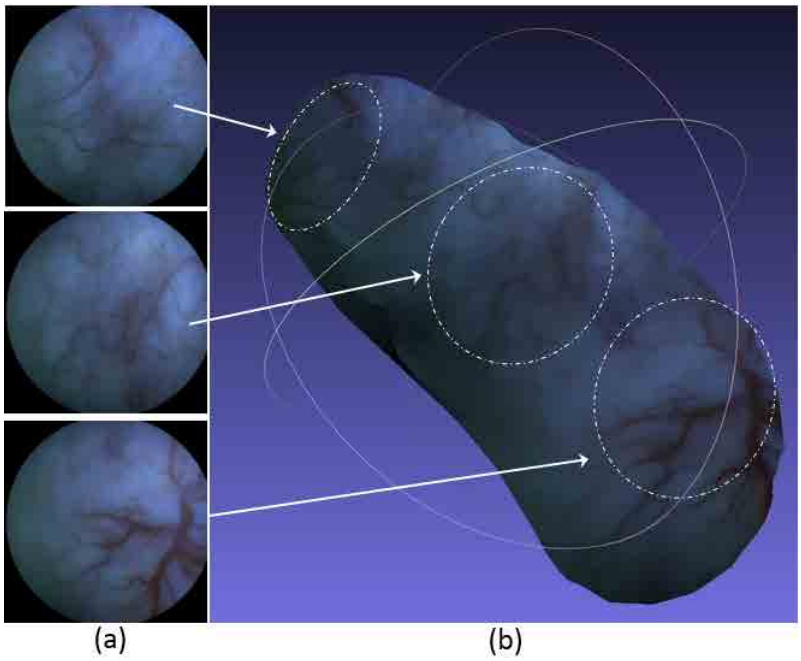


Figure 5.12: Fluorescence cystoscopy surface obtained by the DOF-based SfM method. (a) Three small FoV images. (b) 3D bladder wall: the ellipses delineate the FoV of the three images in (a)

5.2.3 3D skin mosaicing in dermatology

5.2.3.1 Medical context

In dermatology, lesions like pressure ulcers or cancers have to be acquired with a high resolution. Pressure ulcers are usually lesions that are widespread over several images. Besides the fact that extended FoV images are required to represent them with a high resolution, it is also important in dermatology to be able to asses their surface evolution between two examinations. This size evolution assessment is more precise on a 3D surface than on a 2D mosaic. Moreover, in countries as in France, there is a lack of dermatologists in the countryside. A nurse often takes a few images of a wound at the patient's home and transmit them to the dermatologist in the city. An alternative would be to acquire a video-sequence of the interesting skin part and to transmit the data before or after the 3D surface construction. This would also allow for a virtual navigation around the body part under interest without the presence of the patient.

5.2.3.2 Hand and leg surfaces reconstruction

Fig. 5.13 shows the surface construction of a leg part seen in the snapshot given in Fig. 5.13(a). The circle in Fig. 5.13(a) encompasses a small wound which is clearly visible in the top view of the leg given in Fig. 5.13(b). In the bottom view of the leg (see Fig. 5.13(c)), the vertical light gradient is due to the camera viewpoint differences between the first and last images (the number of the last image is 130) which close the loop trajectory required for a 360 degree scan of the leg (see the cross-sectional view of the leg in Fig. 5.13(d)). These colour differences, which do

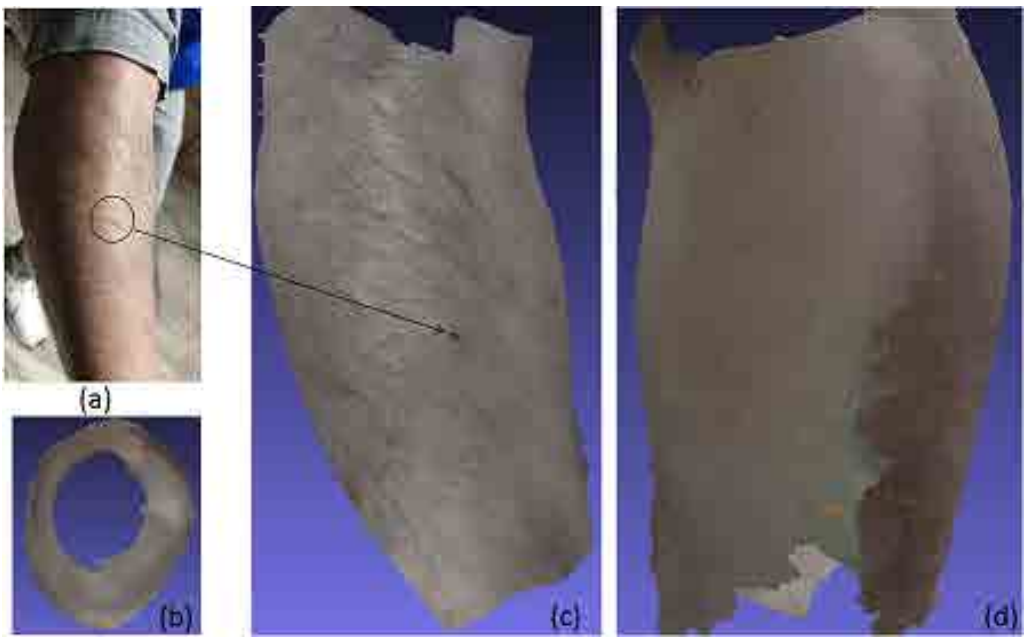


Figure 5.13: Leg reconstruction. (a) Snapshot of the leg and (b), (c), (d) constructed surface under three viewpoints.

not affect the robustness of the proposed SfM-based DOF technique due to the used illumination invariant OF method, can be corrected with various methods (e.g., see the colour correction method in ([WDWR12])).

A reconstruction result is given in Fig. 5.14 for the skin surface of a hand. The FMDOF-based SfM method used 130 images to reconstruct this surface. Among the whole video-sequence, 89 images were exploited by the proposed SfM algorithm to match homologous points using SIFT information. For the remaining 41 images, DOF fields were used to recover a dense point correspondence between image pairs. Increasing the correspondence number does not only allow to avoid the use of a MVS-step, but makes the reconstruction more robust (the SfM performance increases when the number of accurately matched points becomes larger). The shape of the hand is very realistic and each surface part has approximatively the resolution of the images of the corresponding viewpoints. The zoom in Fig. 5.14(c) shows that a small mole can be clearly represented: its colour, shape, and textures are visible by visualizing this surface part.

The reconstruction of the leg and of the hand also shows that the proposed SfM algorithms can deal with different surface geometries. Indeed, the hand in Fig. 5.14 has very different curvature values on the fingers (cylindrical surfaces with a small

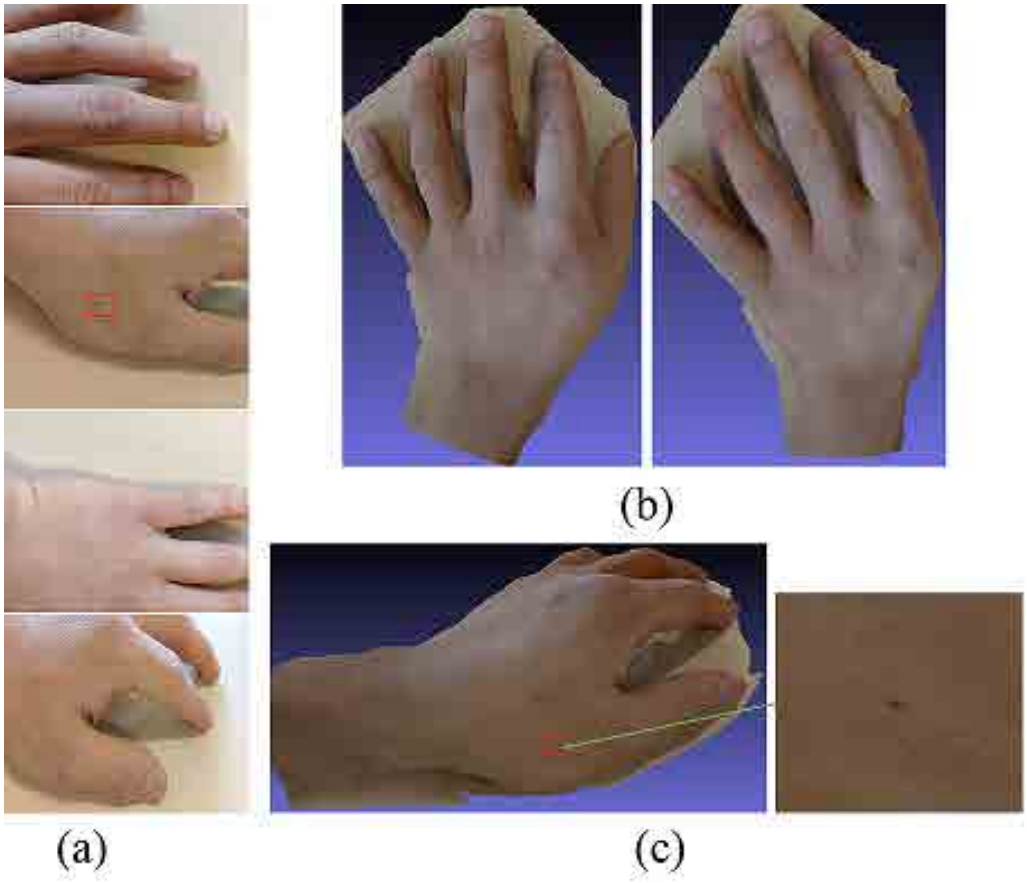


Figure 5.14: Hand surface reconstruction. (a) Four small FoV images of different hand parts. (b) Reconstructed skin surface under two different viewpoints. The hand rests on a table whose planar surface is also partially reconstructed. (c) Zoom on a 3D surface part with a mole.

diameter induce high curvatures and torsions), the back of the hand (almost planar surface), and the beginning of the forearm. On the contrary, the leg in Fig. 5.13 has a smooth shape (with few geometrical disparities) since along the image sequence, the local surfaces have low curvatures, weak curvature changes, and constant curvature signs. The proposed SfM algorithms are able to deal which such different surface types.

5.3 Non-medical scene surface construction

Fig. 5.15 shows the reconstruction of a small kitchen room using a video sequence of 500 high-resolution images acquired with a smartphone. The images have a size of 1080×720 pixels. As seen in Fig. 5.15(b), similarly to an endoscope trajectory inside a hollow organ, the smartphone moves inside the room. Since the video-sequence contains both textured images, and images with less and few contrasted textures (the door and the walls are covered with a plain colour), the FMDOF-based approach was chosen to reconstruct this scene for which the SIFT algorithm determined enough matches for 386 images. DOF-fields were computed for the remaining 114 images.

The global view of the reconstructed surface is shown in Fig. 5.15(b), while some parts of the inner surface are visualized in Fig. 5.15(c). The realistic surface reconstructed for the kitchen room shows the ability of the proposed method to deal also with non-medical scenes.

Similar to gastroscopic or cystoscopic examinations where the intrinsic camera parameters are constant, the smartphone used a constant value for the focal length (the auto-focus option of the smartphone was switched off). The values of the intrinsic parameters were estimated by the SfM approach. That shows again the flexibility of SfM methods when they use uncalibrated cameras.

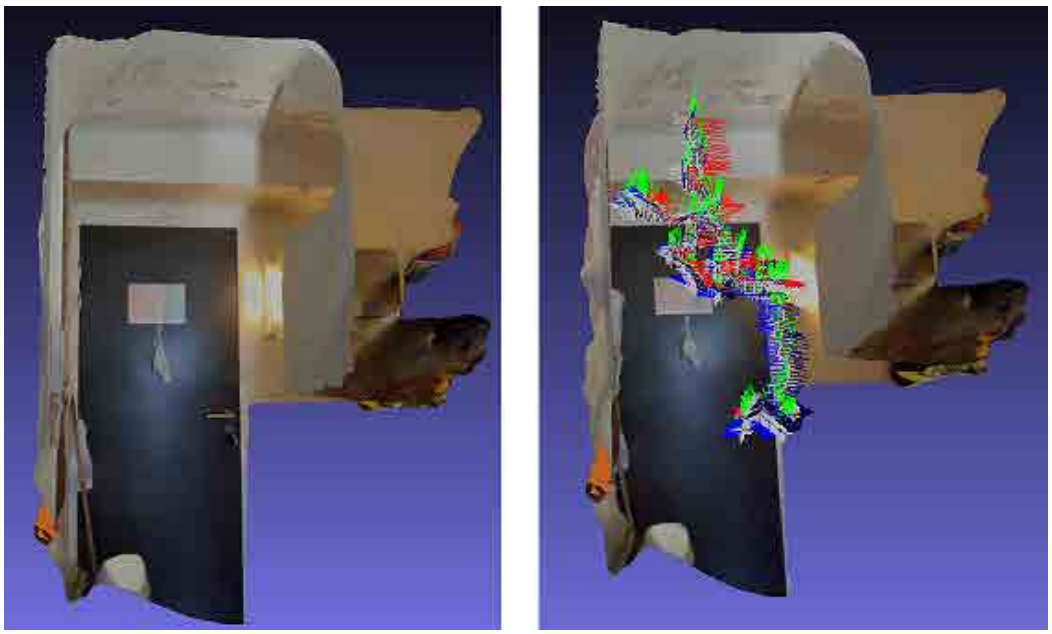
5.4 Main contributions and conclusion

In the medical context of this work, the purpose of the proposed SfM methods were not to construct very precise surfaces because hollow organs have never the same shape between two examinations or from one patient to another. However, the 3D shape must be consistent (in accordance with the anatomy of the organ), without discontinuities of textures, structures or colours, as well as with an acceptable resolution regardless of the location observed on the surface.

In subsection 5.1.3, the results on phantoms show that the precision of the proposed SfM methods can closely approach that of two state-of-the-art methods (COLMAP and VisualSfM) based on the detection of SIFT-features that locate the matched points with a sub-pixel accuracy. In this subsection, it was also shown that the use of an uncalibrated camera is appropriate to ensure a high surface reconstruction accuracy.

In section 5.2, surface construction tests were presented on various patient data. Surfaces with almost no textures (gastroscopy), with rather few textures (cys-

(a)



(b)

(c)

Figure 5.15: Surface construction of a small kitchen room. (a) Three different images among the 500 images of the video-sequence. (b) The image on the left shows the constructed surface, while the right image represents the camera poses along the smartphone (camera) trajectory. (c) Virtual navigation in the surface: two different viewpoints inside the surface represented in (b).

toscopy) and with more textures (dermatology) were successfully reconstructed with the proposed SfM methods. These surfaces were reconstructed for uncontrolled camera trajectories and under strongly varying illumination conditions. Besides that, the reconstruction of a room in section 5.3 shows that the proposed SfM methods

can also deal with non medical scenes.

Moreover, surface construction tests were conducted for very different imaging modalities, i.e., white light for all examination, NBI in gastroscopy and fluorescence in cystoscopy. One point highlighting the robustness of the proposed methods relates to the fact that all surface construction test were performed with constant algorithm parameters (those given in Table 3.4 and Section 4.3), whatever the medical examination, the textures, the illumination changes and the image modality.

The SfM algorithms proposed in the literature were designed for textured scenes and were most often tested on sets of images acquired with rather controlled acquisition conditions: different viewpoints ensuring large parallax values, moderate illumination changes, large and common object parts seen in several images, etc. In endoscopy, the acquisition conditions are by far more uncontrolled and the scenes are often more complex. Besides the lack of textures, the geometrical disparity in the images is low (smooth surfaces), the illumination conditions are strongly changing and the camera trajectory is less controllable so that numerous common regions are seen in pairs of images acquired from similar viewpoints. The results given in this chapter (and published in [PTWD20], [PTL+19b], [PTL+20], and [PTL+19a]) show that computing illumination invariant OF fields, and an appropriate choice of reference images (i.e. with a high number of overlapped images) well spread over the surfaces allows to adapt SfM algorithms to challenging scenes which were barely under consideration in the literature.

List of publications

International journal

- T.-B. Phan, D.-H.Trinh, D. Wolf, C. Daul. Optical flow-based structure-from-motion for the reconstruction of epithelial surfaces. *Pattern Recognition,* Volume 105, 107391, September 2020.

International conference

- T.-B. Phan, D.-H. Trinh, D. Lamarque, D. Wolf, and C. Daul. Dense optical flow for the reconstruction of weakly textured and structured surfaces: Application to endoscopy. IEEE International Conference on Image Processing (ICIP), pages 310-314, Taipei, Taiwan, 2019.

International conference (acceptance on abstract)

- Visualization of extended epithelial tissue surfaces using dense optical flow and structure from motion. T.-B. Phan, D.-H. Trinh, D. Lamarque, W. Blondel, M. Amouroux, D. Wolf, C. Daul. 16th International conference on Laser Applications in Life Sciences. To be held in Nancy, France in 2021 (originally scheduled for April 2020, this conference has been postponed due to Covid-19).

Conclusion and perspectives

The major contribution presented in this thesis lies in the proposal of two novel SfM methods that can deal with complex scene conditions as those encountering in endoscopy. Both the DOF and the FMDOF algorithms were integrated into a SfM scheme and led to realistic results in terms of the reconstruction of hollow organ surfaces. In this work, the medical scenes were almost rigid, all surface construction tests were performed with constant algorithm parameters and no camera calibration step was required to ensure a robust and realistic shape recovery. Moreover, reconstruction tests with known and unknown intrinsic camera parameters led to similar performances in terms of accuracy and robustness. Thus, both the DOF- and the FMDOF-based SfM algorithms have a practical interest since during medical procedures, endoscope calibrations should be avoided (medical examinations should remain unchanged). The proposed 3D reconstruction pipeline is also able to deliver directly a dense 3D point cloud without any multiview stereo algorithm. The latter is an essential step in classical SfM pipelines when dense point clouds are required.

The results obtained for textured phantoms have shown that the precision of the proposed SfM methods can closely approach that of two state-of-the-art methods based on the detection of features. It was important to objectively evaluate the inherent accuracy of the proposed methods before applying them to the real medical scenes which are without known ground truth.

The proposed SfM solutions provided relevant results for very different scene contents and acquisition conditions. In this work, surface construction tests were presented on different data. Surfaces with almost no textures (gastroscopy), with rather few textures (cystoscopy) and with more textures (dermatology) were successfully reconstructed with the proposed SfM methods. These surfaces were reconstructed for hardly controllable camera trajectories and under strongly varying illumination conditions. Moreover, surface construction tests were conducted for very different imaging modalities, i.e. white light for all medical applications, narrow band imaging in gastroscopy, and fluorescence in cystoscopy. Besides the medical scenes, the reconstruction of a small kitchen room performed in Chapter 5 shows the appropriateness of the proposed SfM methods for non-medical scenes.

From the scientific point of view, optical flow was rarely used in SfM due to the accumulating errors when tracking points along consecutive images. This thesis proposes an original strategy to exploit flow fields based on the selection of reference images which allow both for the establishment of numerous and large homologous point groups, and for the construction of surfaces without gaps. The OF fields

between a reference image and each of its overlapped image avoids the tracking of points along the video-sequence. Moreover, this work led to an accurate OF estimation using a new illumination-invariant descriptor to deal with scenes with few textures/structures, and affected by strong illumination changes. All these facets of the DOF algorithm detailed in Chapter 3 not only explain why it is possible to determine numerous groups of homologous points, even without contrasted textures and structures, but also contribute to the fact that at least some homologous points of a group are located in images taken from quite different viewpoints. Such point groups ensure an accurate and robust surface construction with the proposed DOF-based SfM method.

Besides that, a robust FMDOF method was proposed by combining a feature matching algorithm (SIFT) and the DOF-field computation to simultaneously exploit the advantages of both methods for generating large 2D point groups. SIFT is used for the image regions with contrasted textures, while DOF is employed for the other regions (without textures). It is worth noticing that when SIFT works well in all images, the proposed FMDOF reconstruct surfaces in an almost similar way as COLMAP, and, on the contrary, if SIFT is always inoperative, the FMDOF method is equivalent to the DOF method. The FMDOF-based SfM is efficient for reconstructing the scenes whose texture amount and contrast vary between or inside images, as it is the case for the skin and cystoscopic scenes shown in Chapter 5.

From the medical point of view, the proposed SfM methods can reconstruct surfaces acquired for various endoscopic scenes and imaging modalities. The extended FOV images facilitate the detection of abnormal regions (e.g., with polyps) or inflammations (e.g., inflammations around the pyloric antrum region of the stomach). Besides that, the surfaces of the same region reconstructed by the proposed SfM algorithms for two or more examinations help to diagnose a lesion evolution or to assess the remission of a tissue after surgery for instance. For the tested medical applications, the described 3D reconstruction algorithms led systematically to consistent 3D shapes (in accordance with the anatomy of the organ), without discontinuities of textures or structures, as well as with an acceptable resolution regardless of the location observed on the surface. Textured 3D images of the internal organ wall surfaces also support the exchange of information between physicians of different specialties.

Perspectives

In endoscopy, the quality of numerous images is affected by defocusing/refocusing, motion blur, floating objects, etc. The proposed SfM algorithm should be associated with a more complete preprocessing step to improve the image selection [AZB+20]. Such a combination of an elaborated image selection procedure and the SfM scheme will contribute to the automation of the endoscopic surface reconstruction (the image sequence parts were manually chosen in this work).

From the informatics point of view, the algorithms can considerably speed-up by rewriting the code completely in C++, optimizing it, and parallelizing the code.

Recently, the work presented in [MWP+19] has dealt with the real-time visualization of colon chunk surfaces (small parts of the colon) by associating a SfM and a SLAM method. The authors used a recurrent neural network (RNN) to predict depth maps and camera poses for colonoscopic image sequences, the ground truth data (small textured colon surface parts) for the network training being determined with the COLMAP SfM method. The RNN was integrated into a standard SLAM approach to correct the depth maps and camera poses predicted by the RNN network. A global texture mesh is finally achieved after the fusion of the depth maps of colonoscopic frames. The results obtained in [MWP+19] are impressive in the sense that the method achieves real-time reconstructions and allows for the visualization of potentially missing (unscanned or occluded) regions. However, the length of the reconstructed colon parts remains limited due to the ground truth data used in the training process of the RNN which predicts the depth maps and camera poses. Indeed, feature-based SfM methods as COLMAP reconstruct complex colonoscopic scenes with a moderate accuracy (as in the stomach, there is often a lack of textures in numerous images). In the future, the proposed DOF-based SfM method could be used to provide realistic ground truths (organ surfaces) for improving the training of the neural network and to increase the accuracy of the depth map and camera pose prediction in the reconstruction approach proposed in [MWP+19]. Such RNN-based methods could be appropriate for both the colon and the stomach.

A natural extension of the proposed reconstruction algorithms would be to adapt them to non-rigid scenes. The non-rigid scene that will be considered in a close work is located at the junction of the cardia (upper part of the stomach) and of the bottom of the oesophagus. At this junction, the cardiac sphincter muscle acts as a valve that opens and closes itself. When the gastroscope is in the oesophagus and points towards the cardia, it observes an approximately cylindrical shape whose bottom opens and closes. In the case of chronic gastric reflux (acid reflux from the stomach into the oesophagus), there is a change in the colour of the lower oesophageal tissues (pink "trails" rise up on the healthy white tissue without inflammation). The movements due to the cardiac sphincter muscle deforms this area and it is difficult to observe this surface in order to quantify the Barrett's oesophagus. In this context, the aim of a non-rigid structure from motion (NRSfM) algorithm would not be to extend the FoV, but rather to construct a video of a surface whose shape changes with the time. With such a 3D video, the gastroenterologist can choose the surface state which allows him to observe the extent of the Barrett's oesophagus.

A NRSfM method can be used to recover, for each 2D image of a video sequence (i.e. at a given time), the shape and the position of the surface using a set of homologous 2D points seen during the video-sequence which visualizes the deformations of the surface. NRSfM methods are well-known as being an ill-posed problem due to the non-rigidity of the surfaces. In consequence, no general NRSfM method can be developed for all of deformable scenes. In fact, assumptions (the type of deformation is exactly known [PPB18, FAD10]), model simplifications (orthogonal cameras [BHB00, ASKK11]) or restrictions (the shape to be reconstructed can be described on a low-dimensional subspace [THB08]) are usually required to solve the equations relating to the NRSfM method. Three main "criteria" have to

be taken into account when developing a NRSfM method: can the deformable model be statistically or physically described, should the camera model be represented by an orthogonal projection, with a weak perspective projection or a full perspective projection, and what type of missing data must be handled by the reconstruction method. A detailed description and classification of NRSfM methods, as well as their benchmarks can be found in a recent survey paper [JDDA18]. In the frame the Barrett's oesophagus construction, it will be assumed that a perspective camera model can be implemented in the NRSfM scheme in order to reconstruct precise surfaces (most often orthogonal cameras are used to simplify the equations). The group of homologous points taken as input by the NRSfM method will be computed by the DOF method described in this thesis [PTWD20] since only few textures are available at the bottom of the oesophagus (for NRSfM methods, it is also crucial to match homologous points in a robust way). A self-calibration technique [HZ04] and the shape trajectory model described in [GM11] will be exploited to propose an algorithm to determine the video of the 3D deformable shape of the Barrett's oesophagus.

Bibliography

[AAT12] C. Aholt, S. Agarwal, and R. R. Thomas. A QCQP approach to triangulation. In *Computer Vision - ECCV 2012 , Proceedings, Part I*, volume 7572 of *Lecture Notes in Computer Science*, pages 654–667. Springer, 2012.

[ABD12] P.F. Alcantarilla, A. Bartoli, and A.J. Davison. KAZE features. In *European Conference on Computer Vision (ECCV)*, volume 7577, LNCS, pages 214–227, 2012.

[ABD13] S. Ali, W. Blondel, and C. Daul. Tv-L1 based fast and robust mosaicing of cystoscopic images. In *XXIVe Colloque GRETSI Traitement du Signal et des Images, Brest, France*, 2013.

[ADG+16] S. Ali, C. Daul, E. Galbrun, F. Guillemin, and W. Blondel. Anisotropic motion estimation on edge preserving riesz wavelets for robust video mosaicing. *Pattern Recognition*, 51:425–442, 2016.

[ADGB16] S. Ali, C. Daul, E. Galbrun, and W. Blondel. Illumination invariant optical flow using neighborhood descriptors. *Computer Vision and Image Understanding*, 145:95–110, 2016.

[ADWB13] S. Ali, C. Daul, T. Weibel, and W. Blondel. Fast mosaicing of cystoscopic images from dense correspondence: Combined SURF and TV-L1 optical flow method. In *IEEE International Conference on Image Processing, ICIP, Australia*, pages 1291–1295, 2013.

[Ali16] Sharib Ali. *Total variational optical flow for robust and accurate bladder image mosaicing.* PhD thesis, University of Lorraine, Nancy, France, 2016.

[AMN+98] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM*, 45(6):891–923, 1998.

[ASKK11] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(7):1442–1456, 2011.

[ASS08] S. Agarwal, N. Snavely, and S. M. Seitz. Fast algorithms for l∞ problems in multiview geometry. In *IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition, CVPR 2008, Anchorage, Alaska, USA. IEEE Computer Society, 2008.

[ASS+09] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *IEEE 12th International Conference on Computer Vision, (ICCV) 2009*, pages 72–79. IEEE Computer Society, 2009.

[AT17] R. M. Artal and J. D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Trans. Robotics*, 33(5):1255–1262, 2017.

[AZB+20] S. Ali, F. Zhou, B. Braden, A. Bailey, S. Yang, G. Cheng, P. Zhang, X. Li, M. Kayser, R. D Soberanis-Mukul, et al. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Scientific reports*, 10(1):1–15, 2020.

[BBPW04] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *8th European Conference on Computer Vision*, pages 25–36, 2004.

[BDS16] Achraf Ben-Hamadou, Christian Daul, and Charles Soussen. Construction of extended 3D field of views of the internal bladder wall surface: a proof of concept. *3D Research*, 7(3):95:1–95:23, September 2016.

[BETV08] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[BFB94] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *Int. J. Comput. Vis.*, 12(1):43–77, 1994.

[BGCC12] A. Bartoli, Y. Gérard, F. Chadebecq, and T. Collins. On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2026–2033. IEEE Computer Society, 2012.

[BHB00] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2000*, pages 2690–2696, 2000.

[BHSD+13] A. Ben-Hamadou, C. Soussen, C. Daul, W. blondel, and D. Wolf. Flexible calibration of structured-light systems projecting point patterns. *Computer Vision and Image Understanding*, 117(10):1468–1481, Oct. 2013.

[Bou] J. Y Bouguet. Camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/.

[BP92] M. Bichsel and A. Pentland. A simple algorithm for shape from shading. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, pages 459–465. IEEE, 1992.

[BRPM16] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Proceedings of the British Machine Vision Conference, BMVC 2016*. BMVA Press, 2016.

[BRSF09] J. P. Barreto, J. Roquette, P. F. Sturm, and F. Fonseca. Automatic camera calibration applied to medical endoscopy. In *Proceedings of the British Machine Vision Conference, BMVC*, pages 1–10. British Machine Vision Association, 2009.

[BS06] C. Beder and R. Steffen. Determining an initial image pair for fixing the scale of a 3D reconstruction from an image sequence. In *Pattern Recognition, 28th DAGM Symposium*, volume 4174, pages 657–666. Springer, 2006.

[BSGA09] A. Behrens, T. Stehle, S. Gross, and T. Aach. Local and global panoramic imaging for fluorescence bladder endoscopy. In *31th Annu. Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6990–6993, Minneapolis, 2009.

[BSL+11] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *Int. J. Comput. Vis.*, 92(1):1–31, 2011.

[BTS15] C. Bailer, B. Taetz, and D. Stricker. Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4015–4023, 2015.

[CA12] T. Collins and A.Bartoli. Towards live monocular 3D laparoscopy using shading and specularity information. In *IPCAI 2012, Pisa, Italy, June 27, 2012. Proceedings*, volume 7330 of *Lecture Notes in Computer Science*, pages 11–21. Springer, 2012.

[Can86] John F. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.

[CBA+19] R. J. Chen, T. L. Bobrow, T. Athey, F. Mahmood, and N. J. Durr. SLAM endoscopy enhanced by adversarial depth prediction. *CoRR*, abs/1907.00283, 2019.

[CLSF10] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: binary robust independent elementary features. In *European Conference on Computer Vision ECCV 2010*, volume 6314 of *Lecture Notes in Computer Science*, pages 778–792. Springer, 2010.

[COSH13] D. J. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher. SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2841–2853, Dec 2013.

[CP11] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[CS07] R. E. Carroll and S. M. Seitz. Rectified surface mosaics. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE Computer Society, 2007.

[CT90] B. Caprile and V. Torre. Using vanishing points for camera calibration. *Int. J. Comput. Vis.*, 4(2):127–139, 1990.

[CT15] Z. Cui and P. Tan. Global structure-from-motion by similarity averaging. In *IEEE International Conference on Computer Vision (ICCV) 2015, Santiago, Chile, December*, pages 864–872. IEEE Computer Society, 2015.

[CTS95] J. E. Cryer, P. S. Tsai, and M. Shah. Integration of shape from shading and stereo. *Pattern Recognit.*, 28(7):1033–1043, 1995.

[Dav03] A. J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1403–1410. IEEE Computer Society, 2003.

[Dec] Feature dectectors. https://en.wikipedia.org/wiki/Feature_detection_(computer_vision).

[Die16] J. T. Dietrich. Riverscape mapping with helicopter-based structure-from-motion photogrammetry. *Geomorphology*, 252:144–157, 2016.

[DN13] M. Drulea and S. Nedevschi. Motion estimation using the correlation transform. *IEEE Trans. Image Process.*, 22(8):3260–3270, 2013.

[DRMS07] A. J. Davison, I. D. Reid, N. Molton, and O. Stasse. Monoslam: Real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):1052–1067, 2007.

[DS15] J. Dong and S. Soatto. Domain-size pooling in local descriptors: DSP-SIFT. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 5097–5106. IEEE Computer Society, 2015.

[Enda] Endoscopy: What to know. https://www.medicalnewstoday.com/articles/153737#preparation.

[Endb] Upper endoscopy upper GI procedure nebraska patient education. https://www.youtube.com/watch?v=gpxR1ji4fkw.

[EXI] Exif tags. https://exiftool.org/TagNames/EXIF.html.

[FAD10] João Fayad, Lourdes Agapito, and Alessio Del Bue. Piecewise quadratic reconstruction of non-rigid surfaces from monocular sequences. In *ECCV*, pages 297–310, 2010.

[FAT11] S. Foix, G. Alenya, and C. Torras. Lock-in time-of-flight (tof) cameras: A survey. *IEEE Sensors Journal*, 11(9):1917–1926, 2011.

[Fau93] Olivier Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, Cambridge, MA, USA, 1993.

[FB81] M. A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[FBK15] D. Fortun, P. Bouthemy, and C. Kervrann. Optical flow modeling and computation: A survey. *Comput. Vis. Image Underst.*, 134:1–21, 2015.

[Fea] Local feature detection and extraction. https://fr.mathworks.com/help/vision/ug/local-feature-detection-and-extraction.html.

[FFG09a] M. Farenzena, A. Fusiello, and R. Gherardi. Structure-and-motion pipeline on a hierarchical cluster tree. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1489–1496, 2009.

[FFG09b] M. Farenzena, A. Fusiello, and R. Gherardi. Structure-and-motion pipeline on a hierarchical cluster tree. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1489–1496. IEEE Computer Society, 2009.

[FGG+10] J. M. Frahm, P. F. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y. H. Jen, E. Dunn, B. Clipp, and S. Lazebnik. Building rome on a cloudless day. In *European Conference on Computer Vision ECCV*, volume 6314 of *Lecture Notes in Computer Science*, pages 368–381. Springer, 2010.

[FLM92] O. D. Faugeras, Q-T. Luong, and S. J. Maybank. Camera self-calibration: Theory and experiments. In *Computer Vision - ECCV'92*, volume 588 of *Lecture Notes in Computer Science*, pages 321–334. Springer, 1992.

[FP10] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Analysis Machine Intelligence*, 32(8):1362–1376, 2010.

[FT00] Carlos Henrique Quartucci Forster and Clésio L. Tozzi. Towards 3D reconstruction of endoscope images using shape from shading. In *(SIBGRAPI 2000)*, pages 90–96. IEEE Computer Society, 2000.

[FYY+18] Y. Fu, Q. Yan, L. Yang, J. Liao, and C. Xiao. Texture mapping for 3D reconstruction with RGB-D sensor. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22*, pages 4645–4653. IEEE Computer Society, 2018.

[GBC+14] O. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. M. M. Montiel. Visual SLAM for handheld monocular endoscope. *IEEE Trans. Medical Imaging*, 33(1):135–146, 2014.

[GBDH95] P. Graebling, C. Boucher, Ch. Daul, and E. Hirsch. 3D sculptured surface analysis using a structured-light approach. In *Videometrics IV*, volume 2598, pages 128 – 139. SPIE, 1995.

[GFF10] R. Gherardi, M. Farenzena, and A. Fusiello. Improving the efficiency of hierarchical structure-and-motion. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA*, pages 1594–1600. IEEE Computer Society, 2010.

[GHT11] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *Int. J. Comput. Vis.*, 94(3):335–360, 2011.

[GL96] Gene H. Golub and Charles F. Van Loan. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, 1996.

[GM11] Paulo F. U. Gotardo and Aleix M. Martínez. Non-rigid structure from motion with complementary rank-3 spaces. In *CVPR*, pages 3065–3072, 2011.

[GMBR10] A. Gil, O.M. Mozos, M. Ballesta, and O. Reinoso. A comparative evaluation of interest point detectors and local descriptors for visual SLAM. *Mach. Vis. Appl.*, 21(6):905–920, 2010.

[GRA13] R. Garg, A. Roussos, and L. Agapito. A variational approach to video registration with subspace constraints. *Int. J. Comput. Vis.*, 104(3):286–314, 2013.

[HÅ97] A. Heyden and K. Åström. Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 97*, pages 438–443. IEEE Computer Society, 1997.

[Har92] R. I. Hartley. Estimation of relative camera positions for uncalibrated cameras. In *Computer Vision - ECCV'92*, volume 588 of *Lecture Notes in Computer Science*, pages 579–587. Springer, 1992.

[HM18] Sven Haase and Andreas Maier. Endoscopy, 2018.

[HMB+13] L. M. Hein, P. Mountney, A. Bartoli, H. Elhawary, D. S. Elson, A. Groch, A. Kolb, M. A. Rodrigues, J. M. Sorger, S. Speidel,

and D. Stoyanov. Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery. *Medical Image Analysis*, 17(8):974–996, 2013.

[Hor75] B. Horn. *Obtaining shape from shading information.* The Psychology of Computer Vision. McGraw-Hill, New York, 1975.

[HS81] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[HS88] C. G. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the Alvey Vision Conference, AVC 1988, Manchester, UK, September, 1988*, pages 1–6. Alvey Vision Club, 1988.

[HS97] R. I. Hartley and P.F. Sturm. Triangulation. *Comput. Vis. Image Underst.*, 68(2):146–157, 1997.

[HS14] M. Havlena and K. Schindler. Vocmatch: Efficient multiview correspondence for structure from motion. In *European Conference on Computer Vision ECCV*, volume 8691 of *Lecture Notes in Computer Science*, pages 46–60. Springer, 2014.

[HSDF15] J. Heinly, J. L. Schönberger, E. Dunn, and J.M. Frahm. Reconstructing the world in six days. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3287–3295, Boston, MA, USA, 2015.

[HSL16] Y. Hu, R. Song, and Y. Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, Boston, MA, USA, 2016.

[HTP10] M. Havlena, A. Torii, and T. Pajdla. Efficient structure from motion by graph optimization. In *11th European Conference on Computer Vision ECCV 2010*, pages 100–113, 2010.

[HZ04] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2004.

[JDDA18] Sebastian Hoppe Nesgaard Jensen, Alessio Del Bue, Mads Emil Brix Doest, and Henrik Aanæs. A benchmark and evaluation of non-rigid structure from motion. *CoRR*, abs/1801.08388, 2018.

[JP11] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3121–3128, Colorado Springs, CO, USA, June 2011.

[JR12] M.R. James and S. Robson. Straightforward reconstruction of 3D surfaces and topography with a camera: Accuracy and geoscience application. *Journal of Geophysical Research*, 117(F03017):1–17, 2012.

[KBH06] M. M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, pages 61–70, 2006.

[KCCH11] B. W. Kuo, H. H. Chang, Y. C. Chen, and S. Y. Huang. A light-and-fast SLAM algorithm for robots in indoor environments using line segment map. *J. Robotics*, 2011:257852:1–257852:12, 2011.

[KCT+18] A. Krebs, V. Camilo, E. Touati, Y. Benezeth, V. Michel, G. Jouvion, F. Yang, D. Lamarque, and F. Marzani. Detection of h. pylori induced gastric inflammation by diffuse reflectance analysis. In *IEEE 18th (BIBE)*, pages 287–292, 2018.

[Kie] Dang Trung Kien. A review of 3D reconstruction from video sequences.

[KM07] G. Klein and D. W. Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR*, pages 225–234. IEEE Computer Society, 2007.

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012.

[KW08] A. E. Kaufman and J. Wang. 3D surface reconstruction from endoscopic videos. In *Visualization in Medicine and Life Sciences*, pages 61–74. Springer, 2008.

[LAZ+17] K. L. Lurie, R. Angst, D. V. Zlatev, J. C. Liao, and A. K. Ellerbee Bowden. 3D reconstruction of cystoscopy videos for comprehensive bladder records. *Biomedical Optics Express*, 8(4):2106–2123, 2017.

[LCHS03] L. Zhang, Curless, Hertzmann, and Seitz. Shape and motion under varying illumination: unifying structure from motion, photometric stereo, and multiview stereo. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–625 vol.1, 2003.

[LDB+08] R. M. Luna, C. Daul, W. Blondel, Y. Hernandez-Mier, D. Wolf, and F. Guillemin. Mosaicing of bladder endoscopic image sequences: Distortion calibration and registration algorithm. *IEEE Trans. Biomed. Engineering*, 55(2):541–553, 2008.

[LH81] H.C. Longuet Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293, 1981.

[Lin98] Tony Lindeberg. Edge detection and ridge detection with automatic scale selection. *Int. J. Comput. Vis.*, 30(2):117–156, 1998.

[LNF09] V. Lepetit, Fr. M. Noguer, and P. Fua. Epnp: An accurate $O(n)$ solution to the pnp problem. *Int. J. Comput. Vis.*, 81(2):155–166, 2009.

[Low04] D. G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LQ00] M. Lhuillier and L. Quan. Robust dense matching using local and global geometric constraints. In *15th International Conference on Pattern Recognition, ICPR*, pages 1968–1972. IEEE Computer Society, 2000.

[LR85] C.-H Lee and A. Rosenfeld. Improved methods of estimating shape from shading using the light source coordinate system. *Artif. Intell.*, 26(2):125–143, 1985.

[LSKK10] M. Lindner, I. Schiller, A. Kolb, and R. Koch. Time-of-flight sensor calibration for accurate range sensing. *Comput. Vis. Image Underst.*, 114(12):1318–1328, 2010.

[Mat] Mathworks. Pinhole camera model. https://fr.mathworks.com/help/vision/ug/camera-calibration.html.

[MBC11] A. Malti, A. Bartoli, and T. Collins. Template-based conformal shape-from-motion from registered laparoscopic images. In *Medical Image Understanding and Analysis - MIUA 2011*, pages 227–232. BMVA, 2011.

[MBD+04] R. Miranda-Luna, W. Blondel, C. Daul, Y. Hernandez-Mier, R. Posada, and D. Wolf. A simplified method of endoscopic image distortion correction based on grey level registration. In *International Conference on Image Processing, ICIP*, pages 3383–3386, Singapore, 2004.

[MC99] P.R. S. Mendonça and R. Cipolla. A simple technique for self-calibration. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 99, 23-25 June 1999, Ft. Collins, CO, USA*, pages 1500–1506. IEEE Computer Society, 1999.

[MCH+16] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, and J. M. M. Montiel. Orbslam-based endoscope tracking and 3D reconstruction. In *MICCAI*, volume 10170 of *Lecture Notes in Computer Science*, pages 72–83. Springer, 2016.

[MGV09] T. Moons, L. V. Gool, and M. Vergauwen. 3D reconstruction from multiple images: Part 1 - principles. *Found. Trends Comput. Graph. Vis.*, 4(4):287–404, 2009.

[MKA+11] O. E. Meslouhi, M. Kardouchi, H. Allali, T. Gadi, and Y. A. Benkaddour. Automatic detection and inpainting of specular reflections for colposcopic images. *Central Europ. J. Computer Science*, 1(3):341–354, 2011.

[MMM13] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3248–3255, 2013.

[MMMO] P. Moulon, P. Monasse, R. Marlet, and Others. Openmvg. an open multiple view geometry library. https://github.com/openMVG/openMVG.

[Mor78] Jorge J. Moré. The levenberg-marquardt algorithm: Implementation and theory. In *Numerical Analysis*, pages 105–116, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg.

[MRG98] M.Pollefeys, R.Koch, and L. V. Gool. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *IEEE International Conference on Computer Vision (ICCV)1998*, pages 90–95, 1998.

[MRM+14] M. A. Mohamed, H. A. Rashwan, B. Mertsching, M. García, and D. Puig. Illumination-robust optical flow using a local directional pattern. *IEEE Trans. Circuits Syst. Video Techn.*, 24(9):1499–1508, 2014.

[MS04] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vis.*, 60(1):63–86, 2004.

[MWP+19] R. Ma, R. Wang, S. M. Pizer, J. G. Rosenman, S. K. McGill, and J.-M. Frahm. Real-time 3D reconstruction of colonoscopic surfaces for determining missing regions. In *MICCAI*, volume 11768 of *Lecture Notes in Computer Science*, pages 573–582. Springer, 2019.

[NE86] H.H. Nagel and W. Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(5):565–593, 1986.

[Nis04] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(6):756–777, 2004.

[NLB19] S. Nousias, M. I. A. Lourakis, and C. Bergeles. Large-scale, metric structure from motion for unordered light fields. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3292–3301. Computer Vision Foundation, 2019.

[NLD11] R. A. Newcombe, S. Lovegrove, and A. J. Davison. DTAM: dense tracking and mapping in real-time. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2320–2327. IEEE Computer Society, 2011.

[NS06] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006*, pages 2161–2168. IEEE Computer Society, 2006.

[NSR06] N.Snavely, S.M. Seitz, and R.Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Trans. Graph.*, 25(3):835–846, 2006.

[OD97] T. Okatani and K. Deguchi. Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center. *Comput. Vis. Image Underst.*, 66(2):119–131, 1997.

[Par] POSTU 2014 - Paris. Prevention and screening for stomach cancer. https://www.fmcgastro.org/textes-postus/no-postu_year/prevention-et-depistage-du-cancer-de-lestomac/.

[PCI+07] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, 2007.

[Pen88] A. Pentland. Shape information from shading: A theory about human perception. In *IEEE International Conference on Computer Vision (ICCV)*, pages 404–413. IEEE, 1988.

[PFFK12] V. B. Surya Prasath, I. N. Figueiredo, P. N. Figueiredo, and K.Palaniappan. Mucosal region detection and 3D reconstruction in wireless capsule endoscopy videos using active contours. In *IEEE EMBC*, pages 4014–4017. IEEE, 2012.

[PHS+09] J. Penne, K. Höller, M. Stürmer, T. Schrauder, A. Schneider, R. Engelbrecht, H. Feußner, B. Schmauss, and J. Hornegger. Time-of-flight 3-D endoscopy. In *MICCAI*, volume 5761 of *Lecture Notes in Computer Science*, pages 467–474. Springer, 2009.

[PMSE12] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice. Automatic targetless extrinsic calibration of a 3D lidar and camera by maximizing mutual information. In *Proceedings of the Twenty-Sixth AAAI , Toronto, Ontario, Canada*. AAAI Press, 2012.

[PPB18] Shaifali Parashar, Daniel Pizarro, and Adrien Bartoli. Isometric non-rigid shape-from-motion with riemannian geometry solved in linear time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2442–2454, 2018.

[PTL+19a] T.-B. Phan, D.-H. Trinh, D. Lamarque, D. Wolf, and C. Daul. 3D surface reconstruction using dense optical flow combined to feature matching: Application to endoscopy. In *XXVIIe Colloque en Traitement du Signal et des Images (GRETSI)*, Lille, France, 2019.

[PTL+19b] T.-B. Phan, D.-H. Trinh, D. Lamarque, D. Wolf, and C. Daul. Dense optical flow for the reconstruction of weakly textured and structured surfaces: Application to endoscopy. In *International Conference on Image Processing, ICIP*, pages 310–314, 2019.

[PTL+20] T.-B. Phan, D.-H. Trinh, D. Lamarque, W. Blondel, M. Amouroux, D. Wolf, and C. Daul. Visualization of extended epithelial tissue surfaces using dense optical flow and structure from motion. In *16th International conference on Laser Applications in Life Sciences*, 2020.

[PTWD20] T.-B. Phan, D.-H. Trinh, D. Wolf, and C. Daul. Optical flow-based structure-from-motion for the reconstruction of epithelial surfacess. *Pattern Recognition*, 105:107391, 2020.

[QR18] L. Qiu and H. Ren. Endoscope navigation and 3D reconstruction of oral cavity by visual SLAM with mitigated data scarcity. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR Workshops*, pages 2197–2204. IEEE Computer Society, 2018.

[RASC14] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014*, pages 512–519. IEEE Computer Society, 2014.

[RBG+19] M. J. Rakotosaona, V. L. Barbera, P. Guerrero, N. J. Mitra, and M. Ovsjanikov. POINTCLEANNET: learning to denoise and remove outliers from dense point clouds. *CoRR*, abs/1901.01060, 2019.

[RBS+12] Sebastian Röhl, Sebastian Bodenstedt, Stefan Suwelack, Hannes Kenngott, Beat P Müller-Stich, Rüdiger Dillmann, and Stefanie Speidel. Dense gpu-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. *Medical physics*, 39(3):1632–1645, 2012.

[RD06] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision ECCV , 9th , Proceedings, Part I*, volume 3951 of *Lecture Notes in Computer Science*, pages 430–443. Springer, 2006.

[RMG+13] H. A. Rashwan, M. A. Mohamed, M. García, B. Mertsching, and D. Puig. Illumination robust optical flow model based on histogram of oriented gradients. In *Pattern Recognition - 35th German Conference, GCPR*, volume 8142 of *Lecture Notes in Computer Science*, pages 354–363. Springer, 2013.

[RRKB11] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski. ORB: an efficient alternative to SIFT or SURF. In *IEEE International Conference on Computer Vision (ICCV) 2011, Barcelona, Spain*, pages 2564–2571. IEEE Computer Society, 2011.

[RSEM09] R. Rossi, X. Savatier, J. Y. Ertaud, and B. Mazari. Real-time 3D reconstruction for mobile robot using catadioptric cameras. In *IEEE ROSE*, pages 104–109. IEEE, 2009.

[RWdS+19] Jérôme Revaud, P. Weinzaepfel, C. R. d. Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger. R2D2: repeatable and reliable detector and descriptor. *CoRR*, abs/1906.06195, 2019.

[SB97] S. M. Smith and J. M. Brady. SUSAN - A new approach to low level image processing. *Int. J. Comput. Vis.*, 23(1):45–78, 1997.

[SBF15] J. L. Schönberger, A. C. Berg, and J.M. Frahm. PAIGE: pairwise image geometry encoding for improved efficiency in structure-from-motion. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1009–1018. IEEE Computer Society, 2015.

[SCD+06] S. M. Seitz, B. Curless, J. Diebel, D. l Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006*, pages 519–528. IEEE Computer Society, 2006.

[Sch18] Johannes Lutz Schönberger. *Robust Methods for Accurate and Efficient 3D Modeling from Unstructured Imagery.* PhD thesis, ETH Zurich, 2018.

[SCS11] S.Leutenegger, M. Chli, and R. Siegwart. BRISK: binary robust invariant scalable keypoints. In *IEEE International Conference on Computer Vision (ICCV) 2011, Barcelona, Spain*, pages 2548–2555. IEEE Computer Society, 2011.

[SF16] J. L. Schönberger and J. M. Frahm. Structure-from-motion revisited. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 4104–4113, 2016.

[SFS+12] N. Shevchenko, J.A. Fallert, H. Stepp, H. Sahli, A. Karl, and T.C. Lueth. A high resolution bladder wall map: Feasibility study. In *34th Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, pages 5761–5764, San Diego, CA, Aug. 2012.

[SHSP17] J. L. Schönberger, H. Hardmeier, T. Sattler, and M. Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6959–6968. IEEE Computer Society, 2017.

[SLH06] S. Seshamani, W. W. Lau, and G. D. Hager. Real-time endoscopic mosaicking. In *MICCAI*, volume 4190 of *Lecture Notes in Computer Science*, pages 355–363. Springer, 2006.

[SMB00] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *Int. J. Comput. Vis.*, 37(2):151–172, 2000.

[Sna08] Keith N. Snavely. *Scene Reconstruction and Visualization from Internet Photo Collections.* PhD thesis, University of Washington, USA, 2008.

[SPS12] T.D. Soper, M.P. Porter, and E. J. Seibel. Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance. *IEEE Transactions on Biomedical Engineering*, 59(6):1670–1680, June 2012.

[SRB10] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2432–2439, 2010.

[SRB14] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *Int. J. Comput. Vis.*, 106(2):115–137, 2014.

[SS02] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.*, 47(1-3):7–42, 2002.

[SSH+15] C. Sweeney, T. Sattler, T. Höllerer, M. Turk, and M. Pollefeys. Optimizing the viewing graph for structure-from-motion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 801–809, Santiago, Chile, 2015.

[SSPY10] D. Stoyanov, M. V. Scarzanella, P. Pratt, and G. Z. Yang. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *MICCAI*, volume 6361 of *Lecture Notes in Computer Science*, pages 275–282. Springer, 2010.

[ST94] J. Shi and C. Tomasi. Good features to track. In *Conference on Computer Vision and Pattern Recognition, CVPR 1994, 21-23 June, 1994, Seattle, WA, USA*, pages 593–600. IEEE, 1994.

[Ste95] G. P. Stein. Accurate internal camera calibration using rotation, with analysis of sources of error. In *IEEE International Conference on Computer Vision (ICCV) 1995*, pages 230–236. IEEE Computer Society, 1995.

[STF+15] E. S. Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. M. Noguer. Discriminative learning of deep convolutional feature point descriptors. In *IEEE International Conference on Computer Vision (ICCV)*, pages 118–126. IEEE Computer Society, 2015.

[Sto12] D. Stoyanov. Stereoscopic scene flow for robotic assisted minimally invasive surgery. In *MICCAI*, volume 7510 of *Lecture Notes in Computer Science*, pages 479–486. Springer, 2012.

[SYLK18] D. Sun, X. Yang, M.Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, pages 8934–8943. IEEE Computer Society, 2018.

[Sze11] Richard Szeliski. *Computer Vision - Algorithms and Applications*. Texts in Computer Science. Springer, 2011.

[SZFP16] J.L. Schönberger, E. Zheng, J. M. Frahm, and M. Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, pages 501–518, Colorado Springs, CO, USA, October 2016.

[TBD17] D-H. Trinh, W. Blondel, and C. Daul. A general form of illumination-invariant descriptors in variational optical flow estimation. In *International Conference on Image Processing, ICIP*, pages 1263–1267, Beijing, China, 2017.

[TD19] D-H. Trinh and C. Daul. On illumination-invariant variational optical flow for weakly textured scenes. *Computer Vision and Image Understanding*, 179:1–18, February 2019.

[TDBL18] D.-H. Trinh, C. Daul, W. Blondel, and D. Lamarque. Mosaicing of images with few textures and strong illumination changes: Application to gastroscopic scenes. In *International Conference on Image Processing, ICIP 2018*, pages 1263–1267, Athens, Greece, 2018.

[TGFF15] R. Toldo, R. Gherardi, M. Farenzena, and A. Fusiello. Hierarchical structure-and-motion recovery from uncalibrated images. *Comput. Vis. Image Underst.*, 140:127–143, 2015.

[THB08] Lorenzo Torresani, Aaron Hertzmann, and Christoph Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(5):878–892, 2008.

[TK91] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, International Journal of Computer Vision, 1991.

[TM04] A. Talukder and L. H. Matthies. Real-time detection of moving objects from moving vehicles using dense stereo and optical flow. In *IEEE/RSJ*, pages 3718–3725. IEEE, 2004.

[TMHF99] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment - A modern synthesis. In *International Workshop on Vision Algorithms*, pages 298–372, 1999.

[TOF] Tof camera. https://en.wikipedia.org/wiki/Time-of-flight_camera.

[Tri97] Bill Triggs. Autocalibration and the absolute quadric. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 97*, pages 609–614. IEEE Computer Society, 1997.

[Tsa87] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE J. Robotics and Automation*, 3(4):323–344, 1987.

[Tya] Feature-based methods. https://medium.com/analytics-vidhya/introduction-to-feature-detection-and-matching-65e27179885d.

[TZ00] P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Comput. Vis. Image Underst.*, 78(1):138–156, 2000.

[VF10] A. Vedaldi and B. Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 1469–1472. ACM, 2010.

[VLPK12] H.H. Vu, P. Labatut, J.P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(5):889–901, 2012.

[Vu11] H.H Vu. *Large-scale and high-quality multi-view stereo.* PhD thesis, University of Paris-Est, France, 2011.

[WDW+12] T. Weibel, C. Daul, D. Wolf, R. Rösch, and F. Guillemin. Graph based construction of textured large field of view mosaics for bladder cancer diagnosis. *Pattern Recognition*, 45(12):4138–4150, 2012.

[WDWR12] T. Weibel, C. Daul, D. Wolf, and R. Rösch. Contrast-enhancing seam detection and blending using graph cuts. In *Int. Conference on Pattern Recognition (ICPR)*, pages 2732–2735, Japan, 2012.

[Wei13] T. Weibel. *Discrete Energy Minimization Models for Cystoscopic Mapping.* PhD thesis, University of Lorraine, Nancy, France, 2013.

[WKZ+16] K. Wolff, C. Kim, H. Zimmer, C. Schroers, M. Botsch, O. S. Hornung, and A. S. Hornung. Point cloud noise and outlier removal for image-based 3D reconstruction. In *Fourth International Conference on 3D Vision, 3DV*, pages 118–127. IEEE Computer Society, 2016.

[WMG14] M. Waechter, N. Moehrle, and M. Goesele. Let there be color! large-scale texturing of 3D reconstructions. In *European Conference on Computer Vision ECCV 2014 - 13th European Conference on Computer Vision*, pages 836–850, 2014.

[WNJ10] C. Wu, S.G. Narasimhan, and B Jaramaz. A multi-image shape-from-shading framework for near-lighting perspective endoscopes. *International Journal of Computer Vision*, 86(2-3):211–228, Jan. 2010.

[WPZ+17] R. Wang, T. Price, Q. Zhao, J. M. Frahm, J. G. Rosenman, and S. M. Pizer. Improving 3D surface reconstruction from endoscopic video via fusion and refined reflectance modeling. In *Medical Imaging 2017: Image Processing*, volume 10133 in Proc. SPIE, 2017.

[Wu13] C. Wu. Towards linear-time incremental structure from motion. In *IEEE International Conference on 3D Vision (3DV)*, pages 127–134, Seattle, WA, USA, 2013.

[XCJ08] L. Xu, J. Chen, and J. Jia. A segmentation based variational model for accurate optical flow estimation. In *European Conference on Computer Vision ECCV*, volume 5302 of *Lecture Notes in Computer Science*, pages 671–684. Springer, 2008.

[YH15] Y.Furukawa and C. Hernández. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

[YK06] K.J. Yoon and I.S. Kweon. Adaptive support-weight approach for correspondence search. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(4):650–656, 2006.

[YTLF16] K. Moo Yi, E. Trulls, V. Lepetit, and P. Fua. LIFT: learned invariant feature transform. In *European Conference on Computer Vision ECCV*, volume 9910 of *Lecture Notes in Computer Science*, pages 467–483. Springer, 2016.

[YTY99] S. Y. Yeung, H. T. Tsui, and A. Yim. Global shape from shading for an endoscope image. In *MICCAI*, volume 1679 of *Lecture Notes in Computer Science*, pages 318–327. Springer, 1999.

[ZC91] Q. Zheng and R. Chellappa. Estimation of illuminant direction, albedo, and shape from shading. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(7):680–702, 1991.

[ZDFL95] Z. Zhang, R. Deriche, O. D. Faugeras, and Q. T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artif. Intell.*, 78(1-2):87–119, 1995.

[Zha00] Z. Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Analasis Machine Intelligence*, 22(11):1330–1334, 2000.

[ZPB07] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L^1 optical flow. In *29th DAGM Symposium*, pages 214–223, 2007.

[ZPN+16] T. Zhao, Q.and Price, S. Pizer, M. Niethammer, R. Alterovitz, and J. Rosenman. The endoscopogram: A 3D model reconstructed from endoscopic video frames. In *MICCAI*, volume 9900, LNCS, pages 439–447, 2016.

[ZSZ+09] A. Zanchi, F. Salvi, S. Zanchetta, S. Sterlacchini, and G. Guerra. 3D reconstruction of complex geological bodies: Examples from the alps. *Comput. Geosci.*, 35(1):49–69, 2009.

[ZTCS99] R. Zhang, P. S. Tsai, J. E. Cryer, and M. Shah. Shape from shading: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(8):690–706, 1999.

Résumé

Les algorithmes de « Structure from motion » (SfM, structure reconstituée à l'aide du mouvement) représentent un moyen efficace de construction de surfaces 3D étendues à partir des images d'une scène acquise sous différents points de vue. Ces algorithmes déterminent simultanément le mouvement de la caméra et un nuage de points 3D se trouvant à la surface des objets à reconstruire. Les algorithmes SfM classiques utilisent des méthodes de détection et de mise en correspondance de points caractéristiques pour poursuivre les points homologues à travers les séquences d'images, chaque ensemble de points homologues correspondant à un point 3D à reconstruire. Les algorithmes SfM exploitent les correspondances entre des points homologues pour trouver la structure 3D de la scène et les poses successives de la caméra dans un repère monde arbitraire. Il existe différents algorithmes SfM de référence qui peuvent reconstruire efficacement différents types de scènes lorsque les images comportent suffisamment de textures ou de structures. Cependant, la plupart des solutions existantes ne sont pas appropriées, ou du moins pas optimales, lorsque les séquences d'images contiennent peu de textures. Cette thèse propose deux solutions de type SfM basées sur un flot optique dense pour reconstruire des scènes complexes à partir d'une séquence d'images avec peu de textures et acquises sous des conditions d'éclairage changeantes. Il est notamment montré comment un flot optique précis peut être utilisé de manière optimale grâce à une stratégie de sélection d'images qui maximise le nombre et la taille des groupes de points homologues tout en minimisant les erreurs de localisation des points homologues. La précision des méthodes de cartographie 3D est évaluée sur des fantômes avec des dimensions connues. L'intérêt et la robustesse des méthodes sont démontrés sur des scènes médicales complexes en utilisant un jeu de valeurs constantes pour les paramètres des algorithmes. Les solutions proposées ont permis de reconstruire des organes observés dans différents examens (surface épithéliale de la paroi interne de l'estomac, surface épithéliale interne de la vessie et surface de la peau en dermatologie) et dans diverses modalités (lumière blanche pour tous les examens, lumière vert-bleu en gastroscopie et fluorescence en cystoscopie).

Mots-clés : structure from motion, flot optique dense, mosaïquage 3D, endoscopie, dermatologie.

Abstract

Structure from motion (SfM) algorithms represent an efficient means to construct extended 3D surfaces using images of a scene acquired from different viewpoints. SfM methods simultaneously determine the camera motion and a 3D point cloud lying on the surfaces to be recovered. Classical SfM algorithms use feature point detection and matching methods to track homologous points across the image sequences, each point track corresponding to a 3D point to be reconstructed. The SfM algorithms exploit the correspondences between homologous points to recover the 3D scene structure and the successive camera poses in an arbitrary world coordinate system. There exist different state-of-the-art SfM algorithms which can efficiently reconstruct different types of scenes, under the condition that the images include enough textures or structures. However, most of the existing solutions are inappropriate, or at least not optimal, when the sequences of images are without or only with few textures. This thesis proposes two dense optical flow (DOF)-based SfM solutions to reconstruct complex scenes using images with few textures and acquired under changing illumination conditions. It is notably shown how accurate DOF fields can be optimally used due to an image selection strategy which both maximizes the number and size of homologous point sets, and minimizes the errors in the homologous point localization. The accuracy of the proposed 3D cartography methods is assessed on phantoms with known dimensions. The robustness and the interest of the proposed methods are demonstrated on various complex medical scenes using a constant algorithm parameter set. The proposed solutions reconstructed organs seen in different medical examinations (epithelial surface of the inner stomach wall, inner epithelial bladder surface, and the skin surface in dermatology) and various imaging modalities (white light for all examinations, green-blue light in gastroscopy and fluorescence in cystoscopy).

Key-words: structure from motion, dense optical flow, 3D image mosaicing, endoscopy, dermatology.