Image-Laser Fusion for 3D Modeling of Complex Environments

Daniela Craciun

▶ **To cite this version:**

Daniela Craciun. Image-Laser Fusion for 3D Modeling of Complex Environments. Signal and Image processing. Télécom Paris, 2010. English. NNT : . tel-03177071

HAL Id: tel-03177071
https://hal.science/tel-03177071v1

Submitted on 23 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

École Doctorale
d'Informatique,
Télécommunications
et Électronique de Paris

Thèse

présentée pour obtenir le grade de docteur

de Télécom ParisTech

Spécialité : Signal et Images

Daniela CRACIUN

Numérisation conjointe image-laser pour la modélisation 3D des environnements complexes ou habités

Encadrée par Nicolas Paparoditis
Francis Schmitt

Soutenue le 7 juillet 2010 devant le jury composé de :

Peter Sturm	Président
Simon Lacroix	Rapporteurs
Maxime Lhuillier	
Raouf Benjemàa	Examinateurs
Isabelle Bloch	
Nicolas Paparoditis	Directeur de thèse

To my supervisor, Francis Schmitt with all my gratitude.

"Pour être juste, c'est-à-dire pour avoir sa raison d'être, la critique doit être partiale, passionnée, politique, c'est-à-dire faite à un point de vue exclusif, mais au point de vue qui ouvre le plus d'horizons."
Charles Baudelaire (1821-1867)

Remerciements

Les travaux de recherche présentés dans cette thèse ont été effectués au laboratoire de recherche MATIS (Méthodes d'Analyses pour le Traitement d'Images et la Stéréorestitution) attaché à la Direction Technique de l'Institut Géographique National, en collaboration avec Telecom ParisTech - laboratoire Traitement de Signal et d'Images - UMR CNRS 5141.

Je tiens à remercier d'abord les personnes qui ont défini le cadre et le thème de recherche de cette thèse,

à mon directeur de thèse, Francis Schmitt à qui je dédie le présent document et qui impactera toujours les défis que je chercherai à résoudre.

A mon encadrant, Nicolas Paparoditis pour m'avoir donné l'occasion de travailler sur un sujet si passionnant comme la modélisation tridimensionnelle des grottes préhistoriques. Il a su encadrer et motiver les travaux de cette thèse avec un enthousiasme d'exception qui encourage à aller toujours plus loin et à se donner ses propres défis.

A Isabelle Bloch pour m'avoir soutenue moralement et dirigée pendant la dernière année de thèse.

A Didier Boldo - ancien Directeur du laboratoire MATIS et à Patrice Bueso - ancien Directeur du Service de la Recherche de l'IGN - pour leur support technique et administratif.

Je souhaite adresser mes remerciements aux membres du jury,

à Peter Sturm, pour avoir accepté d'être le Président du jury et pour avoir examiné ce travail de thèse.

A Maxime Lhuillier et à Simon Lacroix qui ont contribué à l'amélioration du présent manuscrit.

A Raouf Benjemàa pour avoir examiné les présents travaux avec beaucoup d'intérêt.

Je remercie également les personnes qui ont participé à la mission à la grotte de Tautavel en Octobre 2007, en particulier aux élèves de l'Ecole Nationale Supérieure de Géographie, à Laure Chandelier et aux archéologues du Centre Européen de Recherches Préhistoriques de Tautavel qui nous ont accompagné pendant cette mission.

Mes remerciements vont vers mes collègues chercheurs du laboratoire MATIS qui ont su me ressusciter avec leur humeur pendant les moments les plus durs de la thèse. Je serai toujours reconnaissante à mon collègue de bureau, Nicolas Champion pour son soutien moral.

Je remercie mes nouveaux collègues chercheurs du CEA-LIST/Laboratoire de Vision et Ingénierie des Contenus pour le support accordé pendant la période précédant la soutenance de thèse. Je suis spécialement reconnaissante à Jérôme Esnault et à Steve Bourgeois pour m'avoir permis de visualiser les résultats finaux de ma thèse et aux thésards Pierre Lothe, Julien Michot et Alexandre Eudes.

La thèse conclut huit ans d'études, dont deux effectués à l'Université Politehnica Bucarest et six effectués en France. Je remercie les enseignants-chercheurs de l'Université

Politehnica Bucarest: Catalin Ionita pour m'avoir initié à la recherche et Florin Ionescu pour m'avoir encouragé d'étudier à l'Université de Sciences et Technologies de Lille I.

Je tiens à remercier Damien Muselet et Ludovic Macaire - chercheurs à l'Université des Sciences et Technologies de Lille I - pour m'avoir donné l'occasion de débuter dans le monde de la vision par ordinateur en 2005. Mon intérêt pour ce domaine a été ensuite encouragé par Patrick Fabiani qui je remercie vivement pour m'avoir permis d'effectuer mon stage de fin d'études en vision par ordinateur à l'ONERA - Toulouse sous l'encadrement de Guy Le Besnerais (ONERA - Châtillon). C'est pendant ce stage que j'ai eu l'occasion de réaliser le grand potentiel de la vision par ordinateur pour adresser différents problèmes en robotique et j'ai cherché par la suite de réaliser une thèse dans ce domaine.

Je remercie mes parents Petre et Lica et mon frère, Vladimir qui m'ont toujours soutenue durant mes études. Je leur suis particulièrement reconnaissante pour le support moral accordé à la fin de la thèse.

Je souhaite remercier mes amis Dan Jianu, Razvan Ionica, Julian Galli et Cristian Kreindler pour le support accordé pendant les dernières années.

Abstract

One might wonder what can be gained from the image-laser fusion and in which measure such a hybrid system can generate automatically complete and photorealist 3D models of difficult to access and unstructured underground environments.

In such environments, special attention must be given to the main issue standing behind the automation of the 3D modeling pipeline which is represented by the capacity to match reliably image and laser data in GPS-denied and feature-less areas. In addition, time and in-situ access constraints require fast and automatic procedures for in-situ data acquisition, processing and interpretation in order to allow for in-situ verification of the 3D scene model completeness. Finally, the currently generated 3D model represents the only available information providing situational awareness based on which autonomous behavior must be built in order to enable the system to act intelligently on-the-fly and explore the environment to ensure the 3D scene model completeness.

This dissertation evaluates the potential of a hybrid image-laser system for generating in-situ complete and photorealist 3D models of challenging environments, while minimizing human operator intervention. The presented research focuses on two main aspects: (i) the automation of the 3D modeling pipeline, targeting the automatic data matching in feature-less and GPS-denied areas for in-situ world modeling and (ii) the exploitation of the generated 3D models along with visual servoing procedures to provide mobile systems with autonomous site digitization and exploration capabilities.

We design a *complementary* and *cooperative* image-laser fusion which lead to a *4D mosaicing sensor* prototype. A 4D mosaic represents a 4-channel data structure encoding color (R, G, B) and depth components for each pixel coming from several overlapped images and 3D scans aligned in a global coordinate system. The *complementary* aspect is related to the data acquisition process. In order to deal with time and in-situ access, the proposed acquisition protocol consists in acquiring low-resolution 3D point clouds and high-resolution color images to generate in-situ photorealist 3D models. The use of both sensors rigidly attached leads to a *cooperative* fusion, producing a dual sensing device capable to generate in-situ omnidirectional and photorealist 3D models encoded as *4D mosaic* views, which are not achievable when using each sensor separately.

Each sensor is set to acquire the necessary data to cover a fully spherical field of view from a single 3D pose of the system which are further exploited by laser and image alignment algorithms for generating in-situ 3D and 2D Giga-pixel mosaic views. Although both sensors are rigidly attached, they are related through a global 3D rotation and a small inter-sensor parallax. The proposed image-laser acquisition strategy allows to perform occlusion-free image-laser alignment and texture mapping processes via a mosaic-based framework designed in a coarse-to-fine approach, providing robustness to small inter-sensor parallax.

This leads to a two-steps strategy which addresses the automation of the 3D modeling

pipeline by solving for its main data alignment issues through the image-laser fusion. We first address a simple problem, i.e. same viewpoint and small-parallax data alignment, resulting in automatic 2D and 3D mosaicing algorithms, to provide in a second step image-laser solutions, i.e. the *4D mosaic* views, to solve for wide-baseline 3D modeling alignment using a joint 2D-3D criterion to disambiguate feature matching in feature-less areas.

The proposed automated 3D modeling pipeline gave rise to several solutions for automatic data matching algorithms, from which other stand-alone sub-systems emerged. We propose hardware and software solutions for generating in-situ 2D, 3D and 4D mosaic views in feature-less and GPS-denied areas. They can be employed either as stand-alone processes or included within a 3D modeling process. In our research work we integrate the 4D mosaicing sensor within a vision-based system designed to supply site digitization and exploration to generate in-situ complete and photorealist 3D models in complex environments.

Since the autonomous site digitization and exploration problem is intrinsically related to the unmanned system's autonomy via the world modeling capability, we first investigate in which measure the 4D mosaicing sensor can solve for the system's autonomy problem and propose a vision-based autonomy model to be embedded onboard mobile platforms designed to supply complex missions in challenging environments, site surveys being one of them.

The proposed visual autonomy model was further instantiated to the autonomous site digitization and exploration case, giving rise to the ARTVISYS system which comes together with a 4D mosaic-driven acquisition scenario and automatic data matching softwares for supplying in-situ the entire 3D modeling pipeline. The 4D mosaicing sensor represents the nucleus of the 3D world modeling process which powers visual servoing procedures in-charged with the 3D scene model completeness.

Since the processing blocks composing the visual feedback loop exploit the global 3D scene model, we evaluate the 4D mosaic's potential to address the pose estimation problem. To this end, we propose image-laser solutions for disambiguating the data matching process which is inherent to outliers in feature-less areas. Short-term research perspectives are focused on the remaining processing blocks composing the visual feedback loop, such as view planning and autonomous navigation.

Keywords: image-laser fusion, complex environments, automatic in-situ 3D modeling, automatic data alignment, 3D mosaicing, Gigapixel mosaicing, 4D mosaicing, 3D model matching, 4D mosaic-driven hybrid SLAM, active vision, vision-based unmanned systems, visual-autonomy, autonomous site 3D digitization and exploration.

Contents

Résumé

1 Introduction

Le sujet de cette thèse s'inscrit dans l'action de recherche ARCHI dirigée par Nicolas Paparoditis (IGN-MATIS) et vise à mettre en place un système de vision capable d'effectuer la numérisation 3D automatique in-situ d'édifices remarquables et difficile d'accès pour un opérateur humain, en particulier les sites architecturaux préhistoriques. Néanmoins, la méthode développée peut être utilisée également pour la numérisation d'une grande variété des sites complexes, pour des applications telles que : l'héritage culturel, visites virtuelles, l'analyse des scènes pour la maintenance des sites difficiles à accéder par un opérateur humain (comme par example les mines [Huber and Vandapel, 2003a], [Baker et al., 2004b], [Cole and Newman, 2006], [Survey, 2006], les tunnels [Chaiyasarn et al., 2009]), ou pour l'exploration des environnements sous-marins [Garcias and Santos-Victor, 2000] ou extra-terrestres [Mathies et al., 2007], [Johnson et al., 2007]. La validation des méthodes mises en oeuvre durant la thèse est effectuée sur des sites présentant une architecture complexe, en occurrence les grottes préhistoriques.

2 Problématique

La numérisation exhaustive et photoréaliste d'environnements complexes représente aujourd'hui un grand défi en raison d'une part du besoin d'automatisation des processus d'acquisition et de traitement qui sont encore quasiment manuels et d'autre part en raison de la difficulté de vérifier in situ l'adéquation du modèle avec le cahier des charges. Très souvent on constate a posteriori, une fois les données traitées, que le modèle 3D est incomplet et il n'est souvent pas possible de retourner sur site pour compléter les numérisations.

Dans le cadre de cette thèse, nous nous intéressons à l'automatisation du processus de numérisation 3D d'environnements complexes et en particulier non-structurés qui sont plus difficiles aujourd'hui à traiter avec les outils proposés dans la littérature. Les travaux de recherche réalisés visent d'une part la mise en oeuvre de méthodologies d'acquisition de données et d'autre part le développement d'algorithmes pour le traitement de données in-situ afin d'aider les opérateurs dans leur travail de manière à assurer la bonne numérisation du site.

Comme contexte applicatif, nous nous intéressons aux grottes ornées préhistoriques qui sont des environnements particulièrement difficiles. Dans de tels environnements l'absence des structures habituellement utilisées pour la mise en correspondance et la mise en géométrie des images rend le problème très difficile, voir impossible. L'utilisation de cibles pour faciliter la partie de mise en géométrie des données n'est pas souhaitable d'une part parce qu'elle ralentit fortement les cadences de numérisation (alors que le temps de numérisation autorisé est restreint) et d'autre part car il est difficile voir interdit de poser

FIG. 1 – Campagne de numérisation réalisée dans le grotte Mayenne Science (France) ©IGN-ESGT.

des cibles sur des parois. Par ailleurs, le recours à des solutions basées sur des systèmes de localisation/navigation externe (centrale inertielle etc.) pour aider au géo-référencement est difficilement envisageable et complexe à mettre en oeuvre, par rapport à des chantiers en extérieur où le GPS est disponible.

Le besoin de développer un système de vision pour l'automatisation du processus de numérisation 3D est mis également en évidence par la difficulté d'un opérateur humain d'accéder des tels environnements complexes (pour poser des cibles et guider l'acquisition) et par la nécessité de visualiser in situ le modèle 3D acquis de la scène afin d'assurer la complétude du modèle 3D du site. Pour donner un exemple de la difficulté engendrée par la mise en oeuvre d'un scénario d'acquisition, la Figure 1 illustre un exemple d'acquisition des données réalisé par l'IGN et l'ESGT dans la grotte de Mayenne Science.

3 Solution image-laser proposée pour la numérisation in-situ

Dans le cadre de cette thèse, nous nous proposons d'évaluer le potentiel d'un système de vision photogrammetrique et lasergrammetrique pour la modélisation 3D in-situ des environnements complexes et difficile à accéder pour un opérateur humain. Un tel système doit :

- acquérir et traiter les données sans l'intervention d'un opérateur humain,
- générer des modèles 3D de manière séquentielle
- agir intelligemment afin de complétiser le modèle 3D du site.

Le mémoire est structuré en huit chapitres qui introduisent graduellement la solution portant sur l'utilisation conjointe image-laser pour la modélisation 3D in-situ. Le **Chapitre 2** présente plusieurs applications faisant appel aux techniques de modélisation 3D in-situ et survole les différentes méthodes permettant une numérisation 3D photoréaliste et précise. Ces techniques seront évoquées dans le Chapitre 3 pour justifier le choix des capteurs que nous avons fait pour les travaux de cette thèse. Nous continuons l'état de l'art par une présentation des techniques existantes pour la modélisation 3D en mettant en évidence leur limitation concernant la génération des modèles 3D dans les environnements non-structurés. Le Chapitre 2 conclut avec les aspects clefs qui doivent être résolus pour répondre aux problèmes de la modélisation 3D in-situ ayant lieu dans les environnements complexes.

Les cinq sous-sections suivantes sont dédiées à la description du système de numérisation proposé dans cette thèse dans lequel nous projetons la solution image-laser pour la modélisation 3D in-situ des environnements complexes.

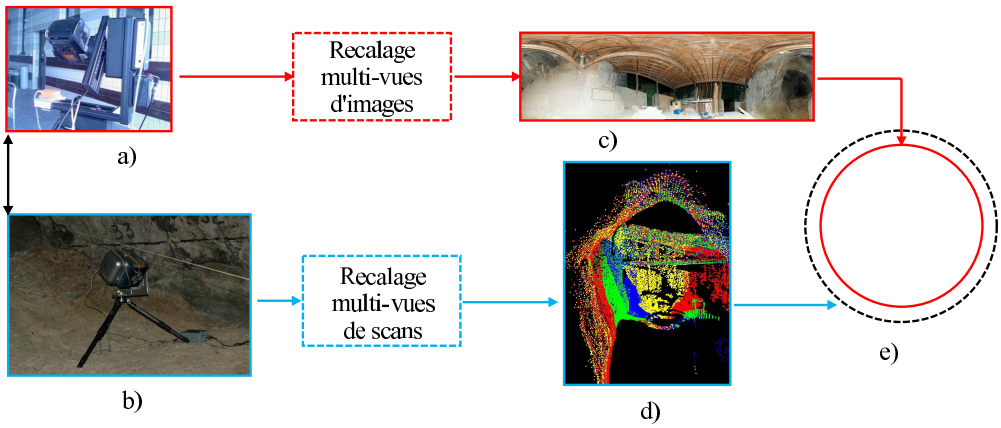


FIG. 2 – Le processus de mosaiquage 4D proposé pour la modélisation 3D in-situ. (a) appareil photo NIKON D70 ® monté sur une plateforme pan-tilt motorisée, (b) laser scanner 3D Trimble ® - campagne de numérisation de la grotte de Tautavel (France) réalisée par l'Institut Géographique National, Octobre 2007, (c) panoramique couleur Gigapixel obtenue via l'algorithme de recalage d'images proposé dans le Chapitre 5 de cette thèse, (d) mosaique 3D obtenue à partir de plusieurs scans avec un recouvrement partiel via l'algorithme de recalage des scans multi-vues proposé dans le Chapitre 4, (e) recalage de mosaique 3D sur la mosaique 2D couleur pour generer une mosaique 4D, i.e. à 4 canaux : rouge, vert, bleu et profondeur. Cette procedure est décrite dans le Chapitre 6 de la thèse.

3.1 Système image-laser pour la modélisation 3D in-situ

On s'intéresse au prototypage d'un système de vision capable de générer des modèles 3D photorealistes et complets in-situ. Le **Chapitre 3** introduit l'architecture matérielle et logicielle, le scénario d'acquisition et les fonctionnalités composant le processus de la modélisation 3D. Nous introduisons un système à double tête, composé par un laser scanner et une caméra haute résolution montée sur une plateforme pan-tilt motorisée. Pour assurer le photoréalisme du modèle, notre approche utilise davantage la complémentarité image-laser, en utilisant un système à double tête, composé par un laser scanner (illustré dans la Figure 2 (a)) et une camera couleur haute résolution montée sur une plateforme pan-tilt tournante motorisée (illustrée dans la Figure 2 (b)).

L'utilisation conjointe des deux capteurs est réalisée de manière complémentaire :

- acquisition et traitement rapide des scans basse résolution,
- photoréalisme via l'acquisition des images couleurs haute résolution.

L'exploitation des deux capteurs est conditionnée par l'étape de recalage des données dans un référentiel commun. Cette étape est encore plus difficile à cause des occultations présentent à la fois dans les données images et les données laser. Le système d'acquisition présenté dans cette thèse évite ce problème en utilisant les deux capteurs attachés de manière rigide mais non-calibré. Le montage proposé réduit les problèmes liés aux procédures de recalage et de construction de modèles 3D dans les zones où les données images et/ou laser sont manquantes. La construction matérielle du système d'acquisition minimise la parallaxe inter-capteurs facilitant ainsi le recalage des données image-laser. Le système proposé nous permet la mise en oeuvre des algorithmes de recalage des données multi-vues guidés via un critère géométrique et radiométrique robuste aux faux appariements qui sont inhérents quand un critère radiométrique seul est utilisé.

3.1.1 Mosaique 4D

Les travaux de cette thèse introduisent les mosaïques 4D comme des modèles 3D omnidirectionnelles encodant l'information radiométrique et géométrique couvrant un champ de vue sphérique complet dans une seule représentation compacte. Pour chaque position spatiale, le système acquiert des données images et laser avec un champ de vue limité. Elles sont ensuite exploitées par des algorithmes de recalage image-image, laser-laser et image-laser afin de produire une représentation unique du sous la forme de mosaïque 4D, unifiant ainsi le photoréalisme et l'information géométrique dans une seule image à quatre canaux : R,G,B et profondeur.

Une mosaique 4D est obtenue en trois étapes :

1. Mosaiquage 3D. Le système commence par acquérir plusieurs scans 3D avec un recouvrement partiel pour ensuite générer une mosaique 3D in-situ via un algorithme automatique de recalage de scans multi-vues. La Figure 2 (d) illustre un exemple de résultat obtenu sur des données acquises dans la grotte de Tautavel (France). Cette méthode est détaillée dans le **Chapitre 4** et résumée en Section 3.2.

2. Mosaiquage optique haute résoluton. Le système acquiert une séquence d'images haute résolution qui sont ensuite injectées dans un moteur de recalage d'images multi-vues pour générer une mosaïque optique RGB Gigapixel. La Figure 2 (c) illustre un exemple de résultat obtenu sur des données acquises dans la grotte de Tautavel. Cette méthode est décrite dans le **Chapitre 5** et résumée en Section 3.3.

3. Recalage image-laser. La troisième étape consiste à recaler la mosaique 3D issue du recalage multi-vues des scans (étape 1) par rapport à la mosaique optique RGB Gigapixel (étape 2). Cette étape correspond à la Figure 2 (e). Pour ceci, il est nécessaire de calculer la transformation rigide qui doit être appliquée à la mosaique 3D pour minimiser un critère de distance (métrique ou radiométrique) entre les zones de recouvrement des deux mosaiques. Cette procédure unifie les deux mosaïques (3D et RGB optique) dans une mosaïque à 4-canaux (R, G, B et profondeur) qu'on appel par la suite une mosaique 4D. Cette procédure est décrite dans le **Chapitre 6** et résumée en Section 3.4.

Comme les techniques existantes de recalage des données imposent des contraintes sur le contenu de la scène, nous avons mis en oeuvre des procédures de recalage automatiques des données images, laser et image-laser indépendantes de l'environnement. Plus précisément, aucune connaissance a priori sur le contenu de la scène (comme par exemple l'existence des primitives radiométrique ou géométriques). De plus, aucune estimation initiale n'est disponible (comme par exemple un recalage grossier manuel ou à partir des données fournies par les capteurs de navigation classiques).

La visualisation des données image et laser est réalisée via la création des mosaïques 4D. Ce sont les images à quatre composantes : rouge, vert, et bleu et la profondeur. Cette représentation compacte de plusieurs millions de points 3D et plusieurs centaines d'images haute résolution fournit une topologie de données qui facilite l'exploitation des données. En l'occurrence, il est possible de visualiser les coordonnées 3D des points localisés dans les images 2D couleur, et donc de trouver les coordonnées 3D d'un objet facilement identifiable dans l'image couleur. Cette visualisation dans l'espace image devient plus pratique quand on souhaite vérifier la complétude de la numérisation du site via l'absence de signal. La création du modèle 3D est réalisée en deux étapes : d'abord une construction du modèle 3D est effectuée via une procédure de triangulation 2D des nuages des points suivie par l'étape de plaquage de texture pour activer le photoréalisme du modèle 3D.

3.1.2 Scénario d'acquisition des panoramiques 4D

En présence des occultations, plusieurs mosaïques 4D doivent être acquises à partir des points de vues différents afin de complétiser le modèle 3D du site. Plus précisément, le système génère des mosaiques 4D in-situ pour chaque position 3D du système (nommée station), comme illustré dans la Figure 3. Pour ces raisons, après l'acquisition de chaque mosaïque, le système procède à son recalage par rapport à un référentiel global afin de créer le modèle global de la scène.

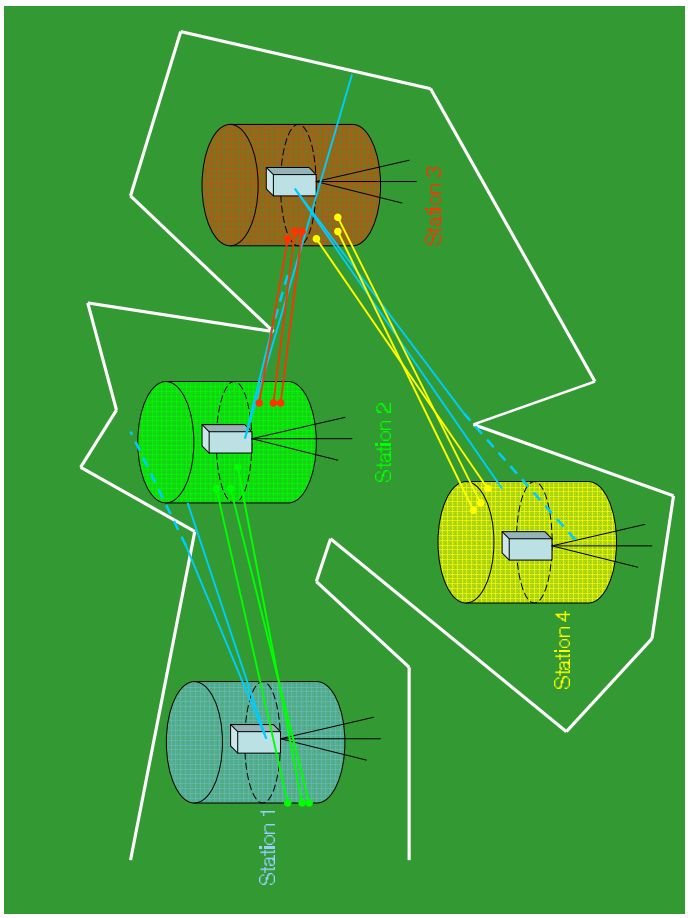


FIG. 3 – Scénario d'acquisition basé sur l'acquisition des panoramiques 4D proposé pour la modélisation 3D in-situ.

Dans un deuxième temps, le système doit déterminer la meilleure prochaine position du système d'où la prochaine mosaique 4D doit être acquise afin de minimiser les occultations (résoudre le problème de Next Best View (NBV) [Klein and Sequeira, 2000]). De plus, le système doit être capable de naviguer de manière autonome entre sa position courante et celle calculée par la procédure de NBV. Il s'agit de mettre en oeuvre des procédures de planification de vue et de trajectoire ainsi que des moyens pour la navigation autonome (détection d'obstacles, moyens rapides de raisonnement et de prise de décision).

Le scénario basé sur l'acquisition des mosaïques 4D permet de résoudre plusieurs problèmes liés à la modélisation 3D photorealiste des environnements non-structurés.

Appariement des données robuste. Dans les environnements complexes, l'absence des primitives radiométriques et géométriques (i.e. zones homogènes ou très texturées)

rend l'appariement très sensible aux mesures aberrantes (outliers), voire impossible. Nous proposons une méthode de recalage des données à travers des mosaiques 4D, en utilisant l'information 2D et 3D, afin de contraintre et desambiguer l'appariement des données dans les zones non-structuré.

Rendu-image basé sur la 3D. Contrairement aux méthodes de rendu basées purement sur l'utilisation de l'image (image-based rendering) pour interpoler et générer des vues synthétiques (ces méthodes n'utilisent pas la géométrie, et de nombreuses prises des vues sont nécessaires pour produire des nouvelles vues cohérentes avec la scène réelle), les mosaïques 4D permettent d'interpoler entre les vues en utilisant l'information 3D pour créer des nouvelles vues cohérentes géométriquement. Ceci facilite la navigation à travers le web pour les applications de type tourisme virtuel, mais aussi pour l'archivage et l'annotation de données.

Le système de numérisation proposé possède une double capacité : il permet de réaliser la numérisation 3D des sites inaccessible aux operateurs humains, fournissant simultanément au système la perception de l'environnement dans lequel il évolue pour réaliser des tâches plus complexes (analyses des scènes, maintenance des sites à risques, etc). Comme cette propriété exploite uniquement les capteurs de vision (active ou passive), nous l'avons dénommé ARTVISYS, l'acronyme de ARTificial VISion-based SYStem, en anglais système de vision artificielle. Nous avons appliqué la méthodologie de numérisation proposée au cas de la numérisation et l'exploration des grottes préhistoriques qui sont des environnements particulierement difficile d'accès pour un opérateur humain.

3.2 Recalage automatique multi-vues de scans pour le mosaiquage 3D

Dans un premier pas, le laser acquiert plusieurs scans avec un recouvrement partiel qui sont recalés pour former une mosaïque 3D avec un champ de vue sphérique complet. Nous avons mis en oeuvre un scenario d'acquisition pour le recalage multi-vues des scans 3D avec un recouvrement partiel, afin de générer une mosaïque 3D complète in-situ. Le scenario d'acquisition facilite considérablement la tâche de recalage en fournissant un recouvrement constant et minimal de 33%.

Cette méthode repose essentiellement sur la corrélation dense des informations fournies par le capteur, ici le laser, i.e. l'intensité et profondeur associé à chaque point 3D. La méthode est capable de fonctionner en deux modes, suivant la type d'information fournie par le capteur. Dans le cas où l'intensité n'est pas exploitable, la méthode utilise la profondeur pour le recalage des scans.

L'originalité de cette méthode réside dans plusieurs aspects :

- l'utilisation du scénario d'acquisition pour création des mosaïques 3D, assurant un recouvrement partiel et constant aux pôles pour faciliter le recalage ;
- recalage des nuages 3D dans l'espace des panoramiques 2D de profondeur ou d'intensité pour exploiter l'information topologique des scans ;
- robustesse à l'absence des primitives radiométriques et géométriques ;
- la procédure de recalage n'utilise aucune connaissance à priori sur l'environnement, et aucune estimation initiale n'est exigée pour le recalage ;
- la méthode proposée remplace les deux étapes de procédures de recalage habituellement utilisées (i.e. dans un premier temps un recalage grossier est réalisé soit manuellement, soit en utilisant des cibles, soit en exploitant des informations fournies par les capteurs de navigation, dans un deuxième temps le recalage fin est réalisé via une technique de type Iteratively Closest Points - ICP [Besl and McKay PAMI92]).

L'approche proposée est une méthode automatique, pyramidale, n'imposant pas des contraintes sur le contenu de la scène. Le choix de la méthode a été influencé par l'absence de primitives dans les environnements non-structurés. Comme aucune connaissance sur le contenu de la scène n'est utilisée, cette technique est indépendante de l'environnement qu'on souhaite numériser (structuré ou non-structuré).

Afin de valider la méthode mise en oeuvre, nous avons réalisé une campagne d'acquisition de données dans la grotte de Tautavel pendant le mois d'octobre 2007. Les résultats de la méthode sont illustrés dans les figures 4 et 5.

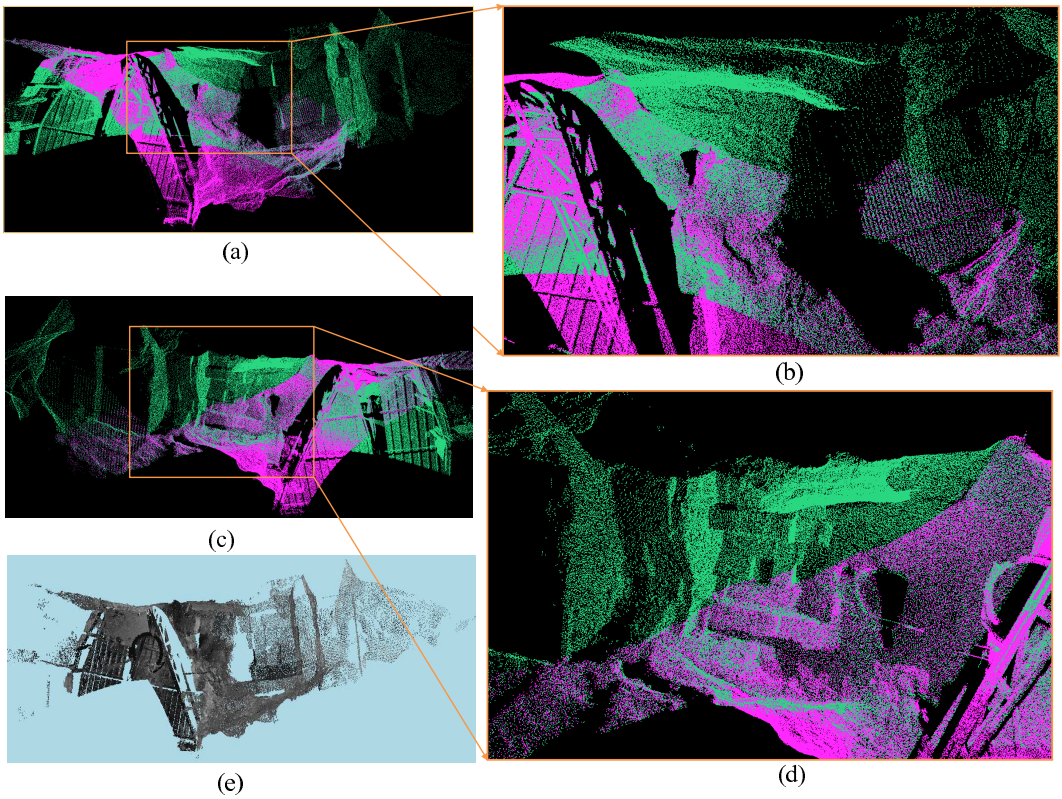


FIG. 4 – Resultats de la méthode de recalage d'un couple des scans - campagne de numérisation réalisée dans la grotte de Tautavel, magenta : scan de réference, vert - scan à recaler. a) Scans récalés - vue de dessus, b) plafon de la grotte, c) scans recalés - vue de dessous, d) vue sur le sol de la grotte, e) l'union des scans recalés en utilisant l'intensité délivré par le laser.

La méthode proposée est décrite dans le **Chapitre 4** et fait l'objet de deux publications [Craciun et al., 2008], [Craciun et al., 2010].

3.3 Recalage automatique d'images pour la création de mosaïques optique Gigapixel

Le deuxième processus composant le mosaiquage 4D est représenté par la mise en correspondance des images couleurs haute résolution acquises par une plateforme pan-tilt motorisée pour générer in-situ une panoramique Gigapixel. Dans cette partie nous cherchons à résoudre une des limitations des algorithmes de mosaïquage existantes portant sur l'appariement des images très haute résolution avec recouvrement faible (pour éviter

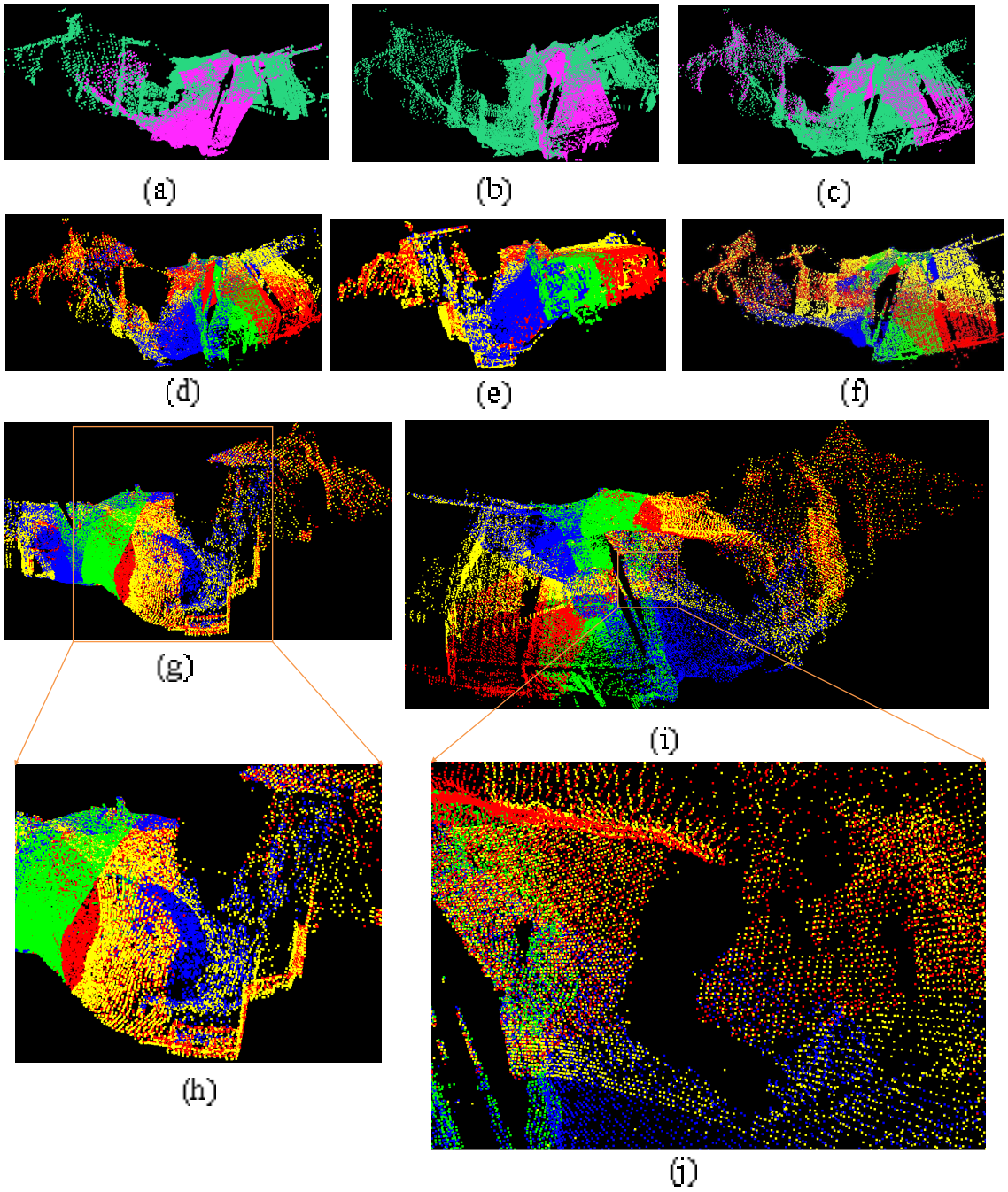


FIG. 5 – Résultats de la procédure de recalage multi-vues des scans acquis dans la grotte de Tautavel. (a) S_1 - vert, S_2 - magenta, (b) S_{12} - vert, S_3 - magenta, (c) S_{123} - vert, S_4 - magenta, (d) Recalage multivues des scans - vue de dessus, S_1 - jeun, S_2 - bleu, S_3 - vert, S_4 - rouge, (e) vue latérale - gauche, (f) vue de dessus, (g) vue latérale - droite, (h) zoom-in (e), (i) vue de dessous, (j) zoom-in - vue du plafon.

les données redondantes) dans les environnements non-structurés. Le **Chapitre 5** décrit une méthode automatique de recalage multi-vues basée sur la mise en correspondance de patches (vignettes), et par conséquent robuste à l'absence des points d'intérêt. L'algo-

rithme proposé permet le recalage multi-vues d'images haute résolution pour la génération automatique d'une mosaïque Gigapixel pour une station donnée [Craciun et al., 2009] et [Cannelle et al., 2009].

Le coeur de l'algorithme est une méthode d'estimation de mouvement inter-image qui prend en entrée un couple d'images avec un recouvrement partiel et donne en sortie une liste des points homologues reliant les deux images. L'algorithme commence par estimer un mouvement global de rotation dans une approche pyramidale basée sur la corrélation de patches extraits que sur les bords de l'image, afin de réduire le temps de calcul. Après avoir compensé le mouvement de rotation, la procédure continue par une mise en correspondance locale des patches réalisée au niveau de résolution le plus fin, afin d'estimer un mouvement local de translation pour chaque patche. Cette procédure donne en sortie une liste des patches homologues pour chaque couple d'images adjacentes qui seront exploités ensuite par d'ajustement par faisceaux pour l'estimation des poses absolues. Comme les points homologues ne correspondent pas aux points d'intérêts habituels, nous les avons dénommé anonymous features. Ces derniers sont injectés dans un algorithme d'ajustement par faisceaux pour permettre une estimation fine des poses globales. L'étape d'ajustement par faisceaux et le rendu de la mosaïque sont réalisés par une approche existante, AutopanoPro [Kolor, 2005].

L'algorithme de mosaiquage proposé exploite la complémentarité des algorithmes de mosaiquage existants [Szeliski, 2006], notamment les algorithmes de corrélation dense [Teller and Coorg, 2000] (précis mais gourmands en temps de calcul pour les images haute résolution) et les algorithmes éparses pouvant bénéficier de la rapidité de recalage multi-vues fin via la procédure de compensation par faisceaux (bundle adjustment [Triggs et al., 1999]). Nous mettons en valeur la haute résolution du mosaiquage en utilisant un visualiseur avec 8 niveaux de détail adapté aux applications de visites virtuelles.

Le résultat final illustré dans la Figure 6 est fortement influencé par l'étape d'ajustement par faiceaux réalisé par AutopanoPro [Kolor, 2005] qui impose une étape de ré-estimation des paramètres intrinsèques de la camera ; par conséquent, l'algorithme procède à la rejection des points homologues qui ne correspondent pas au modèle de mouvement estimé par l'étape d'ajustement par faisceaux. Une deuxième raison à l'origine de la rejection des faux appariements est essentiellement due au critère de mesure de l'erreur résiduelle. Cette dernière étant calculée dans l'espace image, et non pas dans l'espace objet, le critère mesuré est biaisé par l'étape de ré-estimation des paramètres intrinsèques. Par conséquent, la mosaïque finale comporte des artefacts, en particulier des effets de dé-doublement (ghost) et d'éspace entre le début et la fin de la panoramique (gap) due aux faux appariements. Pour cette raison, nous proposons une solution théorique pour l'étape de recalage multi-vues fin, qui estime les quaternions optimaux pour chaque vue en minimisant un critère mesuré dans l'espace 3D.

3.4 Recalage image-laser pour la génération de mosaïques 4D

La dernière étape consiste à recaler la mosaïque couleur haute résolution sur la mosaïque 3D provenant du laser. Pour cette raison, il est nécessaire de calculer la transformation 3D rigide reliant le laser et la caméra. Les deux capteurs sont rigidement attachés et légèrement décalés. Le **Chapitre 6** décrit la mise en oeuvre d'un algorithme de recalage basé sur la corrélation dense multi-résolution pour le calcul de pose camera-laser. Grâce au montage matériel, la parallaxe inter-capteurs devient négligeable aux niveaux basse résolution. Une mise en correspondance de vignettes est réalisée au niveau le plus fin pour

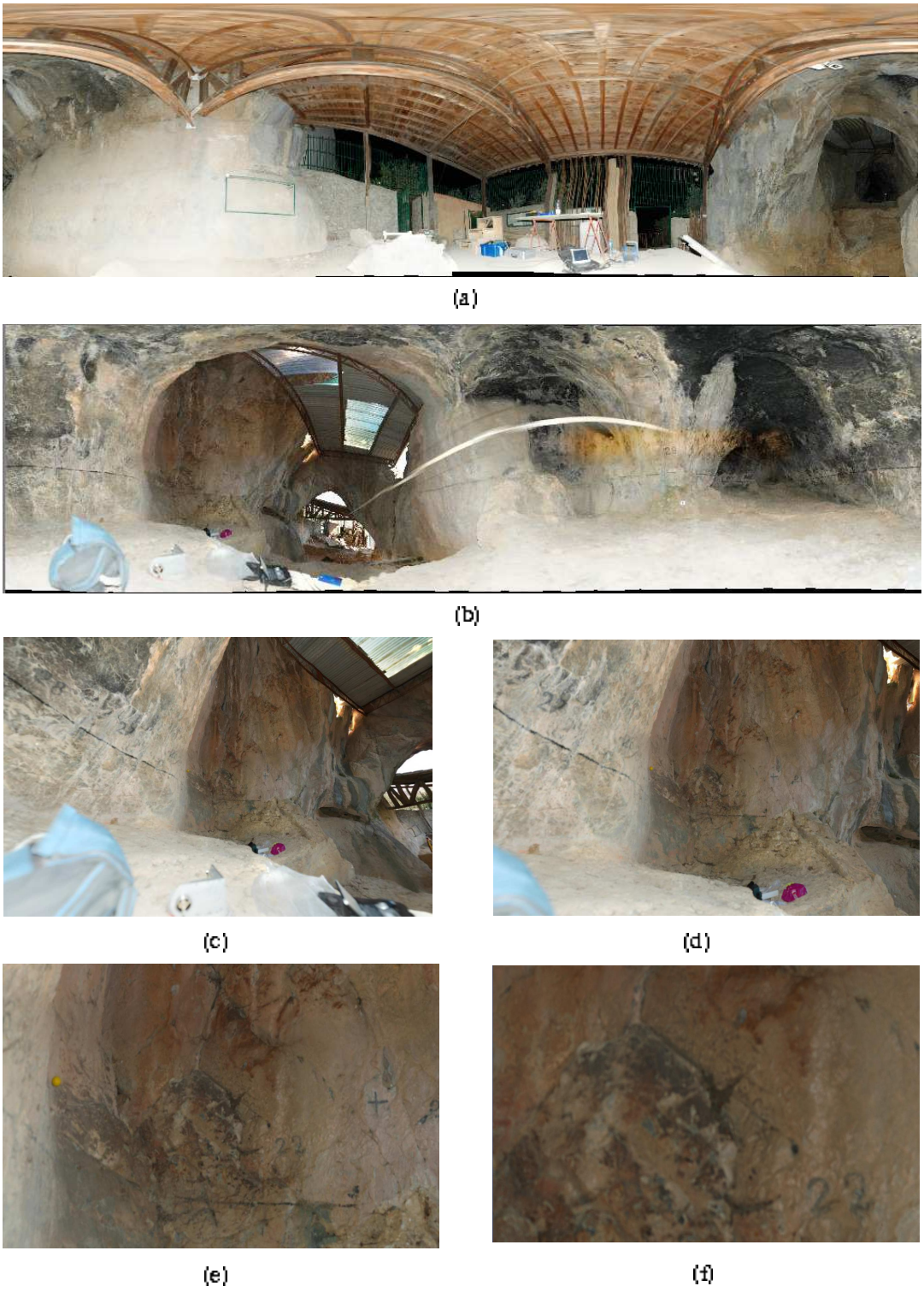


FIG. 6 – Résultat d'une mosaique 2D optique haute résolution obtenue à partir d'une séquence d'images acquises dans la grotte de Tautavel en utilisant la plateforme Rodeon®. Les mosaiques on été générées en injectant les points homologues (AF) dans le processus de compensation par faiceaux intégré dans Autopano Pro v1.4.2. (a) entrée de la grotte, (b) - centre la grotte, (c)-(f) - niveux de détail correspondants au mosaique de la figure (b).

extraire et compenser la parallaxe inter-capteurs dans l'espace image. Due aux différences radiométriques des deux capteurs, seule une pose grossière peut être calculée. Néanmoins, la pose grossière est suffisante pour permettre une visualisation rapide in-situ de données collectées. Pour un recalage fin, un calibrage radiométrique apparaît nécessaire. Des travaux de recherche sur ce sujet ont été démarrés au laboratoire MATIS.

La procédure de recalage des mosaïques 3D et 2D Gigapixel exploite l'information d'intensité de la mosaïque 3D. La Figure 7 illustre les données en entrée de la procédure de recalage des données image-laser. Afin d'obtenir le modèle 3D numérique d'une mosaïque 4D, on associe à chaque point 3D la couleur R, G, B correspondante de la mosaïque couleur. Le résultat final de cette procédure est une mosaïque à 4-canaux (R, G, B et profondeur) qu'on appelle une mosaïque 4D. La Figure 8 illustre le modèle 3D d'une mosaïque 4D.

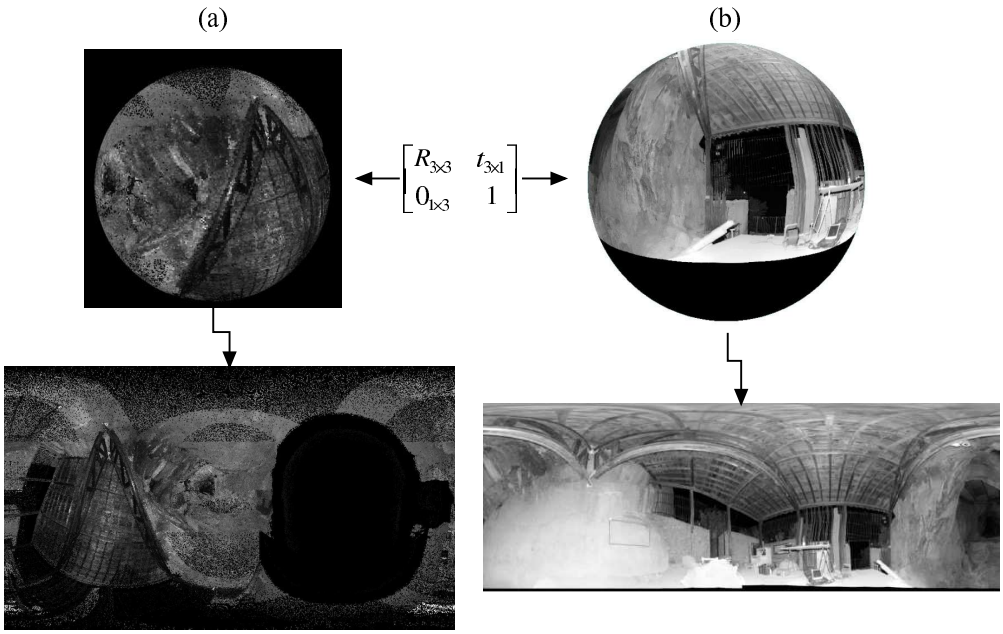


FIG. 7 – Les deux entrées de la procedure de recalage image-laser (résultats obtenus sur les données acquises dans la grotte de Tautavel). Nous illustrons les projections spherique et planaire, dans l'espace image pour chaque donnée en entrée. (a) le mosaique 3D issue de la procédure de recalage décrite dans le Chapitre 4. (b) le mosaique optique haute résolution obtenue en utilisant l'algorithme décrit dans le Chapitre 5.

Le nuage de points 3D transformé passe par une procédure automatique de triangulation 2D suivie par une procédure de plaquage de la texture provenant de la mosaïque Gigapixel sur-échantillonnée. La Figure 9 illustre le modèle 3D d'une mosaïque 4D obtenue après l'étape de plaquage de texture.

3.5 Vers l'exploration des sites complexes basée sur l'acquisition des panoramiques 4D

En présence d'occultations, plusieurs mosaïques 4D doivent être acquises depuis des points de vue bien définis afin de numériser de manière complète le site, dans un intervalle de temps assez limité, tout en évitant les données manquantes et/ou redondantes (trop de recouvrement entre les différents données a comme conséquence des temps de calcul assez longs).

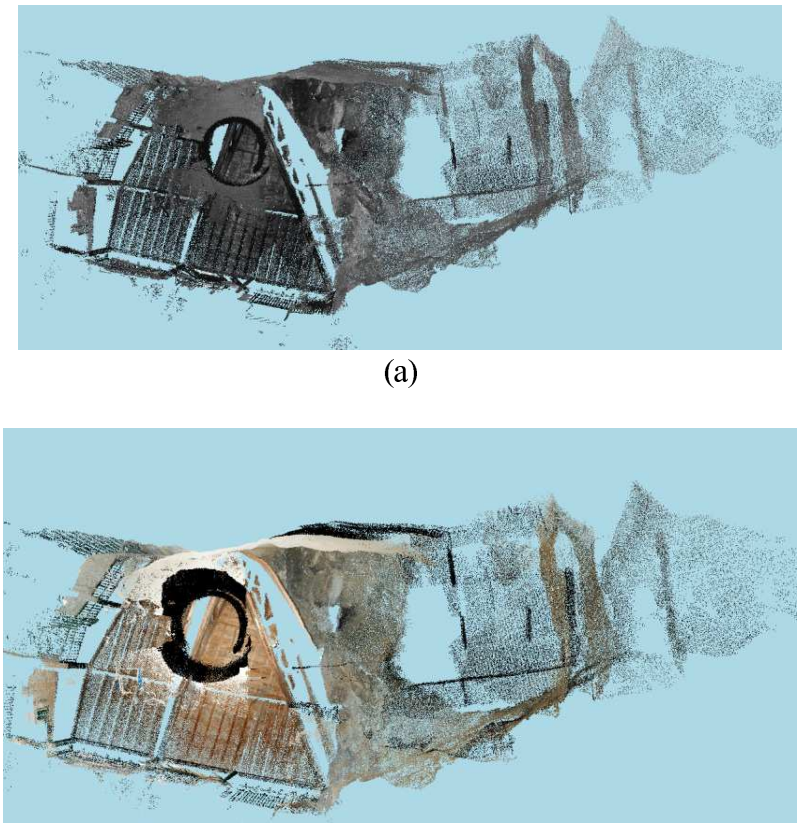


FIG. 8 – Modèle 3D de la mosaique 4D. (a) Le nuage des points affiché en utilisant l'intensité délivrée par le laser scanner. (b) Le nuage des points 3D affiché avec la composante couleur du mosaiquage optique haute résolution sur-échantillonné.

FIG. 9 – Modèle 3D de la mosaique 4D obtenu après le plaquage de texture sur des meshes 2D. (a) Vue de l'extérieur de la grotte de Tautavel. (b) Vue de l'intérieur du modèle 3D de la grotte.

Dans cette thèse nous proposons un scénario de numérisation basé sur l'acquisition d'un réseau des panoramiques 4D qui seront recalées entre elles dans une approche séquentielle.

Le processus de numérisation alterne entre plusieurs phases :
- (i) génération de mosaïque 4D,
- (ii) recalage de mosaïques 4D,
- (iii) mise à jour du modèle global de la scène,
- (iv) calcul de la prochaine meilleure position du système pour la numérisation des parties cachées.

Le **Chapitre 7** propose une solution théorique pour le recalage des mosaïques 4D en utilisant un critère 2D-3D pour desambiguer la mise en correspondance des points d'intérêts. La technique proposée exploite l'avantage offert par le champ de vue sphérique qui permet d'avoir un suivi des primitives stable à long-terme.

Après avoir doté le système de numérisation avec un processus de modélisation fonctionnel, ce dernier doit être exploité pour guider le système à numériser les parties cachées du site. Cette procédure fait appel à une méthode de numérisation intelligente qui implique le calcul de la prochaine meilleure position du système à partir de laquelle la nouvelle mosaïque 4D dois être acquise (problème appelé Next Best View dans la littérature [Klein and Sequeira, 2000]). De plus, dans le cas où l'opérateur ne peut pas intervenir sur le site, le système doit être capable de naviguer de manière autonome entre sa position courante et celle calculée par le module de planification de vue. Nous introduisons un modèle d'autonomie visuelle qui exploite le modèle 3D courant pour extraire des sémantiques (caractéristiques geométriques et radiométriques de l'environnement formant le vocabulaire associé à l'environnement exploré). Cette sémantique peut être exploitée pour asservir le système de numérisation en position, notamment pour fournir les informations nécessaires à la navigation, i.e. pour la détection d'obstacles et la planification de la trajectoire. Ces dernières exploitent une seule et unique entrée : le modèle 3D global de la scène construit en itérant le processus d'acquisition, génération et recalage de mosaïques 4D.

La numérisation complète d'un environnement est intrinsèquement liée à l'autonomie du système qui doit explorer le site en se servant du modèle 3D généré. Pour cette raison, nous avons étudié dans quelle mesure le capteur de mosaiquage 4D peut servir de processus de base pour fournir un modèle d'autonomie fondée sur la vision. En l'occurrence, il s'agit de données image-laser unifiées sous la forme d'une mosaique 4D. Nous proposons une architecture logicielle qui peut être facilement intégrée dans un système de vision artificielle upgradable avec différents capacités pour la réalisation des missions complexes in-situ. Dans notre étude, nous avons instancié le modèle d'autonomie basée sur la vision au cas de la numérisation et l'exploration des environnements complexes. Les modules composant la boucle de contrôle portent principalement sur la planification de vue et la navigation autonome afin assurer in-situ la numérisation complète du site.

4 Conclusions

Le **Chapitre 8** conclut les travaux de recherche réalisés et résume les contributions dediées à l'automatisation de la modélisation 3D. La deuxième partie du Chapitre 8 présente les perspectives de recherche à court- et long-terme portant sur l'asservissement visuel pour assurer la numérisation complète du site.

Les méthodologies de numérisation et recalage des données image, laser et image-laser ont été conçues en tenant compte des environnements complexes. Une caractéristique commune est la robustesse à l'absence des point d'intérêts et aux environnements dans lesquels les capteurs de navigation (GPS, IMU) ne sont pas exploitables. Pour répondre aux contraintes d'embarquabilité et de fonctionnalité in-situ, nous donnons comme example la

parallélisation sur des plateformes de calcul multi-processeurs pour l'algorithme présenté dans le Chapitre 4. Des expérimentations ont été effectuées sur des données acquises dans trois sites préhistoriques ainsi que dans les environnements urbains.

Solutions locales - au niveau des procédures composant le système de modélisation 3D proposé :

- Mosaiquage 3D : scénario d'acquisition des scans 3D et recalage pour la génération de mosaïques 3D robuste au types de capteurs (cameras de distance ou de laser scanners) ;
- Mosaique optique : recalage automatique d'images haute résolution pour la création des mosaïques 2D Gigapixel ;
- Modèle 3D-RGB : technique de recalage des mosaïques image-laser pour la création des mosaïques 3D photorealistes (4D) ;

Solutions théoriques :

- détection de poses globales optimales pour le recalage multi-vues des scans ;
- solution théorique pour l'estimation des quaternions optimaux absolus pour le recalage multi-vues d'images haute résolution en utilisant les patches (anonymous features) ;
- solution théorique pour le recalage multi-vues de mosaïques 4D ;

Contributions globales :

- prototypage d'un capteur dual pour la génération in-situ des mosaiques 4D ;
- architecture logicielle d'un système de vision pour la numérisation autonome des sites complexes et difficile d'accès par un opérateur humain ;

Les avantages offerts par le scénario d'acquisition des plusieurs mosaïques 4D pour assurer la complétude du site sont multiples :

- recalage des mosaïques 4D en utilisant les informations 2D et 3D pour desambiguer les faux appariements (robustesse aux zones non-structurés, très homogènes ou trop texturées).
- les vues panoramiques offrent l'avantage des pouvoir extraire et apparier des primitives stables à long term ;
- l'interpolation des nouvelles vues en utilisant la géométrie rend possible la navigation entre les différentes mosaïques 4D à travers le web (pour des applications comme les visites virtuelles, annotation d'objets, etc.) ;
- les mesures de distances et l'information de texture peuvent être utilisées conjointement pour l'annotation des données (application web-based pour l'indexation d'objets) ou pour réaliser des mesures 3D sur les données acquises pour l'analyse des scènes

Le système de vision proposé dans cette thèse réalise la numérisation du site, donnant des moyens de perception sur l'environnement aux plateformes inhabités, leur fournissant ainsi des capacités pour réaliser des tâches complexes dans les zones difficile d'accès pour un opérateur humain.

Chapter 1

Introduction and Motivation

Embedded 3D scene representation has been opening the possibility to provide vision-based systems with 3D environment perception, allowing them to be aware when evolving in previously unknown environments. Such systems can dynamically generate 3D scene models and act intelligently on the fly in order to pursue the execution of the required tasks. Therefore, a new generation of vision-based systems raises, with intelligent articulation from perception to action. On the other hand, nowadays a wide number of missions taking place in difficult to access environments rely on heavy operator's intervention for piloting and mission validation. Moreover, worse case scenarios imply limited time and in-situ access which can lead to mission failure even when human operator is included within the loop.

Due to the aforementioned reasons, several research works aim at overcoming a technological step by employing vision-based systems dotted with 3D scene representation and fast decision making capabilities. Such systems are aimed at providing assistance to the operator during the mission, or at supplying the entire mission without requiring human operator's intervention. One of the most challenging application seems to be the possibility to employ such systems in complex missions taking place in high-risk environments in order to avoid endangering human operator's life. Supervising and inspection missions are good examples in which unmanned platforms relieve operator's intervention. The same for monitoring missions in which human watchfulness can unexpectedly fail.

Site survey missions in complex and previously unknown environments are a more interesting example in which the system must perform in-situ 3D digitization, while exploring the site. These systems aim at recovering the 3D geometric and photometric information through the jointly use of 3D laser scanners and/or cameras by means of 3D modeling, giving rise to a general-purpose vision-based system capable to supply several functions: first, the 3D modeling capability allows for a wide range of applications, such as cultural heritage, digital archiving, visual effects for virtual reality applications, urban planning, creation of model data-bases for GIS (Geographical Information Systems) and virtual tourism applications. Second, the digital scene representation it-self provides onboard visual perception which is the basic sense required for accomplishing in-situ complex tasks, such as: scene analysis, maintenance of high-risk sites (mines or tunnels), exploration of underground or underwater prehistoric caves and geological studies of extra-terrestrial sites.

This dissertation aims at providing means for automatic 3D modeling to accomplish site surveys missions in complex and difficult to access environments. Such missions are currently difficult to achieve since traditional 3D digitization techniques are highly dependent on human surveyors. Usually, data collection is followed by a post-processing step

performed off-line during which it is often observed that the 3D scene model in incomplete. Since for such endeavors time and in-situ access are major concerns, the 3D scene model completeness must be ensured in-situ in order to avoid to come back on site to complete data collection.

In our research work we focus on developing a vision-based system capable to generate in-situ complete 3D scene models in complex environments. In order to achieve the aforementioned goal, the system must be endowed with automatic procedures in order to generate dynamically the 3D scene model, while autonomously exploring the site to ensure the 3D scene model completeness. Another main requirement of our research goal is to provide in-situ visualization of the acquired 3D scene model to allow verification either in-situ or by a host wirelessly connected to the target. The in-situ 3D modeling process endows mobile platforms with onboard visual perception allowing them to be aware when evolving in previously unknown environments. Such visual capabilities can be further exploited to supply visual servoing resources to provide feedback control to the system for accomplishing autonomously complex missions without requiring human operator's intervention.

1.1 The in-situ 3D Modeling Problem

The *in-situ* 3D modeling problem is concerned with the automatic environment sensing through the use of active (laser) and/or passive (cameras) 3D vision and aims at generating in-situ the *complete* 3D scene model in a step by step fashion. At each step, the currently generated 3D scene model must be exploited along with visual servoing procedures in order to guide the system to act intelligently on-the-fly to ensure in-situ the 3D scene model completeness.

Our research work is focused on developing a vision-based system aimed at automatically generating in-situ photorealist 3D models in previously unknown and unstructured underground environments from image and laser data. In particular we are interested in modeling underground prehistoric caves. In such environments several issues must be addressed: the absence of reliably extractable and trackable features, the non-reliability of navigation sensors, but also the limited time and in-situ access within which the 3D scene model completeness must be ensured.

Systems embedding active 3D vision are suitable for generating in-situ *complete* 3D models of previously unknown and high-risk environments. Such systems rely on visual-based environment perception provided by a sequentially generated 3D scene representation. Onboard 3D scene representation for navigation purposes was pioneered by Moravec's back in the 1980s [Moravec, 1980]. Since then, Computer Vision and Robotics research communities have intensively focused their efforts to provide vision-based autonomous behavior to unmanned systems, special attention being given to the vision-based autonomous navigation problem. In [Nister et al., 2004], Nister demonstrated the feasibility of a purely vision-based odometry system, showing that an alternative for localization in GPS-denied areas can rely on artificial vision basis. Several research works introduced either 2D and 3D Simultaneous Localization and Mapping (SLAM) algorithms using single-camera or stereo vision frameworks [Durrant-White and Bailey, 2006], [Bailey and Durrant-White, 2006]. While gaining in maturity, these techniques rely on radiometric and geometric features' existence or exploit initial guess provided by navigation sensors (GPS, IMUs, magnetic compasses) employed along with dead-reckoning procedures.

Researchers from Computer Vision and Graphics research communities were introducing the 3D modeling pipeline [Beraldin and Cournoyer, 1997] aiming to obtain photorealist

digital 3D models through the use of 3D laser scanners and/or cameras. Various 3D modeling systems have been developed promoting a wide range of applications: cultural heritage [Levoy et al., 2000], [Ikeuchi et al., 2007],[Banno et al., 2008], 3D modeling of urban scenes [Stamos et al., 2008], modeling from real world scenes [Huber, 2002], natural terrain mapping and underground mine mapping [Huber and Vandapel, 2003a], [Nuchter et al., 2004], [Thrun et al., 2006].

Without loss of generality, the 3D modeling pipeline requires automatic procedures for data acquisition, processing and 3D scene model rendering. Due to the sensors' limited field of view and occlusions, multiple data from various viewpoints need to be acquired, aligned and merged in a global coordinate system in order to provide a complete and photorealist 3D scene model rendering. As for SLAM techniques, the main drawback standing behind the automation of the entire 3D modeling process is the *data alignment* step for which several methods have been introduced.

For systems focusing on 3D modeling of large-scale objects or monuments [Levoy et al., 2000], [Ikeuchi et al., 2007],[Banno et al., 2008] a crude alignment is performed by an operator off-line. Then the coarse alignment is refined via iterative techniques [Besl and McKay, 1992]. However, during the post-processing step it is often observed that the 3D scene model is incomplete. Although data alignment using artificial markers produces accurate results, it cannot be applied to high-risk environments due to time and in-situ access constraints. In addition, for cultural heritage applications, placing artificial landmarks within the scene cause damages to the heritage hosted by the site. The critical need for an in-situ 3D modeling procedure is emphasized by the operator's difficulty to access too small and too dangerous areas for placing artificial landmarks and by the need to validate in-situ the 3D scene model completeness in order to avoid to return on site to complete data collection.

Existing automatic data alignment methods perform *coarse alignment* by exploiting prior knowledge over the scene's content [Stamos et al., 2008] (i.e. radiometric or geometric features' existence, regular terrain to navigate with minimal perception) or the possibility to rely on navigation sensors (GPS, INS, odometry, etc.). In a second step, a *fine alignment* is performed via iterative methods.

Since in our research work the environment is previously unknown, features' existence cannot be guaranteed. In addition, in underground environments and uneven terrain navigation sensors are not reliable and dead-reckoning techniques lead to unbounded error growth for large-scale sceneries. A notable approach reported by Johnson [Johnson, 1997] and improved by Huber [Huber, 2002] overcomes the need of odometry using shape descriptors for 3D point matching. However, shape descriptors' computation requires dense 3D scans, leading to time consuming acquisition and processing, which does not cope with time and in-situ access constraints.

A main part of this dissertation focuses on providing image-laser solutions for addressing the automation of the 3D modeling pipeline by solving the data alignment problem in feature-less and GPS-denied areas. In a second phase, we propose to exploit the world modeling capability along with visual servoing procedures to provide feedback control to the system for visual-guided site exploration in order to ensure in-situ the 3D scene model completeness.

1.2 Image-Laser Proposed Solution: 4D-Mosaic-driven 3D Modeling

Our main target is to provide automatic procedures for a vision-based system (VBS) aimed at performing in situ the entire 3D modeling pipeline: (1) acquire and process data without requiring human operator intervention, (2) generate 3D models in a step-by-step fashion, (3) act intelligently on-the-fly in order to improve the scene's completeness. We close the introduction chapter by drawing the remainder of this dissertation along with it's contributions. This dissertation is structured in eight chapters which gradually present how to achieve in-situ 3D modeling through the jointly use of image and laser data.

Chapter 2 lists several applications standing behind the need of an in-situ 3D modeling system and describes several reality sensing techniques which can be employed for building photorealist and highly accurate 3D models from reality. We briefly describe the available techniques for computing digital scene representation and emphasize their limitations for generating in-situ 3D models in unstructured and large-scale underground environments. We end Chapter 2 by listing the requirements which need to be fulfilled when dealing with the in-situ 3D modeling problem in challenging environments and by proposing several image-laser solutions to overcome the existing methods' shortcomings.

The next five chapters are dedicated to a gradual description of the ARTVISYS system in which we project the proposed image-laser solutions willing to provide unmanned platforms with autonomous world modeling capabilities.

Chapter 3 of this dissertation introduces the **ARTVISYS** prototype, an **ART**ificial **VI**sion-based **SYS**tem designed for automatic 3D modeling in previously unknown environments. We describe the hardware and the software architecture together with the acquisition scenario and the on-board functionalities composing the 3D modeling pipeline. We propose a double-head system, composed by a 3D laser range finder and a high-resolution digital camera mounted on a motorized pan-tilt unit (PTU). The system's design exploits the image-laser complementarity featuring: (a) fast acquisition and processing via low-resolution 3D scans and (b) photorealism provided by high resolution color images. When performing 3D modeling from cameras and 3D laser scanners, an additional open issue is raised when the two sensors have different optical centers. Due to occlusions in either image or laser data, a difficult task to solve is the rendering of the aligned image-laser data. In this dissertation we propose a dual camera-laser system rigidly attached in order to overcome the shortcomings caused by high inter-sensors parallax. We provide reliable data matching techniques in unstructured environments by exploiting 3D geometry and appearance information yielding robustness to false matches, which are inherent when exploiting only radiometric information.

Introducing 4D-mosaics: omnidirectional photorealist 3D models. For each 3D spatial position of the system, the proposed 3D modeling scenario unifies photorealism and dense geometry into four-dimensional (4D) *mosaic views*. In this dissertation we solve for the automation of the 3D modeling pipeline by introducing the *4D mosaics* as fully spherical panoramic views encoding surface geometry (depth) and 3-channel color information (R, G, B). A *4D mosaic* is processed within three steps, each of which being described in Chapters 4,5 and 6 of this dissertation and for which we provide a brief description hereafter.

- First, the 3D laser scanner acquires several partially overlapped scans which are aligned and merged into a fully 3D spherical mosaic. Since our work is concerned

with the 3D modeling in unstructured and underground environments, **Chapter 4** introduces an automatic scan matcher which replaces the two post-processing steps usually performed by the currently existing scans alignment techniques (coarse alignment via manual or GPS pose and ICP-like methods for fine alignment). The proposed method does not rely on feature extraction and matching, providing thus an environment-independent method.

- Second, the motorized PTU acquires a sequence of pose-annotated and high-resolution images, which are further exploited to generate in-situ a Gigapixel color mosaic. Since the nowadays image stitching algorithms present several limitations when dealing with unstructured environments, one of our main concern in this dissertation is the ability to match images in feature-less areas. For this reason, in **Chapter 5** of this dissertation we introduce an automatic multi-view image matching algorithm capable to deal with the absence of reliably detectable and trackable features. The proposed mosaicing system is powered by a global-to-local pairwise image alignment algorithm which recovers the rotations relating the overlapping images in a coarse-to-fine approach. The local motion procedure outputs a list of locally matched *anonymous features* which are later injected in a bundle adjustment engine for multi-view fine alignment. The proposed algorithm combines the state of the art mosaicing techniques in a complementary and efficient fashion providing an environment-independent solution for the image mosaicing task. A powerful viewer with 8-levels of detail is employed to enable virtual tourism applications through the world wide web.

- Third, the 3D mosaic and the 2D color Gigapixel one are aligned and fused into a photorealist and geometrically accurate *4D mosaic*. To do so, **Chapter 6** describes a mosaic-based approach for image-laser data alignment. Since the two sensors are rigidly attached and slightly separated, the algorithm computes the 3D Euclidian transformation, which is essentially a 3D rotation and a residual 3D translation. When solving for the image-laser alignment problem, it is difficult to achieve reliable data matching when the two sensors have different optical centers. Since in our research work the image-laser parallax is negligible, our system overcomes the main drawback caused by occlusions in either image and laser data. We match the 3D mosaic against the RGB Gigapixel one by first estimating the global rotation motion within a pyramidal framework followed by a local patch matching procedure performed only at the highest resolution level of the pyramid (where the image-laser parallax is visible) to estimate a global translational model over the entire mosaic space. The estimated pose is exploited to generate automatically a 4D mosaic (4-channel: red, green, blue and depth) which to our knowledge has not been reported until now. The reconstruction of the 3D scene model is performed in two steps: (i) an integration step exploits the 3D mosaic to generate 2D meshes and (ii) a texture mapping procedure enables the photorealist component of the 3D scene model.

The aforementioned algorithms are composing the **ARTVISYS** system. They can be utilized either as stand alone procedures (i.e. for generating in-situ 3D or 2D Gigapixel mosaics) or as a 3D modeling process for generating in-situ photorealist and highly accurate 3D models encoded as *4D mosaics*.

4D-mosaic-driven in-situ 3D modeling. Due to occlusions, several 4D mosaic views must be autonomously acquired from different 3D spatial positions of the system in order to maximize the visible volume, while minimizing data redundancy. In this disser-

tation we propose a mosaic-driven acquisition scenario to be performed in a stop-and-go fashion. As soon as a 4D mosaic is acquired, the latter is matched against the previously generated ones and integrated into a global reference coordinate system in order to build dynamically the global 3D scene model. To do so, in **Chapter 7** we propose a theoretical solution to supply the 4D panoramic alignment process based on a hybrid 2D-3D criterion to disambiguate data matching in unstructured environments.

The proposed data alignment procedures are composing the automatic 3D modeling pipeline. They are featuring robustness to feature-less and GPS-denied areas, being able to cope with embedded computer requirements and to run in-situ. Validation tests are performed in three difficult to access prehistoric caves situated in France (Moulin de Languenay, Tautavel and Mayenne Science). Experimental tests in structured environments are also presented using real data acquired in Paris city.

Toward autonomous 4D-panoramic-driven site exploration. After solving for the automation of the 3D modeling pipeline, the system is now endowed with visual perception means which must be further exploited along with visual servoing techniques in order to guide the system to ensure the 3D scene model completeness. In particular, the system must be able to build dynamically the 3D scene model, while exploring autonomously the site in order to minimize the occluded areas. This calls for an intelligent 3D modeling procedure which implies the computation of the next best 3D pose of the system from which the next 4D mosaic must be acquired in order to minimize the occluded areas (i.e. solve the Next Best View problem [Klein and Sequeira, 2000]). In addition, the system must be able to navigate from it's current position to the estimated 3D pose via path planning and visual-based autonomous navigation procedures powered by the currently generated 3D scene model. The aforementioned procedures are powered by the global 3D scene model obtained via the 4D mosaic matching algorithm, exploiting the image-laser fusion to provide visual feedback to the 4D mosaicing sensor. For this reason, **Chapter 7** provides image-laser solutions which exploit the 3D modeling pipeline for developing visual servoing procedures in order to generate dynamically the 3D scene models, while autonomously exploring the site.

This allows us to evaluate the 4D mosaicing sensor's potential in solving for the automatic in-situ generation of complete and photorealist 3D models in high-risk and complex environments, without operator intervention.

Since the autonomous 3D world modeling capability is intrinsically related to the system's autonomy, we provide a visual-autonomy model which is instantiated to the autonomous site digitization and exploration problem. The software architecture has as main nucleus the 4D mosaicing sensor which used along with visual servoing procedures acquires and integrates on-the-fly 4D mosaic views in order to ensure the 3D scene model completeness.

Chapter 8 closes this dissertation by summarizing our research proposals and by giving short- and long-research directions exploiting the aforementioned results.

Chapter 2

Why and How to perform In-situ 3D Modeling?

We start this chapter by motivating the need for in-situ 3D modeling procedures and listing a variety of applications requiring for an automatic in-situ 3D digitization procedure.

The next three sections present the available technological means allowing to tackle the in-situ 3D modeling problem. Section 2.2 presents the existing techniques encoding high-detailed photorealist digital scene representations of the reality. The next section describes the technical background of the 3D modeling process and introduces the main key issues standing behind the automation of the 3D modeling pipeline. Since the in-situ 3D modeling task implies automatic environment mapping and localization capabilities, Section 2.4 introduces the Simultaneous Localization and Mapping problem and lists several existing solutions, emphasizing their limitations for dealing with unstructured, large-scale and difficult to access environments.

We end this chapter by proposing several image-laser joint solutions for the aforementioned techniques aiming to overcome their main drawbacks when dealing with unstructured and difficult to access environments. More important is that the proposed solutions are the main ingredients which are likely be embedded on a vision-based system in order to achieve autonomously in-situ 3D modeling tasks.

2.1 Why In-situ 3D Modeling?

Systems embedding digital scene representation are well suited for many civilian and military applications. The most challenging one seems to be the possibility to employ such vision-based systems (VBSs) in complex missions tacking place in unknown and complex environments. To do this, onboard perception and decision capacities need to be improved in order to provide autonomous behavior during the mission. A great number of missions in hostile environments are still unavailable due to the autonomous system feasibility, which is limited and thus request operator's intervention.

Embedded 3D scene representation has been opening the possibility to provide unmanned systems with 3D environment perception, allowing them to be aware when evolving in previously unknown environments. Such systems can dynamically generate 3D scene models and act intelligently in order to pursue the execution of the required tasks. Therefore, a new generation of unmanned system arises, with intelligent articulation from perception to action, enriched with new visual-based capabilities: see, detect, plan and

avoid within an information and decision network.

Designing systems embedding 3D vision is one of the nowadays multi-disciplinary challenges which puts together researchers from Computer Vision and Robotics communities. Such VBSs are about to be employed in high-risk environments for performing complex missions. Supervising and inspection missions are good examples in which robots equipped with adapted sensors relieve operator's intervention, the same for monitoring missions in which human watchfulness can unexpectedly fail. The most challenging ones seem to be the possibility to employ unmanned systems in complex interventions, such as disaster response as well as searching and rescuing operations.

Several 3D modeling systems reported tractable solutions for large-scale objects or monuments relying on heavy human operator intervention. 3D modeling of large-scale urban scenes was successfully solved by Stamos [Stamos, 2001] by imposing orthogonality constraints on scene's content. While improving the state of the art, these methods remain limited by the application type: 3D modeling of structured environments. Very few research works attempt to generate automatically in-situ photorealist 3D scene representation in complex and unstructured environments for field robotics applications. Photorealist digitization and understanding of the heritages hosted by prehistoric caves, maintenance of tunnels or mines, site surveys and exploration of underwater, under-ice or extra-terrestrial environments are several applications requiring for an accurate and automatic 3D modeling procedure.

Through the following description we review several ongoing 3D modeling projects promoting the aforementioned applications in order to emphasize the critical need for developing an autonomous vision-based system capable to generate in-situ photorealist and complete 3D models of its surroundings.

Cultural heritage of large-scale sites. Figure 2.1 a) illustrates a site survey campaign performed by the French Mapping Agency[1] and ESGT[2] in the Mayenne Science prehistoric cave (France) by human surveyors to create exhaustive maps for cultural heritage archiving and virtual tourism applications. Figure 2.1 a) emphasize the operator's difficulty to access too small areas. A first difficulty which the human operator must face is finding the adequate number of viewpoints from which the environment must be sensed in order to ensure the 3D scene model completeness. A second concern is that the data acquisition time must fit the granted in-situ access time. Due to the complexity of the site, the standard assumptions for automating the 3D modeling process, such as parallelism, perpendicularity or symmetry are not adequate. Therefore, the data processing step is performed off-line by human operators which often observe that the final 3D scene model is incomplete.

Figure 2.1 b) depicts a site survey campaign undertaken by the French Mapping Agency and ENSG[3] in the Tautavel prehistoric cave, France. The purpose of the site survey was to produce highly accurate cross sections of the 3D model to allow archeologists to analyze the heritage hosted by the cave. During the mission, constraints for surveyors' access and system's positioning were imposed in order to avoid damaging the heritage hosted by the cave. In addition, it was difficult and sometimes impossible to place the system in particular viewpoints to acquire the missing data in order to ensure the 3D scene model completeness. During such endeavors, airborne and ground-based sensors are likely to be employed in order to preserve the cultural heritage hosted by the site. The same problem

[1] Institute Géographique National - IGN
[2] École Supérieure des Géomètres et Topographes
[3] École Nationale Supérieure de Géographie

was reported in [Ono and Ikeuchi, 2007] when attempting to create digital 3D models of Bayon Temple shown in Figure 2.1 c). Due to the high architectural complexity of the site, ground-based sensors are not suitable for sensing too small or too narrow areas, (see for an example Figure 2.1 d)). For this reason, authors in [Banno et al., 2008] designed an aerial mobile sensing platform using a balloon as a base, resulting in a Flying Laser Range Sensor (FLRS). The sensor was employed during the Bayon Temple digitization mission in order to accomplish the 3D scene model completeness.

This dissertation aims at endowing a vision-based system with automatic environment sensing functionalities for generating in-situ complete and photorealist digital 3D models in complex and underground environments. We validate our system on real tests performed in three prehistoric caves situated in France for an ongoing project promoting cultural heritage, virtual reality and scene understanding applications. Such a vision-based system has a double utility: on one hand, when the site allows easy access, the system performs in-situ data processing and view planning, providing guidance and assistance to human surveyors for completing the 3D scene model. On the other hand, when site surveys missions are performed in difficult to access environments, the system must be able to digitize and explore autonomously the site through path planning and autonomous navigation procedures based on image and laser data input.

3D Mine mapping. Nowadays, abandoned subterranean voids, especially mines represent a threat to their surroundings. They present risk inundation by water or hazardous gases. A way to combat these risks is to catalog the exist and characterize these voids. This operation relies currently on the available maps and on other geophysical techniques. Several research groups [Baker et al., 2004b] [Nüchter et al., 2004] are focusing on developing mobile platforms capable to autonomously map and explore abandoned mines. Figure 2.1 f) illustrates the Groundhog mobile platform developed by [Baker et al., 2004b] at Robotics Institute-Carnegie Mellon University for autonomous 3D mine mapping. In the proposed framework, the scanning device provides 3D terrain modeling for obstacle avoidance and path planning, whilst odometry and scan matching provide feedback to the motion controller and allow Groundhog to follow a specific path. Nevertheless, when using odometry on such uneven terrains as shown in figure 2.1f), a major issue for safe navigation arises due to slippage on sloping terrain.

Planets' Exploration. Since 2003, computer vision is being successfully applied in Space, within the Mars Exploration Rover (MER) mission with Spirit and Opportunity, twins geologist robots. Employing vision for taking onboard decisions is a key aspect needed to accomplish complex missions taking place in hostile environments. For instance, for large scale site surveys missions performed on Mars [Paar et al., 2009], data acquisition is performed automatically, transmitted on Earth and processed by computer vision experts. Furthermore, rovers are receiving commands from human operator for path planning, navigation and obstacle detection and for dealing with unpredictable situations. This highly-dependency on human interaction is subject to memory bandwidth and communication latency, causing unmanned systems' failure to react rapidly to unpredictable situations. Recently, researchers from Jet Propulsory Lab have reported in [Mathies et al., 2007], [Johnson et al., 2007] the use of stereo-vision, visual-odometry and feature tracking for rover navigation and lander's velocity estimation before touchdown. As for the mine mapping case study, researchers are planning to address one of the performance's issues for MER navigation: wheels' slippage. Since visual odometry solves this problem partially, researchers are now attempting to use learning algorithms to predict the amount of slip to expect from the appearance and from slope angle of hills in front of the robot.

So far we introduced several practical issues revealing the need of a unmanned vision-based system for in-situ 3D modeling. The next section reviews the existing techniques for generating digital scene representation of the reality through the use of active (3D laser scanners) and passive (cameras) 3D vision devices.

2.2 Digital Scene Representation Techniques

Digital scene representation has been intensively addressed within Computer Graphics, Computer Vision and Remote Sensing research communities and three different approaches have been introduced. Researchers in Computer Graphics reported image-based solutions defined as *image-based rendering* (IBR) techniques. Computer Vision and Remote Sensing research communities employ color cameras and 3D laser scanners aiming to recover the 3D geometry from reality, giving rise to *passive* and *active* 3D vision senses, respectively. Through the following subsections we provide a brief description of the aforementioned techniques, while the last subsection emphasizes their complementarity and illustrates how we propose to combine them in this dissertation to overcome each one's shortcomings.

2.2.1 Image-based Rendering

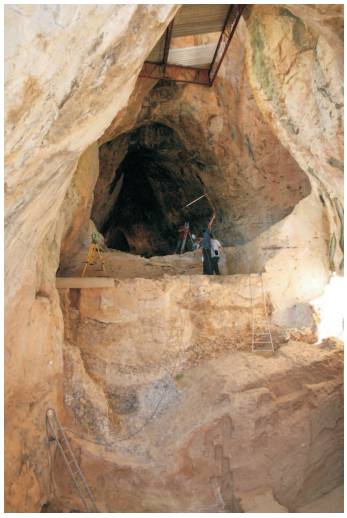
IBR methods exploit exclusively an image sequence together with their associated camera calibration matrix, when available. These techniques are interested in generating synthetic views by interpolating between original images, without aiming to provide the 3D geometry of the sensed surface. Computer Graphics research community aims at producing high-detailed photorealist rendering for virtual tourism and augmented reality applications paying the price of heavy acquisition setup, expensive computation time and semi-automatic frameworks. The proposed rendering pipeline implies the geometry and viewpoint recovery as well as texture, lighting, and shading information.

Image morphing is a widely used IBR technique which generates transitions between original images. Supposing that the rigid transformation which lies between two views is known, image morphing procedure generates an animation which smoothly transforms the initial view toward the final one. Image morphing methods were firstly employed in the early 1990s by Michael Jackson in the clip Black or White for face morphing.

IBR techniques are being successfully employed for Computer Generated Imaging (CGI) to create special effects for movie industry. They produced the science fiction boom at Hollywood starting in the early 1990s with Steven Spielberg's Jurassic Park, in which CGI was used to create special effects, by combining them with live action.

Light Field Rendering (LFR) techniques were introduced by Marc Levoy and Pat Hanrahan in the Computer Graphics community in 1996 [Levoy and Hanrahan, 1996] [Gortler et al., 1996], [Buehler et al., 2001]. Their proposed application was the IBR. The key to this technique lies in interpreting the input images as 2D slices of a 4D function - the light field. This function completely characterizes the flow of light through unobstructed space in a static scene with fixed illumination. Since these methods do not estimate the real 3D structure behind the images, they are unable to cover every possible novel viewpoint and a large number of views is needed to produce undistorted renderings. Therefore, in order to capture, encode and display real 3D scenes based on the light-field principle massive processing power is required. Finally, they cannot model novel illumination conditions.

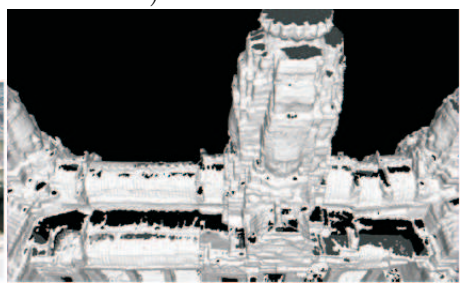
a)



b)

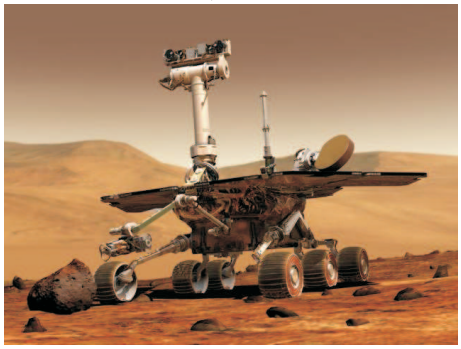
c)

d)



e)

f)



g)

Figure 2.1: In-situ 3D Modeling in Challenging Environments. (a)Data acquisition scenario in Mayenne Science Prehistoric Cave (France) ©IGN-ESGT, (b)Data acquisition scenario in Tautavel Prehistoric Cave (France) ©IGN-ENSG, (c)Bayon Temple - kingdom of Cambodia, (d)Bayon Temple top-down view - image taken from [Ono and Ikeuchi, 2007], (e)Missing data in the 3D model result using only ground-based acquisition platforms - image taken from [Banno et al., 2008], (f) Groundhog mobile platform for mine mapping (Carnegie Mellon University) - image taken from [Baker et al., 2004b], (g) Mars Exploration Rover - Spirit geologist robot ©NASA-JPL.

2.2.2 3D Vision

Following the capturing device's type, two research directions were introduced in Computer Vision and Remote Sensing research communities. Computer Vision research works lead to an extensive use of cameras for interpolating 3D coordinates of a point in space (passive methods), while remote sensing approaches were oriented toward the use of 3D laser scanners (active methods). Both of them provide a digital description of the reality in a defined data structure refereed to as *3D model*. In opposite to IBR methods, both passive and active techniques allow to recover the 3D geometry of the sensed surfaces, which is one of the main concerns of this dissertation, providing the possibility the endow unmanned platforms with 3D vision.

2.2.2.1 Passive

Passive 3D modeling methods are taking as input either calibrated or un-calibrated images and estimate the 3D surface geometry via image motion estimation methods. They are referred to as *image-based modeling* (IBM) techniques. Passive methods employ either single- or multi-camera systems, depending on the application type.

Structure from motion (SFM). In opposite to multi-camera systems, SFM methods require single video camera use, thus offering a low cost and portable solution. They depend on algorithms able to reliably detect and track stable radiometric 2D features over a sequence of images. SFM methods take as input a set of tracked 2D image features and estimate their 3D location and the camera poses simultaneously. When Tomasi and Kanade [Tomasi and Kanade, 1992] reported a batch algorithm for solving the SFM problem, there was when the SFM method was first spread into the Computer Vision community, after being pioneered in the photogrammetry community, in which it is referred to as bundle adjustment [Triggs et al., 1999].

Since the 3D structure estimates are computed only relative to the tracked feature points, the final result is a sparse 3D point cloud. Reported solutions focus on providing dense 3D structure in a separate step [Pollefeys et al., 1998], but poorly textured and unstructured areas are still an open issue and feature matching over non-consecutive frames remains a difficult problem to solve.

Multi-cameras. Multi-camera systems are putting together range and intensity images. For poorly structured scenes, stereo-vision systems provide dense but noisy 3D point clouds of the sensed surface. A multi-camera system allows for noise reduction by integrating multiple range estimates independently computed [Kanade et al., 1995]. At Carnegie Mellon University researchers have reported in [Rander, 1998], [Saito et al., 1999] an extreme case in which the scene is entirely surrounded by cameras giving rise to a virtualized room. In a first step, range estimates are obtained using a multi-baseline stereo framework performed on multiple subsets of the cameras. Then, prior camera calibration knowledge is used to automatically register range images. A real-time multi-cameras system is reported in [Rander, 1998] which allows for dynamic scenes 3D modeling. In [Pollefeys et al., 2008] authors reported real-time 3D reconstruction from a multi-camera system mounted on a vehicle exploited along with navigation sensors (GPS, IMU).

Shape from silhouette. Back in the '70s, silhouettes were firstly used for 3D modeling purposes by Baumgart [Baumgart, 1974], while the silhouettes-based surface's properties were first studied by Laurentini [Laurentini, 1994], defining the visual hull concept. The visual hull algorithm takes as input a sequence of images of an objet acquired from known camera positions. The 3D model computation rely on complex volume intersection tech-

niques needed to discard parts of the viewing volume which do not belong to the object using background constraints. In [Matusik, 2000] authors reported an efficient method to perform visual hull computation entirely in the 2D image space.

Space carving. Instead of exploiting background region constraints, shape from carving methods rely on photo-consistency constraints of the object [Kutulakos and Seitz, 2000], [Matsumoto et al., 1997]. These methods generate a volume and iteratively discard voxels based on color coherency in all images. In [Broadhurst et al., 2001] authors proposed a probabilistic approach for voxel occupancy. The output is a voxel-based 3D model difficult to transform in a 3D mesh representation. In [Zhang and Seitz, 2001] and [Yezzi et al., 2002] authors attempt to solve this problem. Nevertheless, another issue arises when using photo-consistency: namely, the sensitivity to light color variation when comparing absolute color values. A different way of exploiting color by studying local variation via cross-correlation techniques can be found in [Faugeras and Keriven, 1998], [Sarti and Tubaro, 2002]. By formalizing the problem using an approach such as visual hull, it is not possible to model complex shapes (such as concavities) while space carving methods can.

Shape from silhouette and space carving methods are usually employed for small-scale object 3D modeling when the camera captures the entire objet from known viewpoints using a controlled background. Due to the aforementioned constraints, these methods cannot be employed in our research work.

Shape from shading. These methods [Horn and Brooks, 1989] exploit the diffusing properties of Lambertian surfaces making use of several acquisition constraints (orthographic cameras or punctual light sources), making them useless for practical applications.

Fusion-based approaches. Several authors exploit jointly color and additional information such as: silhouettes [Liedtke et al., 1991], [Fua and Leclerc, 1996], [Matsumoto et al., 1997], [Cross and Zisserman, 2000], [Isidoro and Sclaroff, 2003] , shading [Fua and Leclers, 1995], [Jin et al., 2000] or radiance [Yezzi and Soatto, 2001], [Soatto et al., 2003], but extracting 3D information still remains an open issue. Hernandez [Hernandez, 2004] employs a novel data fusion technique for silhouettes-stereo joint 3D modeling which leads to high quality reconstruction results.

2.2.2.2 Active

In opposite to IBM techniques, the *3D scanning* process produces 3D digital representation of the surface geometry encoded as 3D cartesian coordinates expressed wrt the sensor's 3D position. Laser-range sensing implies a physical contact between a controlled source of energy and the 3D surface, leading to a response measurement. These techniques proceed by sampling regularly the nearby surfaces of a scene and by emitting regular energy patterns beams into the scene to measure the visible 3D geometry wrt to the sensor's position and field of view. Range-sensors exploit the spatial and time properties of the reflected laser beam to compute the depth of the 3D point in the beam direction. Appendix A.1 resumes the main laser-range sensing techniques, while further details on early range-sensing can be found in [Besl, 1988], [Poussart and Laurendeau, 1989].

The scanning device employed in the research work presented in this dissertation falls in the category of time-of-flight 3D scanning techniques. Basic features defining a 3D laser scanner are the scanning resolution and its accuracy. The absolute value of error increases with the distance between the sensor and the surface to be scanned. Following the application type, it is possible to set the laser to acquire either one-shot scans, which

are fast but less accurate, or multiple-shots scans, which are more accurate but time consuming. The resolution bounds the dimensions of the 3D scene model and influence the choice of the data structures employed for merging multiple scans into an unified scene representation.

2.2.3 Taxonomy and Image-Laser Joint Solutions

After reviewing the existing approaches allowing to produce digital representations from reality, Figure 2.2 provides a taxonomy of these methods with respect to our main interest: recover reliably 3D geometry and appearance information in unstructured environments for generating in-situ photorealist and accurate digital models in difficult to access environments.

The first class, **IBR methods** are mainly focused on generating novel views from original images. In order to generate novel views geometrically coherent, a high amount of images need to be acquired under fixed illumination conditions. Therefore, in order to capture, encode and display real 3D scenes based on the IBR techniques, massive processing power is required. In addition, in high-risk and difficult to access environments, a human operator cannot access into the site to set fixed illumination conditions. A possible way to improve IBR techniques' performances is to acquire simultaneously geometric information with a 3D laser scanner, which allows to accurately interpolate between views, overcoming the need of a high amount of views.

The second class, **passive 3D vision techniques** detain a main advantage over the IBR methods which is the fact that they do aim at recovering accurately the 3D geometry of the sensed surface. Reported solutions rely on feature extraction which yield reliable and accurate results for environments rich in radiometric and geometric features. Since our research work deals with the 3D modeling problem in unstructured environments, we cannot guarantee the above feature-existence hypothesis.

In previously unknown environments, in absence of stable detectable and trackable features and in presence of unstructured and texture-less areas, such algorithms fail during the data alignment process. Furthermore, in feature-less areas they provide fast, but sparse and noisy 3D point clouds, and multi-cameras systems for dense 3D mapping are highly complex, not portable and computationally extremely expensive. In unstructured environments, image matching techniques lead to outliers, which are inherent when employing only radiometric criteria for image matching. A possible solution to this problem is to employ a joint criterion, using both: radiometry - from color cameras, and 3D geometry - from 3D laser scanners, to disambiguate data matching. Such a technique is proposed in Chapter 7 for 3D model matching purposes.

Active 3D vision techniques are mainly limited by the fact that they cannot capture high-quality color information required to enable the photorealism component of the 3D scene model. This justifies the recent trends reported in computer graphics and vision research communities, which employ image-laser fusion to yield photorealist and accurate 3D models.

Image-laser joint solutions. Figure 2.2 illustrates our choice highlighted in blue and concerns essentially the fusion of active and passive 3D vision techniques for capturing both, 3D geometry and low cost color information, allowing to improve the aforementioned techniques' performances: (1) improve IBR techniques with 3D geometry from 3D laser scanner needed for interpolating more accurately novel views with low cost processing, (2) overcome IBM techniques shortcomings raised by image matching algorithms sensitive

to unstructured environments by employing a joint radiometric and geometric criterion to ensure reliable data matching in feature-less environments, and (3) enrich 3D laser scanners capabilities with color information for photorealist 3D modeling large-scale and unstructured environments. To this end, Chapter 3 of this dissertation proposes a dual environment sensing device endowed with a high-resolution color camera rigidly attached to a 3D laser-range-finder which allows the implementation of the aforementioned image-laser solutions.

2.3 The 3D Modeling Pipeline

This section covers the background material of the 3D modeling pipeline and emphasizes the major issues standing behind its fully automation in order to allow for in-situ 3D modeling. Figure 2.3 illustrates the existing 3D modeling pipeline consisting in laser and image data acquisition, alignment and 3D scene model rendering. The following description briefly summarizes each step of the traditionally employed 3D modeling pipeline, while means for its automation proposed throughout this dissertation are resumed in the final section of this chapter.

2.3.1 Data Acquisition

Following the capturing devices' positioning, two types of acquisition systems can be distinguished. Systems embedding a color camera and a laser scanner rigidly attached (RACL) and those performing data collection with a freely moving camera and laser (FMCL). For both cases, due to the sensing devices' limited field of view, multiple partially overlapped scans and images need to be acquired either automatically or by an operator in order to ensure the 3D scene model completeness.

Figure 2.4 illustrates a classification of the existing data acquisition procedures with respect to the application type. There are manual - usually designed for 3D modeling of small-scale objects [Huber, 2002], semi-automatic methods suitable for large-scale objects [Levoy et al., 2000] and monuments [Ikeuchi and Sato, 2001], [Banno et al., 2008], teleoperated and automatic acquisition scenarios exploiting calibration constraints [Huber and Vandapel, 2003b], navigation sensors [Nuchter et al., 2005], [Thrun et al., 2003] and path planning procedures [Klein and Sequeira, 2000] for 3D modeling and exploration purposes in large-scale environments.

When dealing with the in-situ 3D digitization problem in unstructured and difficult to access environments, several key issues have to be taken into account within the acquisition scenario, such as: the non-reliability navigation sensors for GPS-denied areas and non-flat terrain, the impossibility to place artificial landmarks (for data matching purposes) and the need for intelligent 3D digitization techniques in order to ensure ins-situ the 3D scene model completeness.

2.3.2 Data Alignment

Since each data is described in the sensor's coordinate systems, the 3D scans and color images need to pass through a so-called alignment process which integrates all data into a global 3D scene model. The *data alignment* process consists in estimating the sensor's poses from which each scan (or image) was acquired with respect to a global coordinate system. A 3D pose estimate is a rigid body transform which encodes 6 degree-of-freedom (DOF),

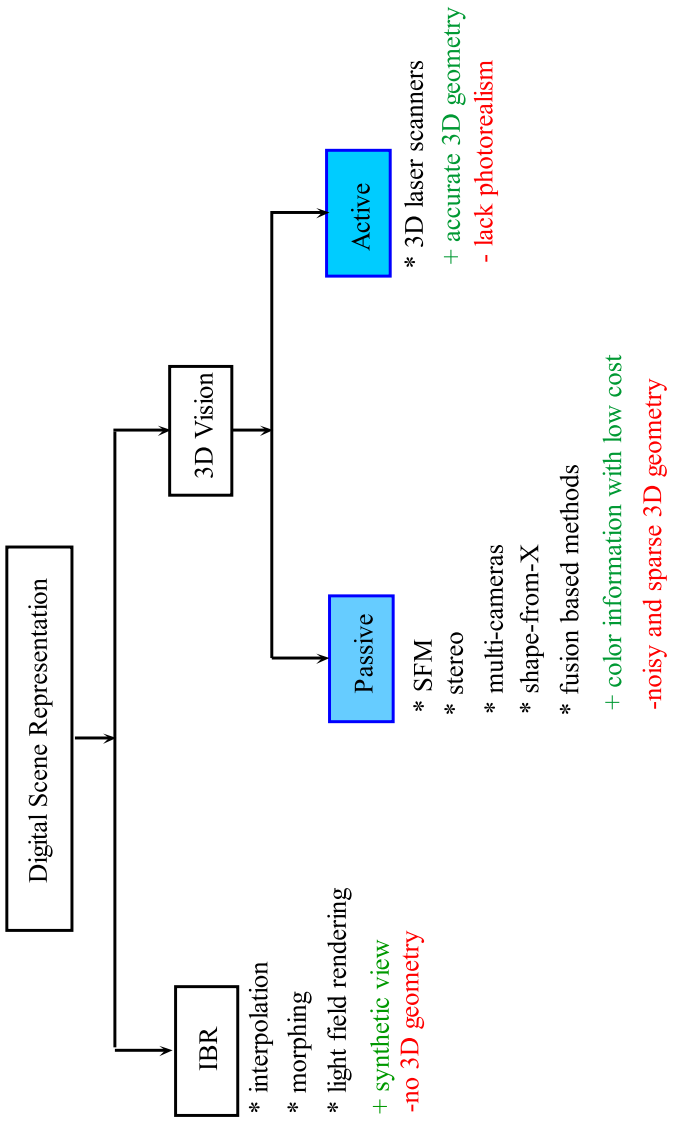


Figure 2.2: A taxonomy of the existing digital scene representation techniques wrt their performances for recovering 3D geometry and photometric information in unstructured environments. While cameras-based methods allows to capture high resolution color information rapidly, 3D lasers scanners captures highly accurate and densely sampled 3D geometry in unstructured environments.

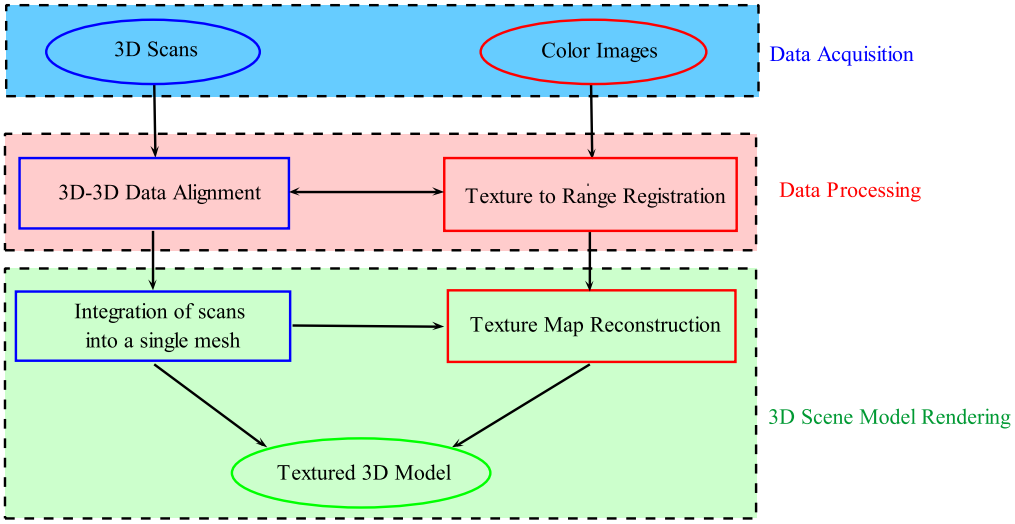


Figure 2.3: The 3D modeling pipeline.

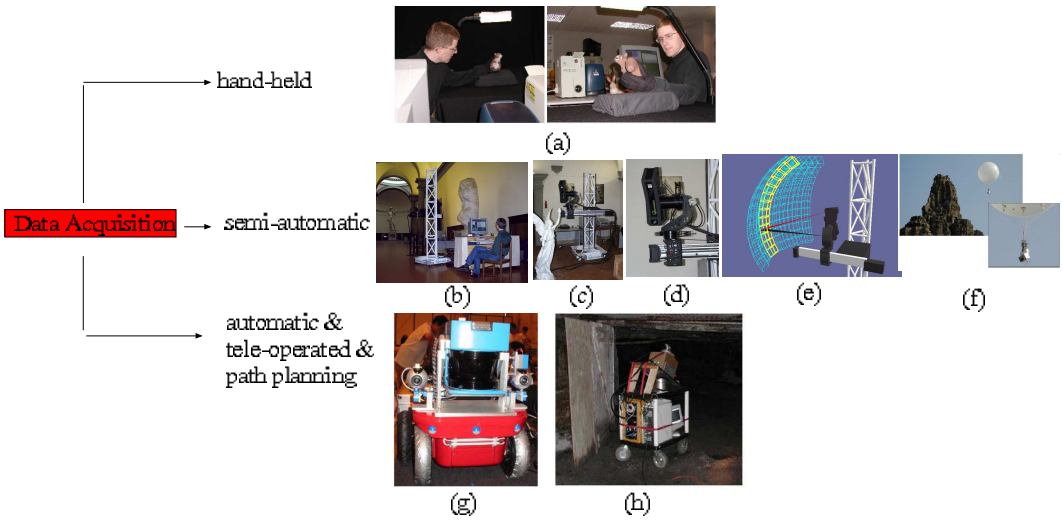


Figure 2.4: Data Acquisition. a) hand-held acquisition for small-scale objects modeling, image taken from [Huber, 2002], b)-f) semi-automatic acquisition for large-scale objects modeling b)-e) capturing devices-images taken from [Levoy et al., 2000], f)3D modeling of large-scale monuments using a flying range sensor, image taken from [Banno et al., 2008], tele-operated or automatic acquisition via path planning: g)Kurt3D - multi-purposes mobile platform - image taken from [Nuchter et al., 2005], h)cart-mounted laser for 3D mine mapping - image taken from [Huber and Vandapel, 2003b].

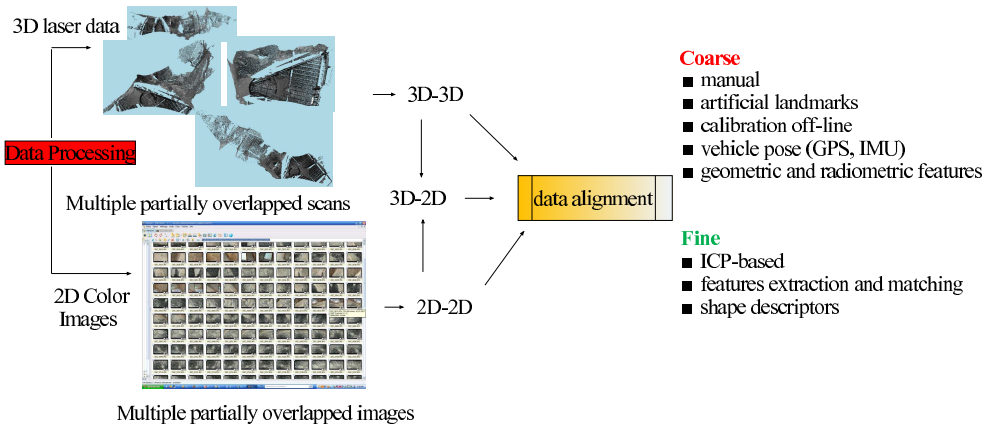


Figure 2.5: Data alignment between overlapping color and range images through coarse-to-fine alignment. Following the application type, we can distinguish two methods providing data matching: either manual or semi-automatic for large-scale objects and automatic based on navigation sensors for large-scale environments.

(three for rotation: yaw - ψ, pitch-φ and roll-θ, and three for translation $[t_x, t_y, t_z]^T$) and allows to align two overlapped images or scans in a global coordinate system.

Pose estimates computed wrt a global coordinate system give rise to *absolute poses*. Otherwise, they can be also computed wrt an arbitrary view, leading to *relative poses'* estimates. The pose estimation process performed between two partially overlapped scans or images is refereed to as the *pair-wise alignment*. When multiple overlapped images or 3D scans are aligned wrt a global reference coordinate system, a *multi-view alignment* process is performed. Without loos of generality, being given a sequence of partially overlapped scans and color images, the data alignment process requires three different procedures:

- pair-wise and multi-view scans alignment (3D-3D);
- pair-wise and multi-view image alignment (2D-2D);
- texture to range alignment (2D-3D).

Figure 2.5 summarizes the current approaches available for performing the data alignment task. Usually, the pair-wise data alignment is performed within two steps: the first step is the *coarse alignment*, usually called *data matching*, followed by a *fine alignment* refereed to as *data registration*.

For both steps, different solutions have been introduced, following the application type. Systems aiming to produce high-detailed 3D models of large-scale objects includes human operator's intervention in the 3D modeling loop, while systems operating in large-scale indoors or outdoors environments aim at embedding fast and accurate 3D modeling algorithms onboard unmanned mobile platforms by coupling data with vehicle's pose provided by navigation sensors (GPS, IMU, odometry) to perform a coarse alignment. For both categories, the fine alignment is performed via radiometric and/or geometric features detection and matching. A widely employed technique for scans registration is the Iteratively Closest Point algorithm [Besl and McKay, 1992] which was pioneered by Besl in the early 1990s.

There are several methods for data matching, including calibrated pose measurements, manual matching and verification [Turk and Levoy, 1994], [Ikeuchi and Sato, 2001], me-

chanical measurement (robot arm or a controlled turn-table). However, calibrated pose methods are generally limited to small-scale objects [Blaise and Levine, 1995]. Researchers from Computer Vision and Graphics research communities have reported very ambitious projects aiming to produce photorealist and highly accurate 3D models of large-scale monuments for cultural heritage applications. The proposed framework employ either manual or semi-automatic procedures for the data acquisition and processing steps [Ikeuchi et al., 2007], [Banno et al., 2008], [Levoy et al., 2000] yielding accurate results in a controlled laboratory environment.

Techniques relying on manual data matching and verification are time consuming, since the user must search for corresponding feature points in the view by hand. Other data matching methods are making use of artificial landmarks previously placed in the environment by human surveyors. Such methods cannot be employed in our research context due to two reasons. For cultural heritage applications, the environmental modification can damage the heritage hosted by the prehistorical site. Secondly, for site surveys missions taking place in difficult to access environments such scenarios are not feasible since they endanger human surveyor's life and they do not cope with time and in-situ access constraints.

Several research works attempted to automate the data alignment step by exploiting a-priori constraints wrt the scene structure, such as the existence of radiometric and geometric features [Stamos et al., 2008], [Zhao et al., 2005], [Dias et al., 2003]. While improving the state of the art, these methods remain limited to the application type: 3D modeling in structured environments. Johnson [Johnson, 1997] and Huber [Huber, 2002] employ shape descriptors for coarse alignment, overcoming the need for odometry. Nevertheless, shape descriptors require accurate normals estimates which are difficult to obtained when acquiring fast one-shot and low-resolution 3D scans, such as in our case.

Due to the aforementioned reasons, the data alignment step represents the major bottleneck standing behind the automation of the 3D modeling pipeline. The most general formulation of the data alignment problem makes no assumption on features' existence nor on navigation sensors employability, being extremely hard to solve. In addition, the main challenge is to achieve coarse and fine alignment within one step.

2.3.3 3D Model Rendering

After the alignment stage the 3D modeling process outputs a digital representation of the system's surroundings which is stored in the system's memory, being directly exploitable by an unmanned system without requiring visual display.

Nevertheless, for humans a visual representation is required and for this reason a 3D model rendering process is performed in two stages. The first step merges all models into a global 3D scene representation. This process is usually referred to as *data integration*. In the second stage, texture maps created from color images collected by digital cameras may be added to enable the photorealism of the geometric model via a *texture mapping* procedure.

Data Integration. The data integration stage aims at merging all scans into an unified and non-redundant scene representation. To this end, overlapping areas leading to partial occlusions in the global scene model are detected via ray tracing methods [Reshetov et al., 2005] and carved in order to produce a global 3D scene model.

Depending on the application type, it is possible to improve the 3D model quality by performing additional pre-processing. For instance, a triangular mesh may be computed

and further simplified to reduce the number of points while preserving the shape. A significant amount of research has been done in this direction.

Existing methods can be classified following the input data type. Methods receiving *a set of range images* are exploiting 3D points for extracting more information (surface normal, partial connectivity, sensor position) for better estimating the actual surface. We recall here Delaunay-based techniques [Edelbrunner, 1998], surface-based methods [Soucy and Laurendeau, 1995], [Bernardini et al., 1999], [Gopi et al., 2000], volumetric approaches [Curless and Levoy, 2000], [Hilton et al., 1996], [Lorensen and Cline, 1987], [Reed and Allen, 1999] and deformable surfaces methods [Terzopoulos et al., 1988] , [Pantland and Sclaroff, 1991], [Whitaker, 1998], [Gomes, 2000], [Eck et al., 1995], [Peters, 1994]. A high amount of research work was reported within the Computer Graphics research community yielding accurate surface reconstruction results.

These methods have several shortcomings to our concern. First, when dealing with complex environments, mesh simplification procedure may lead to a loos of details and artifacts within the 3D model. A second issue is raised by the practical implementation due to the high computational complexity and the existing computing resources. Although the aforementioned methods yield accurate 3D models, they require a heavy interactive process which does not cope with time and in-situ access constraints imposed by difficult to access environments. On the other hand, they can be integrated into the final 3D model rendering process which can be performed off-line by a host wirelessly connected to the target. For *unorganized point clouds*, more general methods have been introduced, but they lack robustness in presence of noise and outliers.

Texture Mapping. First, texture maps are created by merging all the acquired images into a single non-redundant map over the entire object via image stitching and blending techniques [Szeliski, 2006], [Brown and Lowe, 2007]. In a second stage, a texture mapping procedure is performed by assigning a color to each vertex composing the mesh.

When performing texture mapping, an important issue arise when the camera and the laser have different optical centres (i.e. FMCL systems). The main problem is how to render the registered image-laser data due to occlusions in either image or laser data. Several authors attempt to solve this problem by providing efficient solutions for the view-dependent texturing, without handling the problem of texture occlusions [Pulli et al., 1997], [Devebec et al., 1996].

2.4 Simultaneous Localization and Mapping

Endowing unmanned systems with in-situ 3D modeling capacities requires autonomous behavior and automatic procedures to (i) dynamically generate 3D scene models, (ii) to localize it-self within the generated map and (iii) to act intelligently on the fly in order to assure the 3D scene model completeness.

The first two functions are related to the robot's capacities *to build a map* of a previously unknown environment, while *localizing* it-self within the generated map. Localization and mapping are intricately coupled problems: automatic map generation requires platform's localization, while position's estimation is impossible without a map. Researchers from Robotics community exploit this mutual dependency by attempting to solve for the Simultaneous Localization and Mapping (SLAM) problem, also known as Concurrent Mapping and Localization (CML).

In order to solve for the third problem, visual servoing procedures are needed in order to control the system for accomplishing complex tasks in hostile environments, such as: scene

understanding, site inspection, monitoring, site surveys, searching and rescuing missions or disaster response. For instance, in our research work the 3D modeling system must be capable to sense the entire environment and to assure the scene's completeness in-situ, by maximizing the amount of information while minimizing the occluded areas and data redundancy.

This section resumes the SLAM process and the existing solutions, while focusing on methods designed for unstructured, large scales and difficult to access environments for field robotics applications.

2.4.1 The SLAM Process

The SLAM framework aims building a map of a previously unknown environment, while simultaneously estimating the platform's location with respect to the generated map. Initially, the map and the system's location are unknown and the environment is populated with artificial or natural landmarks. The system is supposed to embed the necessary devices to sense the environment relative to the landmarks and to provide platform's positioning.

The most popular environment sensing devices employed to supply SLAM frameworks are sonars, 3D laser range finders and color cameras. Additional navigation sensors are employed to localize the platform wrt a local (proprioceptive sensors: odometry, IMU) or a global (exteroceptive sensors: GPS, lasers, cameras) coordinate system.

Exteroceptive devices collect information about the environment in which the system evolves and provide absolute localization of the system. The most popular exteroceptive localization device is the Global Positioning System (GPS) which employs at least four satellites and performs position estimation via triangulation methods. Other localization devices rely on distance measure: sonars, lidar (infrared), radar. Such sensors measure the time of flight t of an impulse $t = \frac{2d}{v}$, where d denotes the distance and v the speed of sound which is known. The lidar and the radar allows for distance computation using the phase difference: $\frac{\phi}{2\pi} = \frac{2d}{v}$. Video cameras capture powerful representation of the environment surrounding the robot providing sparse features or depth information when stereo montage is available.

Proprioceptive localization devices provide relative positioning wrt the robot's referential system. A good example of relative positioning technique is odometry, which computes the distance traveled by the platform by counting the wheels' tours. It allows for accurate localization for short distances, providing high-frequency data with low cost. Nevertheless, such techniques lead to unbounded error growth due to platform's imperfections (such as unequal wheels' diameter) or to terrain slippage and non-flat terrain. In opposite, inertial measurement units (IMUs) are expensive relative localization devices composed by three accelerometers and three gyros. Two major drawbacks make IMU systems unreliable. First, is that they are sensible to platform's inclination, i.e. to undulated terrain and second, is due to the accelerometers' measures which are characterized by a very low signal to noise ratio (SNR) when the low accelerations are encountered. Thereofore, IMU systems are reliable when navigation is performed at high speed and acceleration values. However, this cannot be ensured in uneven terrain, such as in underground or underwater prehistoric caves. Finally, the error of the position estimate causes drift over time.

The SLAM technique represent a fundamental ingredient to provide unmanned mobile platforms with autonomous navigation capabilities for accomplishing complex missions in GPS-denied areas or where inertial measurement systems are inaccurate and dead-reckoning techniques exploiting noisy estimates lead to map building and location estimates

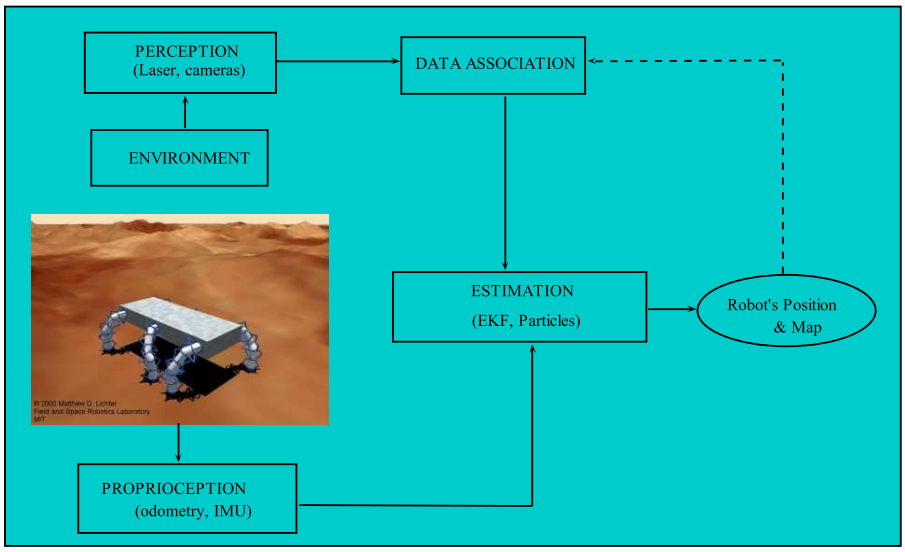


Figure 2.6: The SLAM process [Lemaire, 2004], [Lemaire et al., 2007].

(i.e. position and orientation) characterized by an unbounded error growth. Although, dead-reckoning methods are suitable for local path planning and control execution, they do not allow for global positioning. Figure 2.6 illustrates a general architecture of the SLAM process which consists in estimating jointly the system's location and building a map of the environment. We recall hereafter the feature-based SLAM functionalities introduced in [Lemaire, 2004] and [Lemaire et al., 2007].

- **Perception** about the system's surroundings is a function strongly related to the environment type and to the sensors embedded by the platform. Generally, the perception bloc is concerned with the salient and observable features' selection through the use of cameras or range sensors.

- The **proprioception** bloc estimates the landmarks' location wrt the platform's position within an *observation* process. Then it follows a *prediction* step which integrates propriceptive measures to estimate the platform's location between two observations via dead-reckoning techniques.

- The **data association** process consists in the association of landmarks belonging to different 3D poses of the system in order to make possible the platform's localization. To do so, the landmarks observed from different positions must be accurately matched in order to avoid pose's inconsistency.

- **Estimation** of both, platform's and landmarks' position (i.e. the state vector) wrt a global reference frame, by integrating perception and proprioception data up to the current time. This step represents the core of the SLAM solution for which probabilistic solutions were developed in order to estimate incrementally a posterior probability distribution over the state vector.

In addition, *map management* issues are to be taken into account within the aforementioned functionalities. In particular, means for selecting pertinent features among the

detected ones in order to avoid expensive computation time and high combinatorial of the estimation process.

Solutions for representing the observations and the motion models are generally performed by computing prior and posterior distribution using probabilistic algorithms. In Appendix A.2 we review the existing probabilistic SLAM solutions. Research directions are focusing on both, the *estimation* and the *data association* processes and reported solutions lead to remarkable results for man-made environments. Probabilistic frameworks lead to a complexity which grows quadratically with the number of landmarks, making them unpractical for large-scale environments. The following subsections describes briefly the technical background of the SLAM problem and summarizes the existing visual-SLAM frameworks exploiting cameras and 3D laser scanners.

2.4.1.1 Visual SLAM

When appearance signatures were firstly employed for indexing databases [Rubner et al., 1998], place recognition in topological maps [Argamon-Engelson, 1998], [Ulrich and Nourbakhsh, 2000] and loop detection [Gutmann and Konolige, 1999], there was when computer vision algorithms started to be intensively employed by the Robotics research community.

Vision allows to perform 3D SLAM, implying the estimation of a fully 6DOF robot pose. Visual motion estimation methods produce very accurate platform motion estimates in presence of stable trackable features, outperforming dead-reackoning techniques. Vision algorithms provide solutions for the two over the four problems of the basic feature-SLAM functionalities: *perception* can be provided by the observed interested points, while *data association* can be solved via interest point matching algorithms.

A second component of the SLAM framework is the *loop-closing* stage, which is strongly conditioned by the capacity to recognize revisited areas in order to build consistent maps.

Let us now focus to the most fundamental key topic of the SLAM solution, the data association problem which depends essentially on the capability to match features reliably under any transformation occurring between two partially overlapped images.

Feature matching. The feature matching procedure identifies two features perceived in two different positions at different moments as being the same physical object in the world. SLAM methods rely on robot's pose and landmarks' location estimates to perform data association. Since the errors on these estimates are large for long loop trajectory, the data association becomes ambiguous. This leads to a number of hypothesis which grows exponentially, making SLAM unpracticable for large areas. Computer vision techniques can solve reliably the data association problem without relying on landmarks' estimates, by employing instead feature extraction and matching algorithms. It is mainly due to this reason that SLAM solutions are recently being directed toward purely visual frameworks.

When conceiving a feature matching algorithm, three key aspects must be taken care of [Brown, 1992]: (a) define a feature space (i.e. either the 2D image space or the 3D world's space, (b) define a similarity measure over the feature space and (c) establish a searching strategy for feature matching.

Features can be directly the bi-directional image signal, edges, contours, lines, regions detected within the image after applying a segmentation step, up to a higher semantic information on the scene content.

Since fragile segmentation algorithms can lead to an alteration of the real region, several research works solve the feature matching problem by associating the whole image content to the feature space. Approaches widely employed are minimizing implicitly or explicitly

a similarity measure between the local gray values [Shi and Tomasi, 1994b], [Martin and Crowley, 1995], [Zabih and Woodfill, 1994]. There is a big BUT aspect when using these techniques since they assume that the inter-image motion is relatively low in order to limit the searching space (i.e. they rely on a known epipolar constraint [Hartley and Zisserman, 2004] or suppose that the transformation laying between the overlapped images is close to identity). Consequently, these approaches are suitable for matching data between consecutive frames.

The SLAM framework is feeded with observations by means of feature recognition, tracking and 3D reconstruction. In contrast to early attempts employing low-level features (vertical edges, lines, segments, etc.) and artificial beacons, recent advances in computer vision lead to high-level feature extraction techniques [Schmid et al., 1998]. Current research works search for image descriptors encoding powerful discriminants invariant to any image transformation. The feature tracking problem consists in estimating features' locations over an image sequence. A traditional tracking framework is the Kanade-Lucas-Tomasi (KLT) [Lucas and Kanade, 1981], [Tomasi and Kanade, 1991], [Shi and Tomasi, 1994a] feature tracker, which employs the Harris corner detector [Harris and Stephens, 1988] for feature extraction. A recently image matching framework employs Scale Invariant Features Transform (SIFT) [Lowe, 2004] in conjunction with the Random Sample Consensus (RANSAC) algorithm [Fischler and Bolles, 1981] to reject false matches. Since SIFT are computationally too expensive, attempts aiming to speed up the computation time lead to the Speeded-up Robust Features (SURF) [Murillo et al., 2007]. The 3D reconstruction problem is concerned with the estimation of the 3D structure and the camera pose using a pair of partially overlapped images via epipolar geometry and fundamental matrix use [Hartley and Zisserman, 2004].

Current feature matching algorithms exploit 2D information to solve for the data association problem, becoming ambiguous in large-scale and unstructured environments, causing accumulated errors along the trajectory and leading to inconsistent maps. When performing SLAM scenarios in underground or underwater environments, the feature matching problem becomes more complicated due to the absence of reliably extractable features and significant variations in viewpoint, lightening and distortions. In such environments, a reliable data association algorithm is required in order to uniquely recognize already seen landmarks. Current research work including the one presented in this dissertation are directed toward the jointly use of sensors providing 2D and 3D information to yield more robust features [Sàez et al., 2006], [Petillot et al., 2008], [Wu et al., 2008]. Chapter 7 resumes the state of the art and describes our proposal.

Visual-based SLAM raises additional issues over laser sensors, including high input data rate, the lack of 3D measurement and the difficulty of extracting long term features of the map. These challenges define a successful vision-only SLAM system as one able to build consistent maps on-the-fly and to close loops for drift correction. All the aforementioned capabilities are strongly conditioned by the detection of stable features within the image, in order to match them under significant viewpoint changes, different illumination conditions, noise and dynamic scene content.

Following the vision system that the platform is endowed with, the landmarks' position and the platform's pose are estimated utilizing either stereo-camera pairs [Zhang and Faugeras, 1992], [Mallet et al., 2000], [Olson et al., 2000], monocular using SFM recovery techniques [Heeger and Jepson, 1992], [Vidal et al., 2001], panoramic-cameras or multiple cooperative unmanned vision-based systems. They aim at producing highly accurate motion estimates between successive data acquisitions, providing the possibility to build

global and consistent maps of the environment.

Stereovision solution. We resume hereafter the stereovision SLAM principle described in [Lemaire et al., 2007]. When using stereovision, 3D coordinates of the feature wrt the robot are provided by matching points in the stereoscopic image pair. Stereovision process estimates the state of the observed features transforming the latter into a landmark (i.e. a 3D point) via feature matching algorithms performed on the image couple provided by the stereoscopic bench. The SLAM solution can be then developed via the EKF scheme which assigns to the state of the filter the 3D position parameters of the stereovision system (or the robot) and a set of landmarks coordinates. State prediction and update processes are generally performed via the EKF equations which can be found in [Durrant-White and Bailey, 2006].

In [Jung and Lacroix, 2003] authors reported preliminary results of a stereovision SLAM solution on a blimp testbed. Methods reported by Davison [Davison, 1998], [Davison and Murray, 2002], [Davison and Murray, 1998] employed fixating active stereo, promoting real-time processing (at 5 Hz) being capable of building sparse 3D maps of natural landmarks on the fly while controlling the robot. It was also shown that is possible to provide accurate SLAM using a small set of landmarks carefully chosen and well spread over the image space. This approach was extended to the case of non-planar ramps traversing by combining stereo-vision jointly with an inclinometer. In [Se et al., 2002] authors demonstrated the feasibility of a SIFT-based approach designed for room-size area, yielding sparse 3D maps. A patch-based approach is reported in [Kim and Sukkarieh, 2003]. Authors present an aerial scenario willing to map patches of known size on a horizontal ground plane.

Monocular-SLAM. When the robot is endowed with a single camera, only the bearings of the features are observed. In opposite to stereovision approaches, the bearing-only SLAM does not recover directly the full state of the landmark from a single observation, being considered as a partially observable SLAM problem [Lemaire et al., 2007]. This requires a landmark initialization procedure, which integrates several observations over time.

Several contributions were reported attempting to provide either a *delayed* or *undelayed* initial state estimation. In [Bailey, 2003] a delayed initial-state estimation method is reported. Authors evaluate the Kullback distance between two robot's poses which leads to high-complexity. In [Deans and Herbert, 2000] authors designed a framework which rely on a feature initialization powered by bundle adjustment procedure [Triggs et al., 1999] and a Kalman filter. Despite the high-complexity of the initialization step, the proposed method yields better results. In [Davison, 2003] [Davison et al., 2004] authors introduce a PF approach for representing the initial depth corresponding to each feature. In large-scale environments the proposed method is unfeasible due to the high number of particles required which increases linearly with the initialization range. In [Lemaire et al., 2005] the initial probability density function (PDF) is approximated with a sum of Gaussians which are passed through a discarding process until only a single Gaussian remains, which is injected into the Kalman stochastic map. [Lacroix last papier].

In [Kwok and Dissanayake, 2004] authors introduce the first *un-delayed* feature initialization method. Authors approximate the initial state with a sum of Gaussians which is explicitely injected into the state of the Kalman filter. However, the system's convergence when updating a multi-Gaussian feature is not yet proved. An extension of this algorithm using Gaussian Sum Filter can be found in [Kwok et al., 2005]. In [Sola et al., 2005] a method based on Kalman federate filtering can be found. Research work relating non-linear

optimization approaches and standard Kalman filter bearing-only SLAM can be found in [Konolige, 2005].

A pure SFM approach is introduced in [Mouragnon et al., 2006]. The proposed method alternates between two steps: the 3D reconstruction is performed for images separated by large displacements introduced as *key frames*, while pose computation is performed between consecutive frames to ensure reliable feature matching. The bundle adjustment stage is performed locally in order to overcome the computational complexity of the global optimization stage. Davison provides a recent description of the MonoSLAM algorithm in [Davison et al., 2007]). Authors designed a pure vision SLAM methodology using a single camera featuring real-time but drift-free capabilities, which were inaccessible to SFM approaches. The algorithm has a $O(N^2)$ complexity, where N is the number of features bounded to around 100. The proposed system was validated with an application to real-time 3D localization and mapping for a humanoid robot and live-augmented reality with a hand-held camera. A three-step approach monocular vision is reported in [Royer et al., 2007] for localization and autonomous navigation. Authors combine off-line learning and SFM for mapping to perform on-line localization with real-time performances.

In [Clemente et al., 2007] a monocular SLAM system is presented based on the hierarchical map approach [Estrada et al., 2005], able to build independent local maps in real-time using EKF-SLAM and inverse depth parametrization proposed by [Montiel et al., 2006]. Authors attempt to solve for the data association problem in dynamic and complex environments. The proposed method employs the same approach for salient features detection and matching as in [Davison and Murray, 2002], whilst the loop closing constraint is applied at the upper level of the Hierarchical Map in near real-time.

Panoramic images provide a fully spherical field of view of the system surroundings from a single 3D spatial position of the system, allowing to track observations over long distances and under significant viewpoint changes. In addition, they allow to map features far away from the camera, which is currently not achievable with stereo methods. Naturally, scientists from robotic research community directed their researches toward the integration of panoramic images within SLAM frameworks. Deans in his Ph. D. thesis [Deans, 2002] demonstrated the feasibility of a 2D SLAM technique using panoramic images. In [Lemaire and Lacroix, 2007] authors have recently reported the first 3D SLAM approach powered by panoramic images. The proposed framework relies on natural features matching and image indexation for loop closing.

SLAM using multiple cooperative systems. Vision-based mapping systems with cooperative robots were recently reported [Danesi et al., 2003], [Sujan and Meggiolaro, 2005], [Brown et al., 2008]. These systems aim at performing collaborative mapping scenarios within which each map generated by each unmanned platform is send to a central processing node to merge all the individuals maps into a global one. Such approaches improve considerably the mapping process in terms of rapidity and computational resources, since each platform can have access to the knowledge gathered by others platforms, collaborating within an information and decision network.

So far we have briefly described several visual-SLAM approaches aiming to provide unmanned mobile platforms with localization and map building capabilities. While solving reliably for the observation and data association problems, the reported vision-based SLAM solutions have several limitations when dealing with unstructured, large-scale and GPS-denied environments.

First, the high-complexity of the state-based formulation of SLAM limits the feasibility to room-size scenarios.

Second, data association problem relies on feature extraction and matching whose existence cannot be guarantee in previously unknown environments. In addition, current frameworks exploit data form navigation sensors (GPS, IMU, odometry) which in underground and uneven terrain may have dropouts or carry noisy measurement. Feature matching algorithms yields reliable results in structured environments. A hard-to-solve problem in such environments is to conceive feature matching algorithms capable to deal with feature-less and GPS-denied areas. When dealing with unstructured environments, due to the lack of interest features, one must design a feature matching algorithm capable to disambiguate the data matching process.

A third problem is that the majority of visual-SLAM approaches solve mainly for the platform's localization problem, generating sparse 2D or 3D maps. As stated in Section 2.2.2, this is mainly due to the fact that in unstructured environments stereovision approaches lead to noisy 3D geometry and consequently they cannot allow for accurate 3D measurements. Therefore, endowing a mobile platform with dense 3D scene representation still remains an open issue in the robotics research community. In order to address this critical need, several research works were directed toward the use of range sensors. The next subsection reviews several range-based SLAM methods.

2.4.1.2 Range-based SLAM

Achieving three-dimensional SLAM is a straightforward extension of the 2-D case. Nevertheless, it involves additional complexity due to the more general vehicle pose, increasing sensing and feature modeling complexity. There are three essential forms of 3D SLAM.

The first class is a simple 2D SLAM framework which includes additional map building capabilities in the third dimension. For instance, a widely used system is composed by a horizontal laser and a second orthogonal laser which maps vertical slices [Mahon and Williams, 2003], [Thrun et al., 2000]. However, these approaches are suitable only for vehicles motions which are confined to a plane.

The second form is a direct extension of 2D SLAM, which relies on landmarks extraction and joint estimation of the map and the vehicle's pose.

The third form involves a different SLAM formulation, where the joint state is composed of a history of past-vehicle poses [Newman et al., 2006], [Eustice et al., 2005]. At each pose, the vehicle gathers a 3-D scan of the environment, and their alignment is performed via scans matching methods. These approaches are referred to as *trajectory-oriented* SLAM methods. They are suitable for environments where discrete identifiable landmarks and direct alignments of sensed data is more reliable. When using these techniques, the map is no longer part of the state to be estimated, forming instead an auxiliary data set. Each pose has an associated scan of sensed data and by aligning them it is possible to form a global map.

Also, *topological mapping* gave rise to different trajectory-based SLAM paradigm, where poses are connected in a graphical network rather than a joint state vector. This framework is known as *consistent pose estimation* (CPE) methods [Gutmann and Konolige, 1999], [Konolige, 2004], which based on topological mapping and data alignment procedures yields exemplary results in large indoor environments. Other methods are extracting directly 3D features which are further inserted into a map or matched against others overlapping scans to accurately measure robot displacement and build a map of historic robot locations each with a local scan reference [Thrun et al., 2000], [Konolige and Gutmann, 1999].

Recent trends apply probabilistic frameworks to 3D mapping. In [Katz et al., 2006]

authors employ a probabilistic approach of ICP scan matching. An extended Kalman filter is integrated within the mapping framework in [Weingarten and Siegwart, 2006], [Cole and Newman, 2006]. In [Olson et al., 2006] authors build globally consistent maps by minimizing the global non-linear constraint network on a set of poses. Their approach was employed by [Triebel et al., 2006] to create multi-level surface maps.

Several research groups aim at generating highly accurate 3D maps using immobile 3D laser scanners [Allen et al., 2001], [Georgiev and Allen, 2004], [Sequeira et al., 1999]. In [Sequeira et al., 1999] authors introduce the RESOLV project aimed at modeling interiors promoting virtual reality and tele-presence applications. Their system is composed by a robot endowing a RIEGL® scan laser and scan matching algorithms through means of ICP. In [Allen et al., 2001] authors have developed a robot for modeling urban environments using a CYRAX® scanner and a feature-based algorithm for 3D scans alignment. However, in their recent work [Georgiev and Allen, 2004], authors do not exploit the 3D laser data for the platform's localization.

In opposite to visual-based SLAM methods mainly designed for localization purposes, range-based methods provide highly-accurate dense 3D maps, beside the localization. However, current visual- and range-based SLAM frameworks present two common drawbacks: first, they exploit initial estimation provided by navigation sensors which are unreliable on non-flat terrain and underground environments. Second, they utilize prior knowledge on the scene's content, relying on features' existence for data matching task, which cannot be guarantee in previously unknown environments.

Nevertheless, visual and range-based SLAM solutions have complementary advantages. On one hand, visual-SLAM techniques provide appearance information allowing to solve for data association, loop closing and localization problems, being unsuitable to produce dense 3D maps for areas where interest points are not detectable (such as unstructured zones). On the other hand, range-based SLAM solutions ensure accurate and dense 3D geometry recovery suitable for generating in-situ rich environment perception and for inferring semantics about the robot's surroundings. In addition, both radiometric and geometric information can be efficiently combined to disambiguate the data matching task in feature-less and GPS-denied areas for field robotics applications.

2.4.1.3 Fusion-based SLAM Solutions

In 2009, the Workshop on Visual Navigation and Mapping held in conjunction with the IEEE International Conference on Robotics and Automation has clearly showed that SLAM techniques have now reached a considerable state of maturity.

The standard state-space approach to SLAM in now well understood and representation, association and computation issues appear to be theoretically solved. Nevertheless, the computational complexity limits such methods to room-size areas. As for the visual-based SLAM solutions, since they rely on feature extraction and matching, tractable solutions can be obtained for scenes containing strong structures. Consequently, the accuracy of the 3D structure and platform's motion estimates are subject to the scene's content. Such approaches yield sparse maps, being suitable for accurate localization purposes rather than mapping. In order to obtain accurate and dense 3D point clouds from passive 3D vision techniques the environment has to be strongly structured in order to allow for stable detectable features. In unstructured environments, it is therefore difficult to achieve both: accurate and dense 3D mapping, and 6DOF localization using purely visual SLAM solutions. In opposite to visual-SLAM methods, range-based approaches yield the pos-

sibility to provide dense and accurate 3D maps and 6DOF localization in unstructured environments. A common drawback of the reported visual and range-based SLAM frameworks is that they integrate noisy measurements provided by navigation sensors (GPS, IMU, odometry) which are not reliable and even absent in unstructured and underground environments.

Much less research works attempt to solve for the SLAM problem in complex and difficult to access environments for field robotics applications. In such environments, special attention must be given to the absence of reliably detectable and trackable features but also to the impossibility to rely on navigation sensors. Appearance- and range-based SLAM methods are opening a radically new paradigm for mapping and location estimation, without the need of strong geometric or radiometric landmark descriptions. As stated in [Bailey and Durrant-White, 2006], these methods are opening up new directions and making links back to fundamental principles in robot perception. The key challenges for SLAM are in larger, more persuasive implementations and feasibility demonstrations of the autonomous system. Ongoing research works on SLAM aim to demonstrate their reliability within complex missions taking place in hostile environments, such as driving hundreds of kilometers in large-scale and increasingly unstructured environments where GPS-like solutions are unavailable such as, forest canopy [Nister et al., 2004], underground, underwater mapping and ship hull inspection [Kim and Eustice, 2009], under-ice exploration [Kunz et al., 2009] and automatic mapping of Martian physiography [Stepinski and Bagaria, 2009].

Recent trends of SLAM approaches are directed toward the jointly use of the existing solutions. In [Kim et al., 2009] an integration of grid and topology map is reported. Recently, the use of a sonar-based SLAM solution in conjunction with neural networks for object classification was reported in [Conte et al., 2008]. In [Nuchter et al., 2005] authors employ low-level semantic knowledge (i.e. segmentation of the surrounding areas into ceiling, floor and in-between) to perform 3D mapping using the Kurt3D unmanned mobile platform.

The first hybrid SLAM framework was reported in [Newman et al., 2006]. Authors use an actuated laser scanner for geometric map building and vision for loop closing. Each image is passed into a feature extraction and matching procedure to estimate the rigid transformation between two robot's poses. The latter becomes the initial solution for an iterative laser scan registration procedure reported in [Cole and Newman, 2006]. Authors reported the use of SIFT [Lowe, 2004] and Harris Affine Detector [Mikolajczyk and Schmid, 2004], but other descriptors robust to wide baseline are likely to be employed. The proposed feature matching procedure yields robustness to repetitive texture in outdoor urban scenes, such as trees and windows.

2.5 Proposed Image-Laser Solutions for in-situ 3D Modeling

So far we provide a description of the existing methods for generating digital 3D models of the reality through the use of cameras and laser range finders, i.e. digital scene representation, 3D modeling and SLAM techniques, showing that recent trends are now directed toward the fusion of the different techniques and image-laser fusion is one of them.

In Appendix A.3 we review several 3D modeling systems grouping the aforementioned techniques and highlight their shortcomings with respect to our research goal: *in-situ 3D modeling in unstructured and difficult-to-access environments*. Currently existing approaches were conceived under the assumption that the system disposes of a prior knowl-

edge on the scene's content (i.e. radiometric or geometric features' existence, regular terrain to navigate with minimal perception) or the possibility to rely on navigation sensors (GPS, INS, odometry, etc.).

Our research work aims to study and evaluate the potential of the image-laser data fusion for addressing efficiently the in-situ 3D modeling problem when dealing with difficult to access and unstructured environments. In this context, we suggest that a successful in-situ 3D modeling framework can be supplied by a combination of the aforementioned techniques improved with image-laser solutions for handling feature-less areas in order to respond to worse-case in-situ 3D modeling scenarios.

The image-laser complementarity has been widely emphasized by several research works focusing on photorealist and highly-accurate 3D modeling frameworks. This dissertation exploits the image-laser complementarity, providing a fully automatic 3D modeling system designed to generate in-situ photorealist and complete digital 3D models in feature-less and GPS-denied large scale environments. The proposed system takes advantage of the Nister [Zhao et al., 2005] system's design (i.e. camera-laser baseline negligible wrt the scene's depth), by employing a RACL system, overcoming therefore the shortcomings caused by FMCL systems caused by high inter-sensors parallax. Beside this main contribution, it provides image-laser solutions aiming to fill the gap between nowadays image- and laser-based world modeling techniques to overcome their limitations when dealing with unstructured, large-scale and difficult to access environments. We summarize hereafter the image-laser joint solutions achievable by the proposed system.

2.5.1 Digital scene representation

Figure 2.2 emphasizes the complementarity of the nowadays digital scene representation techniques. In particular, *IBR methods* lack geometry (for generating novel coherent views from original images), which can be recovered via either passive or active 3D vision techniques. However, *passive 3D vision* methods cannot recover dense and accurate 3D geometry in areas for which stable features are barely detectable, such as unstructured environments. In exchange, *active 3D vision* sensors captures densely sampled geometry of the sensed surface, but they lack photorealist photometric information. Consequently, the aforementioned techniques have several shortcomings when dealing with the unstructured and difficult to access environments, which can be overcame through their joint use.

To do so, this dissertation introduces the *4D-mosaics* defined as fully spherical panoramic views unifying color and geometric information. They allow to generate novel views geometrically coherent with the real scene by accurately interpolating between 4D-mosaics, by-passing therefore the costly acquisition step required by IBR approaches. Passive and active 3D vision techniques complementarity becomes more obvious when dealing with unstructured environments. Active 3D vision techniques recover more reliably the 3D geometry, while passive methods provide texture information. The proposed 4D-mosaicing sensor builds a bridge between active and passive methods, providing photorealist and geometrically accurate panoramic view for a single 3D pose of the system by stitching automatically several overlapping scans and color images.

2.5.2 Automation of the 3D modeling pipeline

Data acquisition. In this dissertation we propose means for automatic data acquisition and visual servoing procedures for intelligent digitization. To this end, Chapter 3 introduces a mosaic-driven acquisition scenario, while visual servoing procedures for ensuring the

3D scene model completeness are proposed in the Chapter 7 of this dissertation. Since panoramic views provide long-term feature matching, the proposed mosaic acquisition scenario facilitates data matching task in feature-less areas and ensures in-situ the 3D scene model completeness.

Data Alignment. When dealing with unstructured and difficult to access environments, the main issue standing behind the automation of the 3D modeling pipeline is represented by the *data alignment* stage. In such environments, automatic data alignment is still an open issue since the available data alignment algorithms rely on the assumption that the scene contains stable trackable features whose existence cannot be guaranteed. In addition, most algorithms exploit either appearance or geometric criteria separately which limits the robustness to false matches.

In this dissertation we address key issues for solving the data alignment problem in feature-less and GPS-denied areas. We introduce fully automatic data alignment techniques, giving rise to a 3D modeling pipeline capable to run in-situ without human operator's intervention. The proposed techniques supply fully automatic functionalities for multi-scans (Chapter 4), multi-image (Chapter 5) and laser-image alignment (Chapter 6), without relying on features' extraction nor on navigation sensors, yielding therefore an environment-independent solution for the data alignment task. In addition, data matching under wide view-point variation is performed using a joint 2D-3D criteria to eliminate false matches.

Beside the in-situ generation of 3D scene models, the proposed automatic data alignment procedures provides solutions for multiple onboard autonomous functionalities in feature-less and GPS-denied areas, such as: localization, 2D or 3D mapping, SLAM, view and path planning, obstacle detection, place recognition, etc.

Rendering. In our research work we aim at performing a fast in-situ 3D scene model rendering to validate the completeness of the global 3D scene model. Chapter 6 of this dissertation describes the 3D model rendering process integrated within the proposed 3D modeling pipeline. The proposed method exploits essentially the geometry and the texture recovered by a RACL system.

In order to cope with time and in-situ access constraints, we propose a multi-level 3D model rendering method. Depending on time constraints, the available computational resources and the desired accuracy, the 3D model rendering process can be performed at different resolution levels. For each level, we provide two methods for the 3D model rendering step. The first method assigns the color the each 3D point to yield a photorealist digital 3D model. The second method performs texture mapping onto 2D meshes generated from 3D point clouds to produce a photorealist 3D model rendering. Data integration into 2D meshes is performed via an automatic algorithm able to handle both type of inputs: unorganized 3D point clouds or a set of range images. In order to save computation time and power resources, additional processing to include lighting and shading effects can be performed off-line by a host wirelessly connected to the target to produce highly accurate and photorealist 3D model for virtual tourism applications.

2.5.3 4D-Mosaic-driven Dual SLAM Solution for Complex Environments

Since the in-situ 3D modeling problem is strongly related to the systems' capacities to acquire, generate and merge partially overlapped 3D models and to localize it-self within the global 3D scene model, in this dissertation we tackle the SLAM problem for complex and difficult to access environments and propose image-laser means to solve for it without

relying on feature detection nor on navigation sensors (GPS, IMU, odometry).

We propose a dual SLAM solution through the jointly use of a color camera and a 3D laser range finder. The term *dual* denotes the complementary image-laser fusion, which combines sparse 3D point clouds and ultra-high resolution color images.

The proposed system embeds only active and passive 3D vision sensors, being designed to generate in-situ photorealist and complete digital 3D models in large-scale and unstructured underground environments. Chapter 3 introduces the hardware and the software architecture of the proposed system. The platform is aimed at generating in-situ complete and photorealist 3D models using a *4D-mosaic-driven* SLAM scenario, without relying on navigation sensors nor on feature extraction and matching.

For each spatial position of the platform, the system generates in-situ a fully spherical *4D mosaic* (i.e. 4-channel R,G,B and depth), encoding a photorealist and dense 3D digital panoramic view. The 3D scene model completeness is ensured via a 4D-mosaic-driven acquisition scenario by employing visual servoing resources to provide feedback control for view planning, path planning and autonomous navigation to supply complete site exploration and digitalization. The latter calls for SLAM, obstacle detection and fast decision making capabilities. A SLAM doze would be obtained by matching and merging several 4D-mosaics wrt a global coordinate system, resulting in a global 3D scene model which powers the visual control loop.

A *4D-mosaic* is performed within three-steps of data alignment which are detailed throughout this dissertation. First, several partially overlapped 3D scans are automatically aligned and merged into a 3D fully spherical 3D mosaic. This method is described in Chapter 4. Second, a PTU delivers a sequence of high-resolution color images which are stitched into a 2D Gigapixel mosaic via a multi-view image alignment technique. Technical details on this method are provided in Chapter 5. Third, an image-laser data alignment algorithm is proposed in order to align the 3D mosaic onto the 3D Gigapixel one. This method is described in Chapter 6. All the aforementioned are featuring data matching capabilities for feature-less and GPS-denied areas. As stated in [Lemaire and Lacroix, 2007], panoramic images provide long-term feature matching. In addition, the main advantage of the 4D mosaic-driven SLAM scenario is that it provides reliable data association in unstructured environments by exploiting both 2D and 3D information to disambiguate feature matching when stable features and not detectable. The use of 4D-mosaic matching provides tractable solution for *data association* and *loop closing*, avoiding the use of image indexation for loop closing suggested in [Lemaire and Lacroix, 2007], which is a reliable, but time consuming solution.

Hereafter, we summarize several capabilities provided by the proposed *4D-mosaic-driven hybrid SLAM* scheme:

- 4D-mosaic views allow for photorealist and dense 3D maps within a hierarchical representation;
- reliable solution for the data association and loop closing problems through the 4D-mosaic matching;
- environment perception by merging appearance and geometry information to infer semantics over the scene's content.

When dealing with the SLAM problem in unstructured and difficult to access environments, it is strongly required to generate high-detailed photorealist and dense 3D maps

of the environment in order to endow the platform with rich perception of the environment. Therefore, one must be able to solve reliably for the data matching (image-image, range-range and range to image alignment) in order to provide mapping, localization, data association and loop closing capabilities, without relying on navigation sensors nor on feature extraction.

Beside providing automatic means for data alignment allowing to perform vision- and range-based omnidirectional 3D vision in unstructured environments, the proposed 4D-mosaicing sensor ensures reliable data matching between different 3D poses of the system using a joint 2D-3D similarity criterions, allowing to disambiguate the data matching process, which is sensible to false matches when using only radiometric criterions. Figure 2.7 synthesizes the existing SLAM solutions and places our proposal within it: *dual SLAM* described throughout this section.

All the aforementioned functionalities allow for several applications to be performed either in-situ (site surveys, inspection and monitoring in high-risk environments via metrology techniques) or through the world wide web (cultural heritage, data annotation for virtual tourism or augmented reality purposes). In addition, the proposed solutions provides unmanned mobile platforms with rich environment perception (i.e. photorealist and dense 3D mapping), allowing them to be aware when evolving in an previously unknown environment.

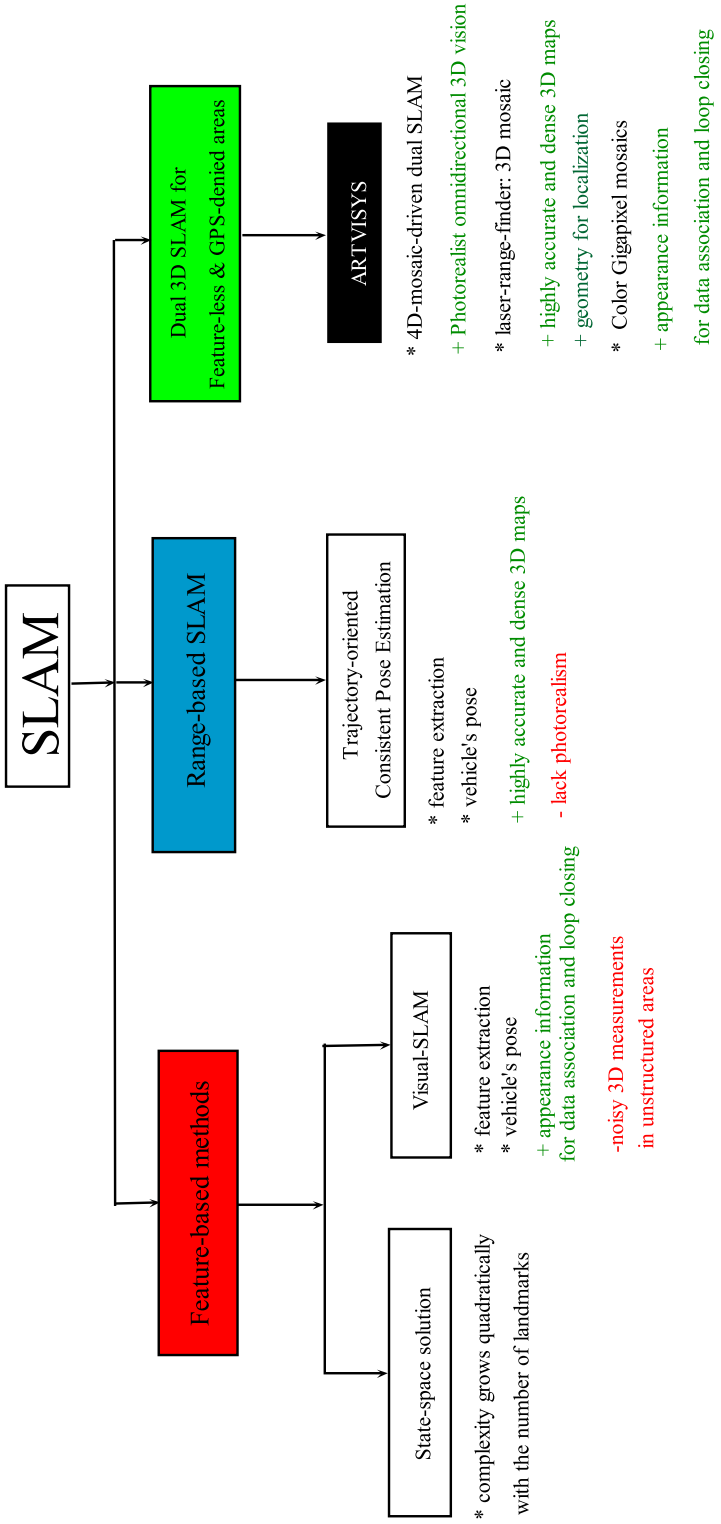


Figure 2.7: SLAM methods classification and our proposed image-laser solution: 4D-mosaic-driven dual SLAM for feature-less and GPS-denied areas.

Chapter 3

ARTVISYS: ARTificial VIsion-based SYStem

This chapter introduces a general-purpose artificial vision-based system aiming to provide vision-based systems (VBSs) with basic functionalities for accomplishing autonomously complex tasks in unstructured and difficult to access environments. An autonomous system can be employed in a wide variety of complex missions and generating in-situ complete 3D models of complex environments is one of them. Since such endeavors are strongly related to a basic sense: *vision*, we endow a site survey system with automatic means for environment sensing, giving rise to a *perception-through-site survey* system.

We start this chapter by listing several key issues which need to be addressed when dealing with the 3D modeling problem in challenging environments. The next section solves for the automation of the 3D modeling pipeline by introducing the *4D-mosaic* - a fundamental ingredient of our proposed in-situ 3D modeling pipeline. In order to ensure in-situ the 3D scene model completeness, Section 3.3 proposes a *4D-mosaic-driven in-situ 3D modeling process* having as main scope the automatic digitization and exploration of the site. Section 3.4 presents the ARTVISYS's software architecture, whilst Section 3.5 describes the in-situ operating modes of the ARTVISYS system. Section 3.6 summarizes ARTVISYS's capabilities comparing to the state-of-the-art 3D modeling systems, while Section 3.7 concludes on the ARTVISYS's potential to be used as a general-purpose vision-based system which can be upgraded with secondary capabilities to perform specific tasks in complex environments.

3.1 Key Issues for In-situ 3D Modeling in Challenging Environments

Nowadays, unmanned systems equipped with active and passive 3D vision sensors are about to be deployed in *previously unknown* and *difficult-to-access* environments for performing complex tasks in order to avoid operator's intervention. During such endeavors it is often not possible to provide the platform with a detailed map of the environment. Therefore, the system is required to create a dimensionally accurate geometric model of its surroundings by sensing autonomously the environment.

In our research work, we aim at developing a vision-based system capable to generate in-situ photorealist and digital 3D maps in unstructured and difficult-to-access environments in order to avoid operator's intervention. To achieve the aforementioned research goal, this

dissertation is willing to answer the following question:

How can an unmanned vision-based system generate in-situ photorealist and geometrically accurate and high resolution digital models of the real world in previously unknown, unstructured and difficult-to-access environments?

When referring to the state-of-the-art 3D digitization techniques described in Chapter 2, additional key aspects must be taken care of in order to endow an unmanned platform with autonomous functionalities for supplying in-situ 3D modeling missions in complex environments. We list hereafter these key aspects and how we propose to address them in this dissertation.

Underground, underwater, under-ice, extra-terrestrial environments. Such environments lead to the impossibility to rely on navigation sensors (GPS, INS, odometry, dead-reckoning). Several research works have attacked this problem but the proposed frameworks rely on features' extraction and matching (Harris, SIFT) yielding reliable results for structured scenes, or on shape descriptors which require dense 3D scans, being unaffordable for in-situ processing.

In this dissertation we provide fast acquisition and data processing by acquiring low resolution 3D scans and high-resolution color images. It is important to note that in our research work no additional sensors are used beside a 3D laser scanner and a camera. Both capturing devices are sensing the environment providing unmanned systems embodied with such sensors with onboard visual capability. The proposed algorithms exploit the jointly use of image and laser data alone to design environment-independent methods for map building and platform's positioning, without exploiting knowledge from navigation sensors.

Unstructured terrains. In such environments geometric and radiometric features are barely extractable. In addition, zones containing repetitive texture or too homogeneous areas are difficult to be reliably matched using the currently existent feature extraction and matching methods, due to the separate use of either 2D or 3D information. In this dissertation we propose means to disambiguate data matching in unstructured environments by employing a joint 2D-3D criterion, yielding robustness to feature matching. To this end, in Chapter 6 we introduce the *4D-mosaic* data structure in order to ensure reliable data matching under significant viewpoint variations, which encode photogrammetric and geometric information, allowing for long-term primitive matching. The research work related in this dissertation introduces two categories of descriptors to solve for data matching problem in unstructured environments.

- The first one corresponds to 2D features encoding radiometry and the bearing of the sensor. These features were developed for image stitching purposes and a detailed description of them is provided in Chapter 5. Since they do not correspond to any interest point, we introduce them as *anonymous features* (AF).

- The second class of features is a hybrid 2D-3D descriptor encoding texture and shape properties (i.e. geometric and statistical distribution of the 3D point clouds) of the sensed surface. They are introduced in Chapter 7 as *viewpoint invariant hybrid descriptors* (VIHD) for matching 4D-mosaics acquired from different 3D poses of the system, under significant viewpoints variation and illumination changing.

Dangerous areas. From Appendix A.3 we can note that several research works have reported semi-automatic frameworks for 3D modeling of complex environments [Banno et al., 2008], relying on heavy operator's intervention for placing artificial landmarks, for

guiding the data acquisition and processing, and for validating the final 3D scene model. In our research work, the 3D modeling missions are undertaken in hostile environments, precluding therefore human surveyors' access. The critical need for a vision-based automatic 3D modeling system is emphasized by the operator's difficulty to access too small or too dangerous areas. As in the work proposed by [Nuchter et al., 2005], [Paar et al., 2009], [Miura and Ikeda, 2009], a possible solution is to send data to a host wirelessly connected to the target in order to allow data processing and validation. However, such methods are usually limited by memory bandwidth and communication latency, being therefore vital to develop fully onboard autonomous functionalities to be performed in-situ. To this end, this dissertation proposes automatic algorithms to automate the 3D modeling pipeline, which leads to in-situ generation of *4D-mosaics*, unifying photorealism and geometry into a fully spherical view.

Limited time and in-situ access. A very ambitious goal in mobile robotics is to deploy unmanned systems in hostile environments to perform complex missions. The goal of such endeavors is to avoid endangering human's surveyor's life and to ensure that the mission is accomplished within the granted time. In order to avoid to come back on site to complete data collection, the limited time and in-situ access within which the 3D scene model completeness must be ensured are major concerns. In this dissertation we address this issue by proposing an automatic 3D modeling pipeline able to cope with time and in-situ access constraints. The 3D modeling capability is further exploited along with visual servoing procedures in order to guide the system to act intelligently on-the-fly to ensure in-situ the 3D scene model completeness.

Dealing with the aspects listed above is the main concern of our research work. To this end, this dissertation provides image-laser solutions for the aforementioned issues and proposes their integration within an unmanned vision-based system architecture. Since the proposed system exploits active 3D vision and color camera inputs alone, this dissertation introduces it as ARTVISYS, the corresponding acronym for ARTtificial VIsion-based SYStem.

Our research work aims at addressing the aforementioned issues by introducing a purely-vision-based system able to perform in-situ the entire 3D modeling pipeline: (1) data acquisition and processing, (2) generate 3D models in a step-by-step fashion, (3) act intelligently on-the-fly in order to improve the scene completeness.

This dissertation focuses mainly on the system's design and solves for the automation of the 3D modeling pipeline through the use of 4D-mosaics. The research perspectives aim at exploiting the 4D-mosaic structures to supply visual servoing functions to provide unmanned platforms with autonomous behavior for site exploration and to ensure complete site digitization.

The proposed system embeds a *perception-through-site-digitization* capability, being able to perform site surveys missions in previously unknown environments, while perceiving the environment. The system's design allows its use in either *autonomous* or *manual* mode (to provide assistance to the human surveyors). Tests run on real data acquired in three prehistoric caves from France (Moulin de Languenay - Chasteaux, Tautavel and Mayenne Science) are presented to evaluate the performances of the proposed system. In addition, in order to illustrate the environment-independent character of the proposed methods, we perform tests in structured environments using data sets acquired by a vehicle equipped with several cameras and LRFs designed for city mapping purposes.

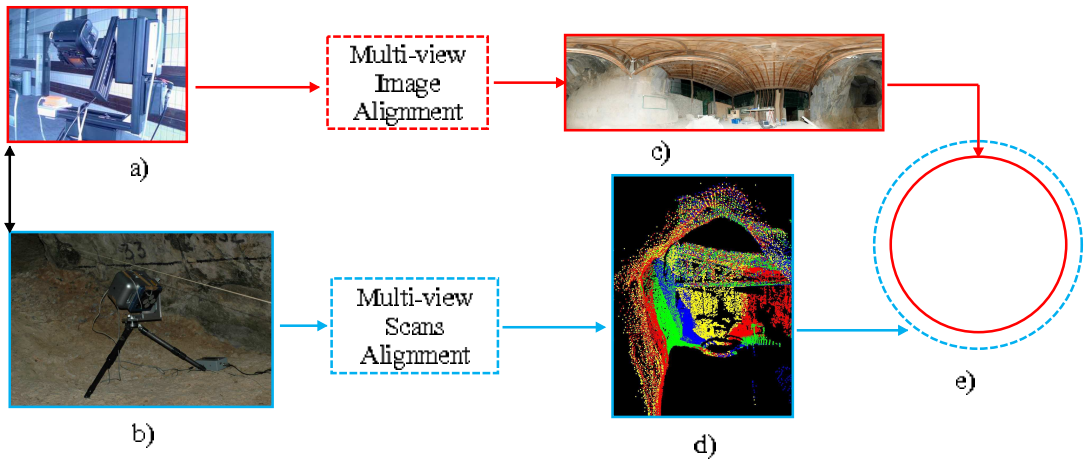


Figure 3.1: The 4D-mosaicing process proposed for integration onboard ARTVISYS. a) NIKON D70 ® digital camera mounted on Rodeon ® motorized pan-tilt unit, b) Trimble ® 3D laser-range-finder during a data acquisition campaign undertaken in the Tautavel prehistoric cave (France) by the French Mapping Agency in October 2007, c) a Gigapixel color mosaic resulted from an image sequence acquired in the Tautavel prehistoric cave using an automatic image stitching algorithm which we introduce in Chapter 5 of this dissertation, d) a 3D mosaic resulted from several overlapped scans acquired in the Tautavel prehistoric cave, matched by an automatic multi-view scan-matcher proposed in Chapter 4, e) alignment the 3D mosaic onto the Gigapixel one to produce the 4D mosaic, process described in Chapter 6 of this dissertation.

3.2 Automatic 3D Modeling through 4D-Mosaic Views

This section resumes how we solve for the automation of the 3D modeling pipeline through the use of 4D mosaics. The following description introduces the hardware design and summarizes the 4D-mosaicing process.

We designed a *dual* system for performing in-situ 3D modeling tasks in large-scale, complex and difficult to access underground environments. Since in such environments navigation sensors are not reliable, the proposed system embeds only 2D and 3D vision sensors, unifying photorealism and high resolution geometry into *4D mosaic views*. Figure 3.1 illustrates the ARTVISYS's hardware along with a *4D-mosaicing* process. We describe hereafter several ARTVISYS's features and justify the proposed design.

RACL dual-system. The proposed hardware architecture falls in the category of the RACL dual sensing devices, embedding a high-resolution color camera mounted on a motorized pan-tilt unit and a 3D laser-range-finder, which are depicted in Figures 3.1 a) and b), respectively. There are several reasons for choosing a RACL design:

- image-laser complementarity has been widely emphasized and investigated by several research works. There is no doubt that employing the two sensors separately, none can solve for the 3D modeling problem reliably.

- RACL systems overcomes several shortcomings raised by FMCL ones. In particular, image-laser alignment and texture mapping procedures are difficult due to occluded areas in either image or laser data.

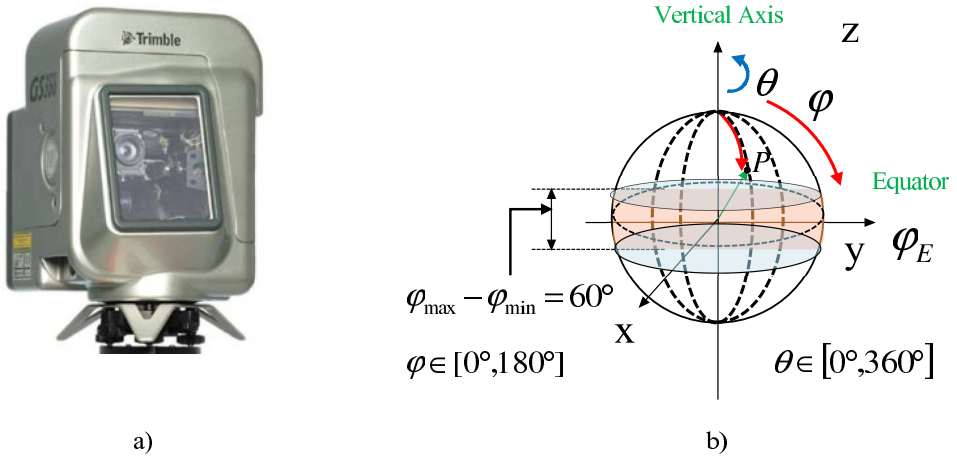


Figure 3.2: a) Trimble ® laser range finder delivering 5000 points per second with an accuracy of 3mm at 100m. The dimensions of the laser range finders are: 340mm diameter, 270mm width and 420mm height. The weight of the capturing device is 13.6kg. b) the field of view covered by the sensor.

3D Mosaics from laser-range-finders. As mentioned in Section 2.2.3, in opposite to stereovision techniques, laser range finders allow dimensionally accurate and high resolution geometry recovery in unstructured environments, where stable features are not detectable. The scanning device employed in this research work belongs to the time-of-flight 3D scanning techniques for which a brief description is provided in Appendix A.1. We dispose of a Trimble ® laser range finder depicted in Figure 3.2 a) providing a 3D point clouds and their associated light intensity backscattering within a field of view of 360° horizontally × 60° vertically, as shown in Figure 3.2 b). ARTVISYS acquires several partially overlapped scans and matches them into a fully 3D spherical mosaic via a multi-view scan matching algorithm for which a detailed description is provided in Chapter 4. Figure 3.1 d) illustrates an example of a 3D mosaic obtained from real data acquired in the Tautavel prehistoric cave.

Gigapixel panoramic head. ARTVISYS is capable of generating photo-realist 3D models through the use of fully spherical Gigapixel mosaics. Our system is equipped with a Gigapixel panoramic head for which we designed an automatic software for generating in-situ Gigapixel mosaics. The Gigapixel mosaicing system illustrated in Figure 3.1 a) delivers a sequence of ultra-high resolution and pose-annotated images which are further automatically stitched into a Gigapixel mosaic via a multi-view image matching algorithm which will be described further in Chapter 5. Figure 3.1 c) depicts an example of an Gigapixel mosaic obtained from real data acquired in the Tautavel prehistoric cave.

4D-Mosaicing. For each 3D spatial position, ARTVISYS aligns the 3D mosaic onto the Gigapixel color one and integrates them into a photorealistically textured 3D panoramic view encoded as a *4D-mosaic*. This is the last step in the 4D-mosaicing process corresponding to Figure 3.1 e) and for which a detailed description is provided in Chapter 6. The main advantage of 4D-mosaic views is represented by the fact that they encode explicit color information as 3-channel components (i.e. red, green and blue) and implicit shape description as depth for a fully spherical view of the system's surroundings. The four dimensional components are required in order to ensure reliably further processing, such as unambiguous data matching under wide viewpoint variation.

Addressing time and in-situ access constraints. An in-situ 3D modeling system must be able to supply fast data acquisition and processing while assuring the 3D scene model completeness in order to avoid to return on site to collect new data. To do so, ARTVISYS fulfills the aforementioned requirements by integrating several features within the data acquisition and processing stages.

- **Low-resolution geometry.** As already emphasized in Chapter 2, dense 3D point clouds require computationally expensive data acquisition and processing. In order to cope with time and in-situ constraints, ARTVISYS acquires low-resolution 3D geometry, whilst fast processing is ensured via a multi-resolution framework.

- **High resolution texturing.** Since cameras allows for instantaneous acquisition, ARTVISYS captures high-detailed texture and generates in-situ Gigapixel mosaics through the use of a fast multi-resolution image stitching algorithm.

- **2D, 3D and 4D mosaics**. ARTVISYS generates in-situ 2D, 3D and 4D fully spherical view of the system for a single 3D pose. This makes them suitable for map-building and localization tasks. In addition, they provide long-term features tracking, ensuring reliable data matching in feature-less environments. The aforementioned advantages are exploited by the ARTVISYS system within a *4D-mosaic-driven acquisition scenario* aiming to ensure the 3D scene model completeness. To this end, a 4D-mosaic matching procedure described in Chapter 7 in order to aligns and integrates them into a global 3D scene model.

- **Partial 3D mosaic acquisition for scene completeness avoiding data redundancy.** The 3D mosaicing scenario starts by acquiring several partially overlapped scans and integrates them into fully spherical 3D mosaics. However, when occlusions are encountered, the system senses only the occluded areas, giving rise to partial mosaics. This allows to ensure the 3D scene model completeness while avoiding data redundancy and long processing.

The next section introduces the 4D-mosaic-driven in-situ 3D modeling process performed by ARTVISYS aiming to ensure in-situ the 3D scene model completeness.

3.3 4D Mosaic-driven In-situ 3D Modeling

The proposed 3D modeling pipeline leads to a vision-based system capable to generate in-situ photorealist and highly accurate 3D models encoded as 4D mosaics for each ARTVISYS's spatial position, called *station*.

When dealing with the in-situ 3D modeling problem in large scale complex environments, one has to generate dynamically 3D scene models and to deal with occluded areas on-the-fly, in order to ensure automatically the 3D scene model completeness. Moreover, since our research work is concerned with 3D modeling in difficult to access environments, time and in-situ access are major concerns. Therefore, once accessing the site, the system must be capable to ensure in-situ the 3D scene model completeness in order to avoid returning on site to collect new data. This calls for an intelligent 3D modeling system, which implies the computation of the Next Best View (NBV) position from which the new 4D mosaic must be acquired in order sense the occluded areas. In addition, the system must be able to navigate from it's current position to the next best estimated 3D pose from which

the next 4D mosaic must be acquired. This implies path planning, autonomous navigation and fast decision making capabilities. To this end, the 4D-mosaicing process in integrated within a *4D-mosaic-driven in-situ 3D modeling process* for which a global description is provided in Figure 3.3.

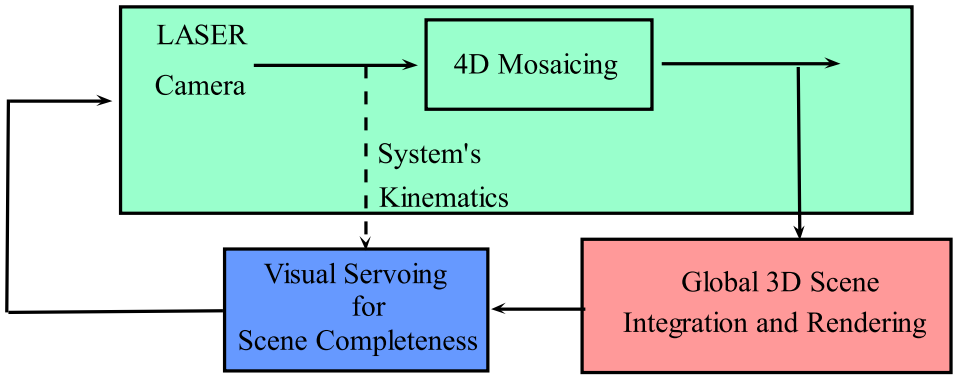


Figure 3.3: The global software architecture of the 4D-mosaic-driven in-situ 3D modeling process proposed for integration onboard ARTVISYS.

Software design. The global architecture of the ARTVISYS is composed by the main process corresponding to the 4D mosaicing acquisition and processing, and the control loop which provides feedback to the system in order to ensure the 3D scene completeness by means of visual servoing. The latter capability exploits the currently generated 3D scene model, being subject to the system's kinematics. The next section provides a detailed description of the onboard autonomous functionalities composing the proposed global architecture.

4D-mosaic-driven acquisition scenario. The above software design is supplied by a *4D-mosaic-driven acquisition scenario* performed in a stop-and-go fashion, as illustrated in Figure 3.4. Due to occlusions, several 4D mosaics must be autonomously acquired from different 3D spatial positions of the system in order to maximize the visible volume, while minimizing data redundancy. The acquisition scenario starts by acquiring a 4D-mosaic which is further exploited to detect the occluded areas. In Figure 3.4, they corresponds to the blue segments representing depth discontinuities associated to each station. In a second step, the system must estimate the 3D pose from which the next 4D-mosaic must be acquired in order to maximize the visible volume. In a third step, the 4D mosaics are matched and integrated within a global 3D scene model which is further exploited to iterate the two aforementioned steps until the 3D scene model completeness is achieved. Figure 3.4 illustrates the matching process between successive stations which establishes homologous triads belonging to the overlapping area. A description of the matching process is provided in Chapter 7.

3.4 On-board Functionalities

This section zooms into the ARTVISYS's software design depicted in Figure 3.3 to provide a detailed description of its onboard autonomous functionalities in Figure 3.5.

Main process: automatic 3D modeling through 4D-mosaics. The main process corresponds to the 3D modeling pipeline for which this dissertation proposes a 4D-mosaicing framework. During the main process, the system acquires a sequence of N

partially overlapped scans and M high-resolution color images, for a given 3D position of the platform. A multi-view scan-matcher exploits the 3D point clouds to automatically generate in-situ a low resolution (LR) 3D mosaic. In a second step, a multi-view image matching algorithm stitches the M high-resolution images into a Gigapixel mosaic. The two aforementioned blocs have stand-alone capabilities, providing the possibility to employ them outside of the 3D modeling process. A third procedure aligns automatically the LR 3D mosaic onto the Gigapixel one to generate in-situ a 4D mosaic view. The result is further exploited by a fourth bloc which improves the 3D model quality by mapping the texture obtained from the Gigapixel mosaic onto 2D meshes computed from 3D mosaic.

Control loop: visual servoing for 3D scene model completeness. The system iterates the main process described above, while servoing the 3D position of the system from which a new station must be acquired in order to fulfill the 3D scene model completeness.

Supposing that several 4D mosaics were acquired, the 4D-mosaic matching bloc computes the 3D pose to unify them into a single 3D entity via the *global 3D scene model integration* process.

In order to complete the 3D scene model, the systems performs automatically *occlusions' detection*. If occluded areas are detected, the systems employs a *view planning* procedure in order to compute the spatial position from where the next station must be acquired in order to minimize the occluded areas, while minimizing data redundancy. Otherwise, the system ends by computing the final 3D scene model rendering. In order to reach the NBV position, we provide image-laser solutions for embedding the system with visual-based autonomous navigation capabilities.

Chapter 7 studies and evaluates image-laser solutions for supplying the aforementioned processes composing the control loop of the ARTVISYS system, leading to a vision-based system embedding complete site digitization and exploration capabilities.

3.5 In-situ Operating Modes

Following the application type, the proposed 4D-mosaic-driven in-situ 3D modeling process can run in either *autonomous* or *manual* mode. Each of them requires automatic *4D-mosaicing matching* and *integration* processes to generate in-situ the global 3D scene model which is further exploited for detecting the occluded areas within the generated scene. The main difference between the two modes is that the first one detects automatically the occluded areas and estimates the NBV from which the new 4D-mosaic must be acquired, while the latter relies on human operator intervention to supply the above mentioned operations.

This dissertation focuses on both operating modes. First, we provide a solution for the *4D-mosaic matching* in Chapter 7, which is common requirement for both operating modes. In a second time, we focus on the *autonomous mode* by introducing an visual-based autonomy model and future research directions to supply visual servoing procedures (i.e. view and path planning, autonomous navigation) powered by the 4D mosaicing module to provide feedback control to the system for ensuring autonomously complete site digitization and exploration.

3.6 System's Capabilities vs. State-of-the-Art

We end this chapter by listing the main capabilities offered by ARTVISYS wrt the state-of-the-art closely related approaches. We compare our system by looking at the global and local levels of the software architecture. At the global level, we compare ARTVISYS's performances wrt the 3D modeling pipeline performances, while at the local level we list its several stand-alone functionalities.

Global level: automatic 3D modeling pipeline through 4D mosaic views

- **4D Mosaics: photorealist omnidirectional 3D models.** In this dissertation ARTVISYS's main purpose is to automate the 3D modeling pipeline in order to generate in-situ photorealist 3D models. Our 3D modeling framework introduces *4D-mosaics* as photorealist panoramic 3D models, which to our knowledge have not been introduced by now.

- **Fast data acquisition and processing.** The 3D laser is set to capture low-resolution 3D point clouds for rapidity purposes, while high-detailed photorealism is enabled by fast acquisition of high resolution color images. We employ a pyramidal data matching framework along with calibration constraints and optimization techniques to ensure fast data processing. The proposed technique overcomes the limitations of the existent techniques which reported dense 3D scans acquisition and time-consuming semi-automatic methods [Levoy et al., 2000], [Ikeuchi et al., 2007] aiming to produce highly accurate 3D models in a controlled laboratory environment, or automatic frameworks employing computationally expensive algorithms to produce highly realist 3D models of the scene [Stamos, 2001].

- **Automatic data alignment in feature-less and GPS-denied areas.** Concerning the fully automation of the 3D modeling pipeline, we would like to first take a look at the data alignment step, because there is where the main issue lies. Our system embeds three procedures for (1) image, (2) laser and (3) image-to-laser data alignment. As presented in Section 2.3.2, the existent data alignment techniques employ either semi-automatic or automatic frameworks and rely on navigation sensors readings or feature existence hypothesis. ARTVISYS provides environment-independent methods for (1,2,3) data alignment steps, yielding robustness to feature-less and GPS-denied areas, overcoming therefore the main drawback of the current techniques.

- **Occlusion-free image-laser fusion.** We proposed a dual RACL system which overcomes the shortcomings of FMCL systems caused by occlusions in either image or laser data. As a matter of fact, we exploit Nistér's idea [Zhao et al., 2005] for the 3D modeling system design, for which the inter-sensors baseline is negligible wrt the scene's depth. When using the proposed hardware design, the rigid transformation separating the two sensors is essentially a 3D rotation and a residual translation, for which is easy to solve since occluded areas minimized.

Local level: stand alone capabilities

- **Scene completeness through the use of fully and partial 3D mosaics.** In [Rekleitis et al., 2009] a Lidar-based path planning algorithm for Mars exploration is proposed. The system acquires directly 3D mosaics from different 3D poses of the system and matches them in a sequential manner. This process provides 3D maps of

the environment which are further exploited for path planning purposes. The main drawback of the proposed system is that for each 3D pose, the system acquires a fully 3D spherical mosaic, leading to a computationally expensive step for the 3D mosaic matching process and global map computation. Our proposed solution consists in acquiring fully 3D mosaics when the system encounters large displacements, and partial 3D mosaics in order to cover the occluded areas, ensuring therefore the 3D scene's model completeness in a feasible computation time onboard mobile platforms.

- **Automatic Gigapixel mosaicing.** There are several key aspects which have to be taken care of when performing high-resolution image stitching. Since feature-based algorithm fails to detect and match poorly textured areas - which is the case in our research work, we develop an automatic Gigapixel mosaicing algorithm powered by a pair-wise image motion estimation algorithm robust to feature-less environments.
- **4D-mosaic hybrid SLAM.** ARTVISYS embeds a 4D-mosaic matching algorithm enabling simultaneous photorealist and dense 3D mapping along with 6 DOF localization through the jointly use of (1) cameras - allowing to solve for data association and loop closing problems, and (2) 3D laser-scanners - suitable for localization purposes. Furthermore, the use of 4D mosaic allows to disambiguate the feature matching process which is inherent to outliers in unstructured terrains. One may argue that due to computational requirements, a hybrid SLAM scheme would be unaffordable for online processing. Computational issues are discussed along this dissertation, providing several optimization schemes and possible improvements.

3.7 Conclusion

This chapter introduces the ARTVISYS system, a vision-based system with an application to automatic environment sensing. This provides unmanned platforms with a basic sense, *vision* - which is undoubtedly a powerful resource for mapping and localization tasks, offering the possibility to infer semantics about the environment in which the platform evolves. The aforementioned functionalities give rise to a *perception-through-site-survey* capability which is basic requirement for endowing unmanned platforms with elementary intelligence (i.e. reasoning and fast decision making capacities) powered by a computer vision engine.

In addition, vision and conceptual perception form the artificial intelligence skeleton to which visual servoing procedures need to be added to supply autonomous onboard functionalities, giving rise to unmanned platforms able to accomplish complex missions in previously unknown environments.

A wide research literature was reported aiming to develop intelligent systems for performing civilian and military applications. However, current approaches are introducing binary reasoning solutions to accomplish specific tasks, without developing perception and reasoning resources in order to solve for the system's autonomy problem. Consequently, such systems fail to deal with unpredictable situations, precluding the feasibility of complex missions in hostile environments, where human operator's intervention is highly undesirable.

Instead of developing such special-purpose systems which are operational within a limited application field, a more durable solution is to solve first for the system's autonomy problem and to enrich them in a second step with specific onboard functionalities. Therefore, in order to provide a long-term solution to this issue, we believe that a general system

dotted with an elementary intelligence level and autonomy has to be designed first. In a second stage, the system can be upgraded with specific onboard functionalities in order to accomplish particular tasks. The aforementioned design enriches unmanned systems with new articulations, from perception to action. For this reason, Chapter 7 investigates the ARTVSYS's potential for addressing unmanned system's autonomy problem and establishes a vision-based autonomy model.

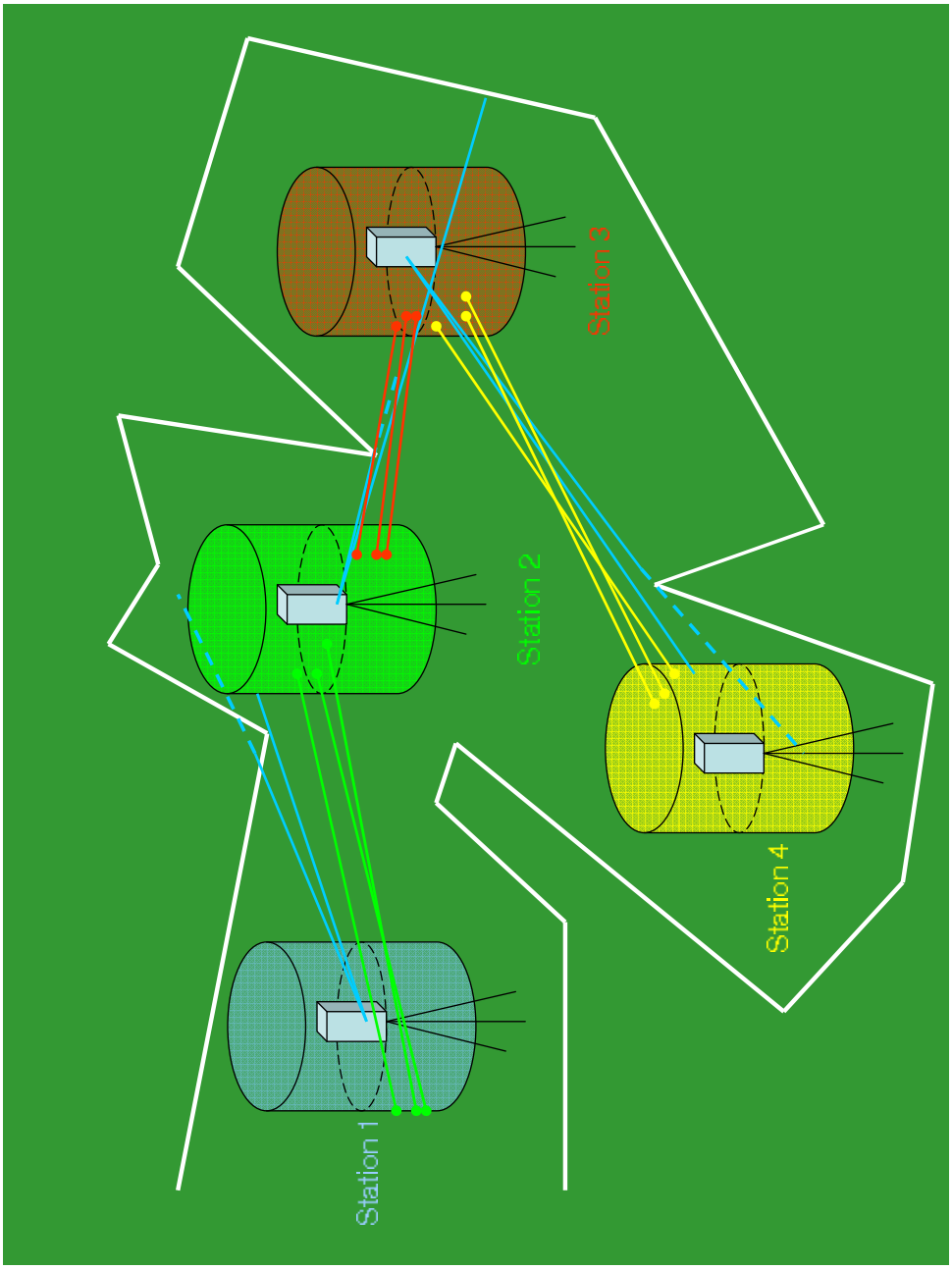


Figure 3.4: The 4D-mosaic-driven acquisition scenario performed by ARTVISYS.

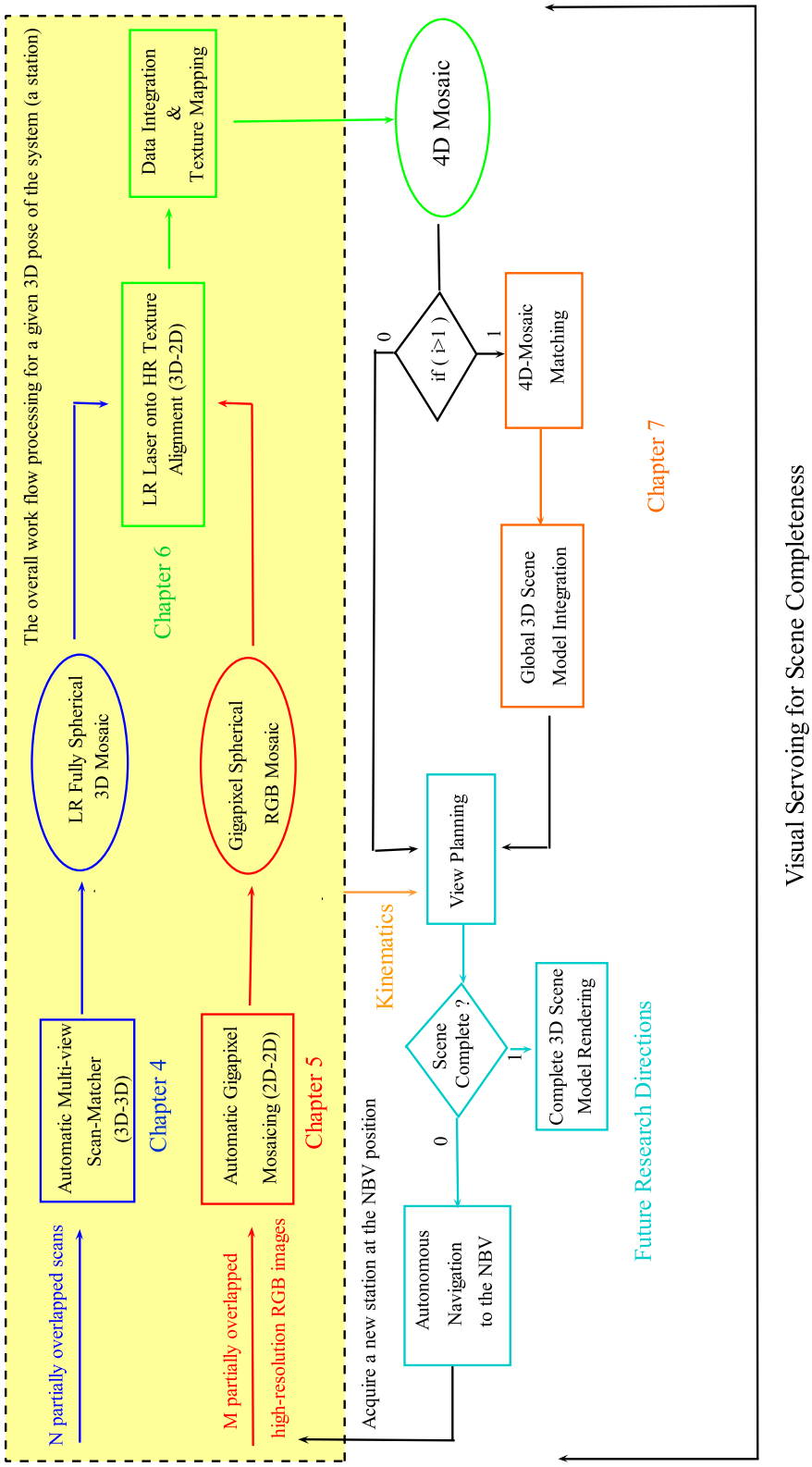


Figure 3.5: A detailed overview of software architecture of the ARTVISYS system.

Chapter 4

Multi-view Scans Alignment for in-situ 3D Mosaicing

This chapter focuses on one of the main processing blocs composing the ARTVISYS system and deals with the automation of the 3D modeling pipeline by introducing an automatic multi-view scans alignment technique for generating in-situ fully 3D spherical mosaics. This chapter presents extends the research work introduced in [Craciun et al., 2008] and [Craciun et al., 2010].

We start this chapter by stating the 3D scan matching problem and by presenting the available methods solving for it. Section 4.3 introduces the 3D mosaicing acquisition scenario, while the following section provides an overview of the proposed multi-view scans alignment technique associated to it. Section 4.5 presents our pair-wise alignment proposal: a *free initial guess* pair-wise scans alignment designed to achieve *precise rigid poses estimates*. The experimental results of the proposed method are presented in Section 4.6. The next section describes how we integrate the proposed pair-wise scan alignment technique within a multi-view global alignment framework to yield *optimal absolute poses*, while Section 4.8 presents the quality assessment of the multi-view scans alignment algorithm. Section 4.9 introduces the embedded design supplying multi-core onboard processing of 3D mosaics. We end this chapter in Section 4.10 by summarizing our research proposal and its main contributions.

4.1 The Multi-view 3D Scans Alignment Problem

3D scanning devices capture the 3D structure of the environment from a single view-point. However, due to the sensor's limited field of view and occlusions, multiple scans from various viewpoints need to be acquired, either automatically or by an operator, in order to sense the scene in its entirety.

Let $S_0, ..., S_{N-1}$ be N partially overlapping scans acquired from different viewpoints. Since each scan is expressed wrt the sensor's coordinate system, the multi-view scan matching problem requires to recover each sensors' viewpoints wrt a global coordinate system in order to align and integrate all views into a single 3D entity. As a matter of fact, the multi-view scans alignment problem is pretty much like a 3D puzzle game, since it consists in assembling several overlapped scans without knowing how the sensed scene looks like.

Generally, the first scan in a sequence can be chosen as the origin, so that the global coordinate system is locked to the coordinate frame of that scan. However, we show

further in this dissertation that this approach can yield optimal results only when a specific acquisition scenario is associated to it.

The multi-view scans alignment process is a fundamental step within the 3D modeling process which requires to solve for the absolute poses. An absolute pose $\mathbf{T}_i, i = \{0, .., N-1\}$ is defined as the 3D linear operators which rigidly transforms the 3D coordinates of a point $\mathbf{p} \in S_i, \mathbf{p} = (p_x, p_y, p_z, 1)^t$ from the local coordinate system of scan S_i to the global (or world) coordinate system: $\mathbf{p}_w = \mathbf{T}_i \mathbf{p}_i$.

Absolute poses' computation is highly dependent on the relative positioning of the overlapped views and their associated overlap. As a matter of fact, when dealing with the multi-view scans' alignment problem, one has to solve for two inter-related sub-problems: (a) find the overlapped views and compute the overlaps $\mathbf{O}_{ij}, j = \{0, .., N-1\}, i \neq j$ when they are detected, and (b) find the relative poses $\mathbf{T}_{ij}$ between the overlapped views. As for the SLAM problem, the relative poses estimates and the overlap are intricately related problems: pose estimation algorithms are generally powered by corresponding point pairs detected in the overlap region of each scan and vice-versa, the precise overlapping areas are not findable without the pose's knowledge.

The aforementioned tasks are generally supplied by pair-wise scans alignment procedures. However, due to the mutual dependency which lies between the overlaps $\mathbf{O}_{ij}$ and the relative poses $\mathbf{T}_{ij}$, the multi-view scan matching is a difficult task. The most general case of the problem does not assume any prior knowledge on the original sensor viewpoints nor on which views share the same scene region, being extremely hard to solve.

Since the multi-view fine alignment problem is powered by the pair-wise alignment process, in the next section we review the existing methods supplying both pair-wise and multi-view scans alignment functionalities.

4.2 Related Work

In the past few years several research studies have been presented within a wide range of communities interested in building 3D models using 3D Laser Range Finders (LRFs). In Chapter 2.3 we pointed out that the main issue standing behind the automation of the 3D modeling pipeline is the data alignment step. This section presents a deeper analysis of the related work attempting to solve for the automation of the 3D scans alignment process, being perfectly applicable to the 2D image alignment problem.

This section focuses first on the main ingredient of the multi-view 3D scans alignment method, i.e. the *pair-wise alignment* step. The second part of this section describes how pair-wise methods are integrated within global multi-view frameworks in order to solve for the *multi-view scans alignment* problem.

4.2.1 Pair-wise Alignment

Given a pair of partially overlapping scans, we shall refer to the reference scan as the *model* and to the scan to align as the *data*. The goal is to compute the six degrees of freedom (6-DOF) (3 for rotation and 3 for translation) of the Euclidian rigid transformation $\mathbf{T}$, which will register the *data* scan with the *model*. Following the available input data, two types of scans alignment methods were reported designed to take into account whether an initial guess of $\mathbf{T}$ is provided or not to produce *coarse* (i.e. *scan matching*) or *fine alignment* (i.e. registration), respectively.

4.2.1.1 Matching

If no initial estimation of **T** is given, *direct* and *feature-based* methods were introduced to recover an approximate transformation relating two overlapped views.

The first class refers to correlation-based scan matching methods, such as the one reported in [Konolige, 2004] in which authors proceeds to an exhaustive search by varying the parameters of **T**. The combination of parameters that gives the lowest cost is accepted as the optimal match. While affordable if an initial estimation is given or for 2D scan matching, these methods are time consuming for higher dimensional spaces. In [Lucchese et al., 2002] authors reported a frequency-domain-based method which exploits the geometric regularity to solve for the range image matching problem. In [Chen et al., 1999] authors introduced the RANSAC-based DARCES method in which an exhaustive search is performed to match two partially overlapped views.

The second approach used to solve for the scan matching problem establishes correspondences between distinctive features that may be present in the overlapping area. The basic procedure involves features' extraction and matching and pose estimation based on the established correspondences. Different approaches explored a wide variety of features: edge maps [Sappa et al., 2001], lines and planes [Faugeras and Herbert, 1986], bitangent curves [Wyngaerd and Gool, 2002], surface curvatures [Yamany and Farag, 2001], surface orientation [Johnson and Herbert, 1999] and invariant features such as moments and curvatures [Sharp et al., 2002]. In [Stein and Medioni, 1992] Stein and Medioni introduced the *splash structure* - a local map describing the distribution of surface normals along the geodesic circle. Johnson in [Johnson, 1997] introduced the *spin image* idea - a surface descriptor invariant to rigid motions used to solve for the scan matching problem. Since the spin image descriptor requires uniform distribution of point for sensible discrimination, later Huber [Huber, 2002] improved Johnson's idea with *face based spin image* computed on meshes, yielding robustness to scale variations. However, spin image-based descriptors remain quite sensitive to error in the point normals and unfortunately, it is very difficult to compute reliable point normals on noisy scans, which is the case in our research work.

The general formulation of the scan matching problem does not assumes any knowledge about the environment type (i.e. either structured or unstructured), being extremely hard to solve. When dealing with homogeneous surfaces, one can imagine the difficulty to extract features and the ambiguous matches that such surfaces may lead to. The same problem may rise when attempting to match too homogeneous areas using radiometric features. Although a big part of scans alignment are applied for missions undertaken in man-made environments, we strongly believe that an environment-independent method need to be introduced in order to supply reliably the scans' alignment task in both structured and unstructured environments.

Direct and feature-based common. Generally, it is difficult to obtain precise pose estimates when using the aforementioned methods, being usually employed to supply a rough alignment. On the other hand, scans registration methods require a good initial starting point and employ different criterions to evaluate the quality of the refined estimates. Due to these reasons, several attempts reported the combination of matching followed by a registration to achieve automatically precise results [Huber and Vandapel, 2003b], [Sappa et al., 2001] [Chen et al., 1999].

Direct vs. Feature-based. At the first sight, one can propose a feature-based method to detect matches belonging to the overlapping area and estimation step through the jointly use of linear estimators [Plackett, 1972] and non-linear optimizers [Moré, 2006].

One issue may arise when employing feature-based methods, since they are subject to the scene's content which must be rich in features. For poorly structured or completely feature-less areas such methods usually lead to either ambiguous or false matches (also called false alarms or outliers). Traditionally, outliers rejection algorithms are applied to increase the accuracy of the estimates. A well known algorithm for outlier rejection is the probabilistic framework RANSAC (Random Sample Consensus) [Fischler and Bolles, 1981], which unfortunately leads to random computational time making it unsuitable for embedding processing on silicon devices. This problem gets more severe when one has to minimize the amount of the acquired data, which is the case in our research work. Since the minimum overlap provided cannot guarantee the existence of corresponding features in that particular area the algorithm may lead to lost features and pose computation failure.

A *safer* solution is to explore the whole poses' solution space to find the best transformation which minimizes the signals' dissimilarity measured via correlation techniques. Although heavy for in-situ processing, when improved in terms of rapidity, this method is the more likely to be employed since it ensures reliable data matching in previously unknown environments.

Both schools present their pros and cons and sustaining one of them is not our purpose in this dissertation. We simply point out that when attempting to provide a general solution for the multi-view alignment problem, one has to provide the needed information to solve carefully for each other's drawbacks in order to avoid algorithm's failure. To this end, this dissertation provides means to overcome both methods' drawbacks, i.e. direct and feature-based, as follwing:

- In this chapter we adopt the *safer* way and decide to deal with the computational issues of the *correlation-based* methods using calibrated constraints to limit the solution space, in conjunction with a pyramidal searching strategy to cut down the combinatory and produce precise rigid pose estimates.

- Feature-based methods' drawbacks are addressed in Chapter 7 which introduces a feature-based method which combines 2D and 3D criterions to disambiguate the feature matching task and to produce reliable pairings within an environment-independent framework.

4.2.1.2 Registration

If an initial estimation is provided, iterative methods are preferred [Besl and McKay, 1992], [Chen and Medioni, 1992], [Zhang, 1994]. The chief algorithm used to supply fine alignment is the Iteratively Closest Point pioneered by Besl [Besl and McKay, 1992]. At each ICP iteration, two steps are performed: (i) correspondences are first established between the *data* points and their nearest *model* points and (ii) the transformation $\mathbf{T}$ is then estimated via Horn's closed-form solution using quaternions [Horn, 1987], or singular value decomposition (SVD) [Hartley and Zisserman, 2004].

The algorithm minimizes iteratively the mean squared error (MSE) between the corresponding points, which is given by the following expression:

$$f(\mathbf{R},\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{p}_i^1 - \mathbf{R}\mathbf{p}_j^2 - \mathbf{t}\|^2 \tag{4.1}$$

where $\mathbf{p}_i^1 \in S_1$ and $\mathbf{p}_j^2 \in S_2$.

As stated in [Besl and McKay, 1992], the correspondences search is the most expensive step within the entire process, taking about 95% of the runtime. A kd-tree data structure [Bentely, 1975] can be used to establish faster correspondences. Another important strategy employed to speed up the registration process exploits sampling techniques to reduce the number of points [Rusinkiewicz and Levoy, 2001]. In [Greenspan and Godin, 2001] authors perform a local search within a small neighborhood around the matches resulted from the previous iteration of ICP and update them.

While being efficient with a $O(n \log n)$ complexity for n-point scans, ICP converges under the assumption that one of the datasets is included in the other. This leads to narrow convergence, implying monotonically convergence to a local minimum and therefore, the need of a good pre-alignment to ensure convergence to the correct solution.

Since its conception, many approaches have been developed aiming to widen the convergence basins. Fitzggibon in [Fitzgibbon, 2003] uses the iteratively non linear optimizer Levenberg-Marquardt (LM) to perform the residual error minimization, which allows for a robust kernel to be applied resulting in a wider convergence basin. Recently, Bae in its dissertation [Bae, 2006] introduces a pair-wise registration algorithm capable to handle up to 10° of error in the rotational pre-alignment and small translation errors through the use of geometric primitives and neighborhood search.

In order to compute more precise transformations, various attempts focus on both steps performed at each ICP iteration, by introducing to new rules for outliers' rejection and quality alignment measures. A good survey of different variations of ICP is presented in [Chen and Medioni, 1992], [Levoy et al., 2000] , [Rusinkiewicz and Levoy, 2001], [Sharp et al., 2002] and [Liu and Rodrigues, 2002]. In [Levoy et al., 2000] authors discard boundary points, while [Zhang, 1994] proposes a method to classify outliers when the point-to-point distance exceeds an automatically computed threshold. Chen and Medioni reported in [Chen and Medioni, 1992] the use of the distance between the point and the tangent plane to the corresponding point in the other view. When a good pre-alignment is available, this method yields better results than ICP. In [Masuda and Yokoya, 1995] authors combine random sampling with least median squares estimator to adjust ICP.

A quality assessment of ICP-variations can be found in [Liu and Rodrigues, 2002], [Dalley and Flynn, 2002], [Rusinkiewicz and Levoy, 2001], [Rodrigues et al., 2002]. However, since algorithms were developed on different databases using different metrics, it is difficult to establish a correct evaluation. Nevertheless, the aforementioned comparative studies lead to a common conclusion suggesting that although there is still room for further extensions of ICP, its improvements may become helpless when initial alignments are not feasible and methods attempting to overcome the pre-alignment requirement may lead to erroneous poses.

Recent trends. The pose-space search algorithm introduced in [Robertson and Fisher, 2002] shows that tractable solutions stand in the development of a new fundamental pose searching strategy, rather than using the correspondence-based search of ICP-like methods. These methods aim at exploring a huge solution space to find the best transformation which aligns two views *precisely* in a reasonable time. Reported frameworks employed stochastic optimization techniques such as Genetic Algorithms (GAs) [Man and Kwong, 1996] and Simulated Annealing (SA) [Kirkpatrick et al., 1983] to supply coarse registration in conjunction with local search heuristics to produce fine alignments. In [Silva et al., 2005] authors focus on obtaining precise alignments using GAs and introduce the Surface Interpenetration Measure (SIM) to compute the interpenetration of the two registered range views.

Figure 4.1 presents an overview of the existing pair-wise alignment methods and highlights our proposal wrt the state of the art capabilities. The pair-wise scans alignment method introduced in this dissertation belongs to the *pose-search* methods, employs calibration constraints and performs a pyramidal searching strategy to cut down the combinatory. The proposed algorithm provides *precise rigid estimates* without requiring for an initial guess by matching either intensity or depth 2D panoramic views using dense-correlation via quaternions. The estimation process starts by first exploring the entire solution space to localize the global minimum which is further refined at higher resolution levels of the pyramidal structure.

4.2.2 Multi-view Alignment

Generally, the multi-view scan alignment process can be performed either *sequentially* or *simultaneously*. *Sequential* methods [Chen and Medioni, 1992], [Turk and Levoy, 1994] are susceptible to propagate and accumulate errors from one iteration to another. Nevertheless, if one can ensure precise alignment, they are computationally more attractive and require less memory resources. A general approach was introduced in the literature [Ikeuchi and Sato, 2001] to perform *simultaneous* multi-view fine alignment which consists in performing an initial alignment between each overlapped pair which are further refined within the global registration phase to distribute all errors among all alignments, being followed by an integration step [Ikeuchi and Sato, 2001], [Masuda, 2002] [Shum et al., 1997] and [Dorai et al., 1998].

Global rigid alignment algorithms have been studied by [Bergevin et al., 1996], [Benjemaa and Schmitt, 1997], [Pulli, 1999] and [Mitra et al., 2004]. In [Stamos and Leordeanu, 2003] an automatic feature-based range image registration technique for 3D modeling of urban scenes is proposed. Authors are extracting high-level entities: 3D lines and planar regions in order to compute a rigid transformation between two overlapping scans. While improving the state of the art, this method remains limited to the application field: 3D modeling of urban structured scenes.

Several global registration methods were introduced based on a physical equivalent model [Stoddart and Hilton, 1996] later optimized using a multi-resolution framework [Eggert et al., 1998], or using a network of views [Bergevin et al., 1996] and [Huber and Vandapel, 2003a]. In [Huber and Vandapel, 2003a] authors introduce a complete system for 3D modeling which employs spin images to obtained initial guess, proposes to find the minimum spanning tree in a graph and uses a topological inference criterion [Sawhney et al., 1998] to verify the consistency of the global alignment. Although, the aforementioned methods have been successfully applied to several cases [Huber and Herbert], [Huber, 2002], [Huber and Vandapel, 2003a], they have $O(N^2)$ complexity in the number of views which limits the processing to sub-maps containing about 50 views. Consequently, these methods are computationally unaffordable when it comes to large-scale and complex environments for which a high amount of views is required to cover the entire site without omitting the occluded areas.

Non-rigid alignment. The aforementioned techniques assumes that the sensing device captures the identical geometry of the scene. However, when it comes to scanner calibration errors or noise, this assumption can easily be violated, resulting in slightly warped data. This causes rigid alignment algorithms to diverge, since they model only rigid motion. In order to ensure robustness to the scanner miscalibration, different *non-rigid alignment* techniques have been introduced using an hierarchical ICP approach [Ikemoto

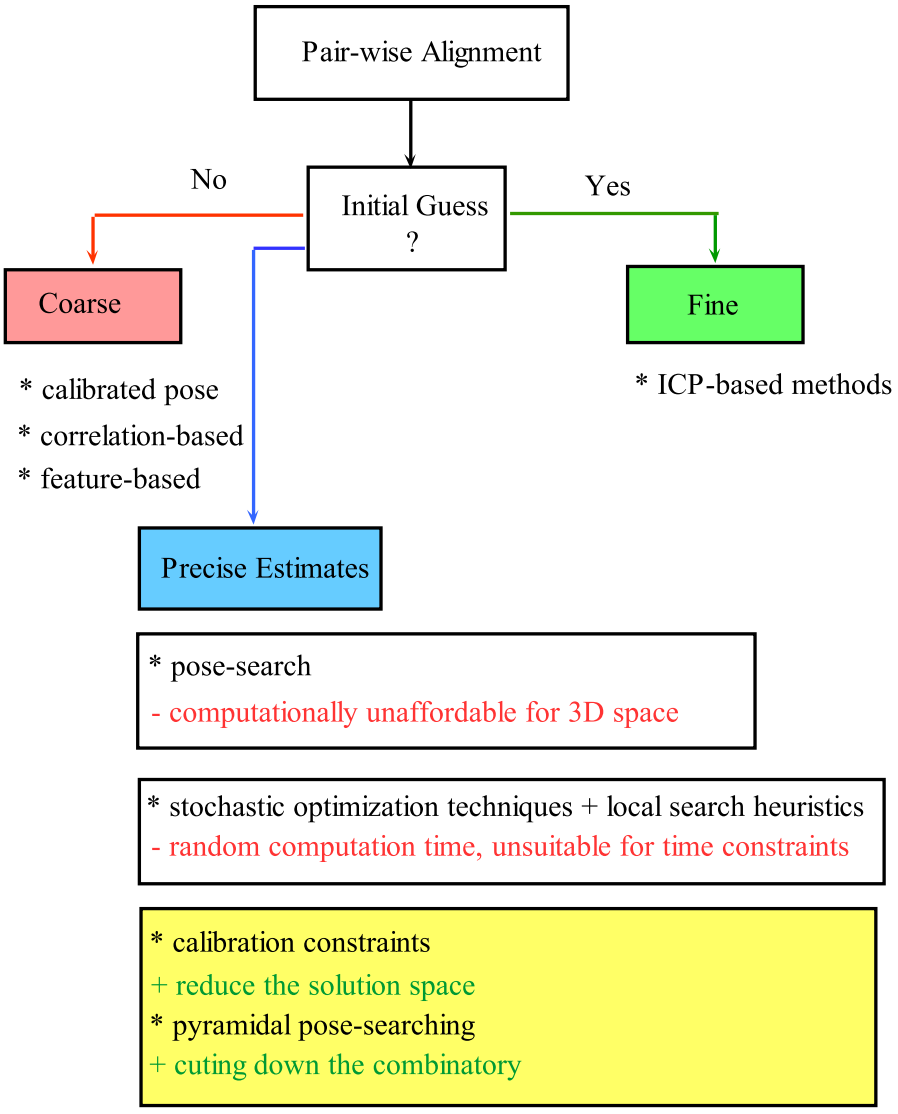


Figure 4.1: Classification of pair-wise scan alignment methods. Our proposal is highlighted in yellow.

et al., 2003], affine transformation to align body scans to a template [Allen et al., 2003a] and for global registration of images [Zhang and Rangarajan, 2004]. On the downside, such methods require dense 3D scans which limits their applicability to small-scale objects.

4.2.3 Taxonomy and Open Issues

After reviewing the available scans alignment methods, we provide a taxonomy of the existing methods and state several open issues which need to be addressed when attempting to solve for the automation of the multi-view fine alignment process.

Taxonomy. Figure 4.2 illustrates a taxonomy of the main techniques employed to integrate the existing pair-wise techniques within the multi-view alignment process. *Sequential* frameworks powered by fine pair-wise alignment techniques are more attractive in terms of accuracy and computational resources. On the other hand, *simultaneous* approaches aim at gaining computation time by performing only the matching phase and thus, paying the price of an accumulated error. Although the following global registration step may distribute the accumulated error among all poses, the residual error requires a global consistency test whose complexity grows quadratically with the number of views, leading to unfeasible schemes for large-scale environments, which is one of our main concerns in this dissertation.

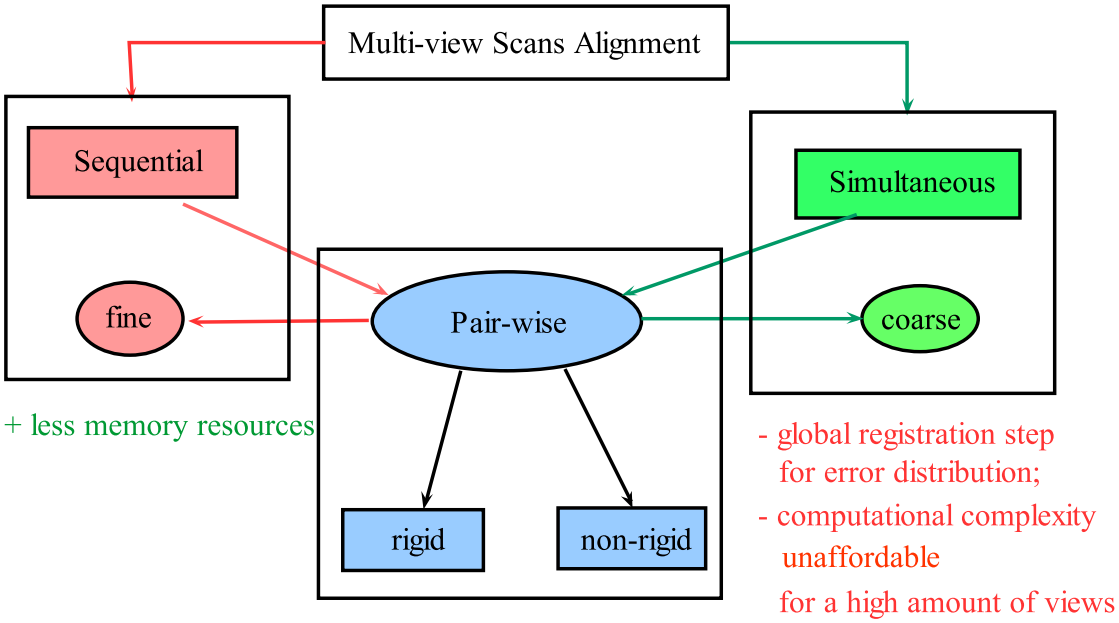


Figure 4.2: Taxonomy of the multi-view scans alignment approaches. Sequential approaches are generally powered by a fine rigid pair-wise alignment, while simultaneous methods employ coarse matching which is further refined via global registration frameworks.

Since the accumulated errors increase with the number of views to be aligned, one important aspect to be taken care of is to minimize the number of views required to recover entirely the 3D scene model. This issue has also been emphasized in [Ikeuchi and Sato, 2001] when attempting to minimize the amount of data to be acquired in order to avoid expensive acquisition and processing. While aiming to minimize the amount of the

acquired and processed data, it is important to ensure the robustness of the algorithms to low-overlapped views. Another issue which needs to be addressed is how to evaluate the minimum overlapping area between two views in order to guarantee a precise alignment. Several papers addressed this problem by calculating the overlapping areas between views and measuring the registration quality [Dalley and Flynn, 2002], [Huber and Vandapel, 2003b], [Silva et al., 2003]. Huber and Herbert in [Huber and Vandapel, 2003b] proposed an exhaustive search over the entire network of views to find the multi-view registration solution giving the lowest global error. Authors defined the overlapped points concept based on two thresholds which cannot define overlapped areas precisely.

Open issues. We resume hereafter several open issues which need to be addressed in order to produce an environment-independent method capable to solve automatically for the multi-view scans alignment task:

- handle the impossibility to supply initial alignment through the use of navigation sensors or manual intervention;
- deal with the absence of reliably trackable features;
- provide automatically the minimum overlapping area required to guarantee a precise alignment while minimizing the amount of the acquired data;
- replace the coarse-to-fine alignment framework by solving simultaneously for the *poses estimates* and the *corresponding points* to produce *precise rigid estimates*;
- since one view may overlap several others, it is necessary to detect which one produces the *optimal absolute poses*;
- when simultaneous global registration methods are used, reduce the combinatory of the global consistency test.

Throughout this chapter we focus on providing several means to solve for the aforementioned problems raised by the automation of the multi-view scans alignment task.

4.3 3D Mosaicing Acquisition Scenario

As mentioned in Chapter 3, ARTVISYS achieves 3D scene model completeness through the use of a *mosaic-driven acquisition scenario* for which a brief description was provided in Section 3.3. This chapter is concerned with the multi-view scans alignment problem for generating in-situ 3D mosaics from several partially overlapped scans acquired from the same 3D pose of the system. To this end, this section introduces a 3D mosaicing acquisition scenario which is integrated within the mosaic-driven acquisition scenario proposed in Section 3.3.

In Section 3.2 we presented the hardware design of the proposed system, which includes a Trimble® scanning device illustrated in Figure 3.2 a) providing a cloud of 3D points and their associated light intensity backscattering, within a field of view of 360° horizontally x 60° vertically, as shown in Figure 3.2 b). When mounted on a tripod, due to the vertical narrow field of view, the scanning device is not suitable for the acquisition coverage of ceiling and ground. Therefore, we manufactured in our laboratory a L-form angle-iron shown in Figure 4.3 a).

The L-mount-laser prototype illustrated in Figure 4.3 b) captures all the area around its optical center within 360° vertically and 60° horizontally as shown in Figure 4.3 c),

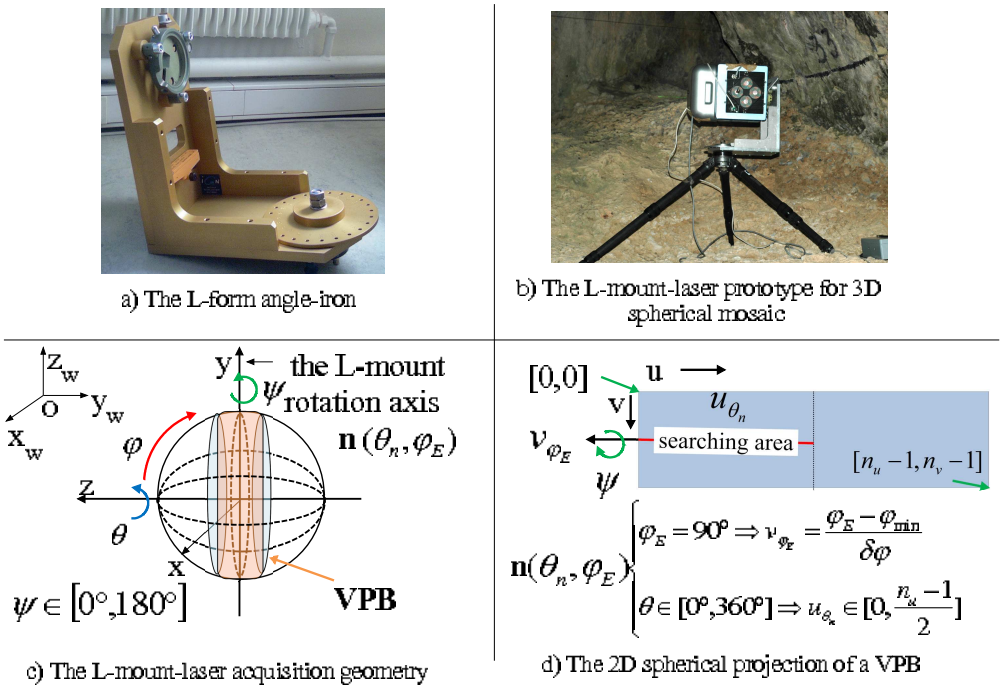


Figure 4.3: The 3D mosaicing acquisition geometry.

which we call a *vertical panoramic band* (VPB). Given a spatial position of the tripod, which we call a *station*, the scenario consists in acquiring multiple overlapping VPBs in order to provide a fully 360° × 180° 3D spherical view. For this purpose, the L-mount-laser is turned around its vertical axis **n** (superposed with the scan equator axis, Oy) with different imprecisely known orientations ψ, acquiring one VPB for each orientation, as shown in Figure 4.4. The L-mount-laser rotation angle ψ may vary within the range of [0°, 180°]. For this experiment the L-mount-laser was turned manually, but using a non-calibrated turning device it is straight forward. Generally, $N_{scenario} = 4$ VPBs are acquired to provide a fully 3D spherical view, separated by a rotation $\psi_{max} \simeq 45°$ providing an overlap of $\simeq 33\%$ which our algorithm can handle (to be compare to the state of the art [Makadia et al., 2006], for which a minimum overlap of 45% is required).

Minimum overlap guaranteed. The proposed acquisition scenario facilitates considerably the scan matching task providing a constant and minimal overlapping area situated at the bottom (ground) and top (ceiling) areas of the 3D spherical view. This is an important key issue when performing 3D modeling tasks in large-scale environments, where the amount of the acquired and processed data must be minimized.

This chapter introduces an automatic scan alignment procedure which aligns the 4 VPBs wrt a global coordinate system and integrates them into a single 3D entity, providing thus *in situ* a fully 3D spherical view of the system's surrounding.

When comparing the proposed 3D mosaicing sensor to several 3D scanning devices designed to acquire directly a 3D mosaic, such as Leica HDS3000®, Faro Laser Scanner Photon® or LiDAR®, which aimed to supply 3D modeling and path planning operations for missions undertaken on Mars [Rekleitis et al., 2009], our system provides completeness through the use of mosaic and avoid data redundancy by acquiring partial mosaics when occlusions are encountered. More precisely, instead of acquiring a fully spherical mosaic in occluded areas - which is computationally unaffordable for in-situ processing - we sense

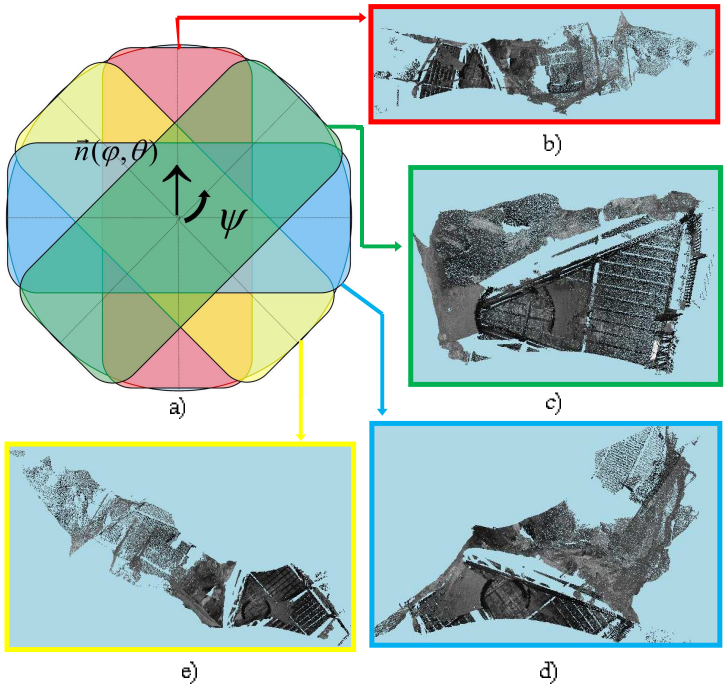


Figure 4.4: Example of the 3D mosaicing acquisition scenario performed in the Tautavel prehistoric cave - France. (a)Top view of the acquisition: the laser acquires 4 VPBs as it rotates around its vertical axis $\mathbf{n}(\theta\mathbf{n}, \varphi\mathbf{n})$ with different values of ψ: (b) VBP 1 corresponding to $\psi \approx 0°$, (c) VPB 2 for $\psi \approx 45°$ (d) VPB 3 for $\psi \approx 90°$, (e) VPB 4 for $\psi \approx 135°$.

only the occluded areas by acquiring partial mosaics, avoiding therefore data redundancy while achieving site completeness.

The next section provides an overview of the proposed multi-view alignment method associated to the proposed 3D mosaicing acquisition scenario.

4.4 Algorithm Overview

Generally speaking, when solving for the data alignment problem, one has to carefully formalize the motion encountered by the sensing device and set a strategy to solve for absolute poses.

Motion parametrization. Theoretically, for a single spatial position, the system delivers a sequence of partially overlapped scans separated by a rotation. However, in practice the system's instability may introduce small amounts of parallax between scans. Moreover, when the center of mass of the the capturing device is not superposed with the rotation center of the platform, *non-rigid motions* are introduced. Other sources of non-rigid motion are the system's vibrations which are amplified by the heavy capturing device, calibration errors, unmodeled sensor's distortions and noise.

Solving for the absolute poses. The proposed multi-view scans alignment is performed in a *sequential* fashion, being powered by a pair-wise alignment procedure which outputs *precise rigid relative poses estimates*, overcoming therefore the accumulated errors caused by multi-view methods integrating coarse pair-wise algorithms. Figure 4.5 synthesizes the main ingredients included in the proposed multi-view alignment method which

are justified hereafter:

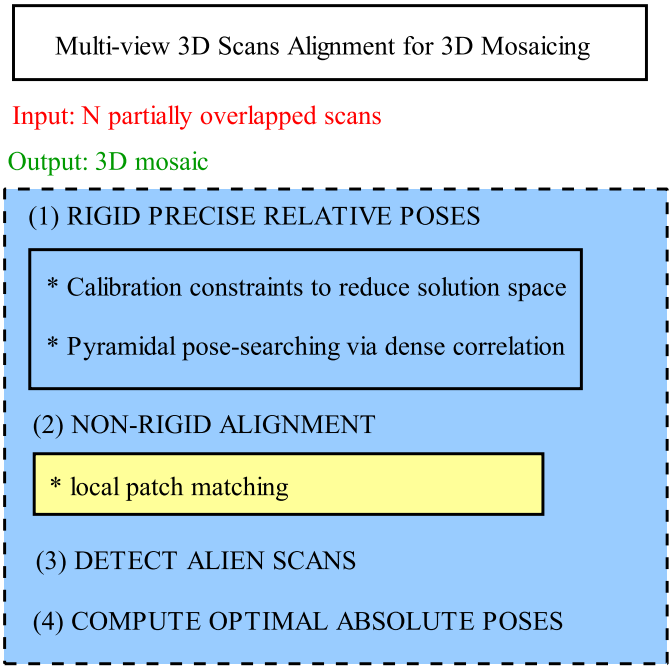


Figure 4.5: Global overview of the proposed multi-view scans alignment method for generating in-situ 3D mosaics. In yellow is emphasized the optional character of the non-rigid alignment phase.

(1) Precise pair-wise rigid estimates. The pair-wise alignment falls in the category of *pose-search methods* and supplies both matching and registration phases to produce precise rigid poses estimates, eliminating therefore the initial guess requirement of the traditionally used ICP-methods. In order to cope with in-situ processing constraints, we reduce the solutions' space using calibration constraints and cut down the combinatory by employing a pyramidal searching strategy.

While focusing on acquiring the minimum amount of data required to ensure a reliable scan matching and fast processing, the algorithm must deal with poor overlaps. To do so, the alignment process must exploit all the existent information available in the overlapping area. For this reason, the core of the algorithm performs pose estimation by matching either intensity or depth 2D panoramic views using dense correlation via quaternions. Two main raisons stand behind the use of the dense correlation instead of the use of feature-based approaches: (i) feature extraction and matching for poor overlaps may easily fail or result in ambiguous matches, (ii) this aspect gets worse when applying such scenarios to feature-less areas, since the poor overlap cannot guarantee the existence of such features in that particular common area. Since in our research work we are interested in providing an environment-independent framework for in-situ processing in previously unknown environments, we aim at acquiring and processing the minimum amount of data needed. Consequently, we employ a correlation-based method in order to exploit the entire information contained in the overlapping region.

As mentioned in the previous section, due to the mutual dependency laying between the relative poses and the overlaps, the multi-view scans matching problem is a difficult task. The proposed pair-wise matching procedure solves *simultaneously* the above interrelated

problems by matching 2D panoramic views using dense-correlation via quaternions. In addition, the 2D panoramic images provides spatial and appearance constraints increasing therefore the robustness of the scan matching process.

The pyramidal searching strategy emphasizes the tradeoff between the two key aspects of any scan matcher: the *accuracy* and the *robustness*. The accuracy is related to the subpixel precision attached to the dense correlation step, while the robustness component is related to the capability to handle large and small motions by performing estimation at the lowest and highest resolution levels of the pyramid, respectively.

(2) Non-rigid motion estimation. Optionally, when non-rigid motions are encountered, a *non-rigid pair-wise registration* step can be applied in order to compensate the eventual non-rigid motions introduced by the system. To this end, Chapter 5 introduces a non-rigid pair-wise registration technique. Although originally designed and tested on 2D color images for 2D optical mosaicing purposes, it is straight forward to apply it on either intensity or range 2D panoramic images.

Since the pair-wise procedure outputs precise poses estimates, there is no need to perform a global registration process for poses' refinement. In exchange, the multi-view alignment process has to detect whether the sequence contains non-overlapped scans and to find out wrt which view the algorithm must register all scans in order to benefit of accurate estimates provided by pairs with high overlaps.

(3) Alien scans' detection. Since no knowledge is provided about the overlapped scans, the multi-view alignment process starts by first detecting the *alien scans*, i.e. scans which do not belong to the currently processed 3D mosaic.

(4) Detect optimal absolute poses. Since one view may overlap several others, the multi-view alignment procedure detects the best reference view which optimally register all views into a global 3D scene model.

The scan matching procedure ends by integrating all scans into a single 3D entity, hence providing a complete spherical view of the scene for a given *station*.

This chapter focuses on the rigid alignment part, describing the pair-wise procedure in Section 4.5 - corresponding to phase (1) in Figure 4.5, and the multi-view scans alignment process in Section 4.7 - corresponding to phases (2) and (3) in Figure 4.5.

4.5 Free-Initial Guess Pair-wise Alignment for Precise Rigid Estimates

This section focuses on the pair-wise alignment process corresponding to phase (1) in Figure 4.5. First, we briefly illustrate how we build 2D panoramic images from 3D point clouds. The following section defines the 3D rigid poses' solution space in the 2D panoramic image space under calibration constraints. In Section 4.5.3 we describe the rotation estimation process which can be performed in either intensity or depth mode, following the input data provided by the capturing device. The next section exploits the rotationally aligned scans to estimate the translation. The pair-wise process ends up by performing an incremental refinement of the obtained 3D pose using a pyramidal pose searching strategy to produce *rigid precise poses estimates*.

4.5.1 From 3D Point Clouds to 2D Panoramics

Let S_1 and S_2 be two sets of 3D points expressed in Cartesian coordinates, with respect to the laser scan reference system. They are defined as $S_1 = \{\mathbf{s}_1^i | i = 0, ..., N_1 - 1\}$,

$S_2 = \{\mathbf{s}_2^j | j = 0, ..., N_2 - 1\}$, where N_1 and N_2 stand for the numbers of 3D points contained by the scans S_1 and S_2, respectively. The four-dimensional vector $\mathbf{s} = (p_x, p_y, p_z, \mathbf{i})$ represents the 3D coordinates of a point and the associated reflected intensity value defined on $[0, 255]$. The proposed pair-wise scan matcher performs pose estimation by correlating either intensity or depth values in the 2D panoramic image space which represents the 2D spherical projection of a VPB.

In order to compute the 2D spherical projection of a generic VPB, we first automatically recover the internal parameters of the spherical acquisition $\mathcal{P}_{\mathcal{I}}$: the acquisition steps $(\delta\theta, \delta\varphi)$ and the field of view ($[\theta_{min}, \theta_{max}]$,$[\varphi_{min}, \varphi_{max}]$) via a triangulation procedure.

The 2D image projection which assigns to each direction $(\theta, \varphi)^t$ a 2D image location $\mathbf{m} = (u, v)^t$ and conversely is defined below:

$$\mathbf{S}_n{:}\mathbf{p}(\theta, \varphi) \rightarrow \mathbf{m}(u, v) \tag{4.2}$$

where, u and v denote the column and the row of a generic panoramic pixel $\mathbf{m}$, as shown in Figure 4.3 d). The quantities $I(\mathbf{m})$ and $D(\mathbf{m})$ are defined as the greyscale and depth value corresponding to the 2D image location $\mathbf{m} = (u, v)^t$. The rectangular support of width n_u and hight n_v of the 2D spherical panoramic corresponds to $\theta_{max} - \theta_{min} = 360°$ and $\varphi_{max} - \varphi_{min} = 60°$, which is the effective laser's field of view. Figure 4.6 illustrates the 2D spherical projections of two overlapped VPBs acquired in the Moulin de Languenay prehistoric cave. Missing data is observed systematically in each view, in the area indicated by the arrow which corresponds to the area situated right underneath the sensor which cannot be digitized due to the system's montage.

4.5.2 Constructing Pose's Space Candidates under Calibration Constraints

The 6 DOF rigid transformation can be written as a 4×4 matrix:

$$\mathbf{T} = \begin{bmatrix} \mathbf{R}_{3\times3} & \mathbf{t}_{3\times1} \\ \mathbf{0}_{1\times3} & 1 \end{bmatrix} \tag{4.3}$$

We solve for the pose estimation $\mathbf{T}$ in two steps, within a hybrid framework: the rotation $\mathbf{R}$ is first computed by matching either intensity or depth data in the 2D panoramic image space, while translation is calculated a posteriori by projecting back in the 3D space the rotationally aligned panoramic images.

Rotation parametrization. There are several choices for representing 3D rotations: 3×3 orthogonal matrix, angle-axis representation or unit quaternions. An unit quaternion is a normalized four-dimensional vector $\dot{\mathbf{q}} = (q_0, q_x, q_y, q_z)$ which provides a more compact representation than an orthogonal matrix (9 parameters), being suitable for numerical optimization techniques. A rotation of angle ψ around an axis $\mathbf{n}$ can be represented by the unit quaternion $\dot{\mathbf{q}} = (\cos\frac{\psi}{2}, \sin\frac{\psi}{2}\,\hat{\mathbf{n}})$ where $\hat{\mathbf{n}}$ is the unit vector $\hat{\mathbf{n}} = \frac{\mathbf{n}}{\|\mathbf{n}\|}$. The orthogonal matrix $\mathbf{R}(\dot{\mathbf{q}})$ corresponding to a rotation given by the unit quaternion $\dot{\mathbf{q}}$ is expressed by:

$$\mathbf{R}[\dot{\mathbf{q}}] = \begin{pmatrix} q_0^2 + q_x^2 - q_y^2 - q_z^2 & 2(q_x q_y - q_0 q_z) & 2(q_0 q_y + q_x q_z) \\ 2(q_0 q_z + q_x q_y) & q_0^2 - q_x^2 + q_y^2 - q_z^2 & 2(q_y q_z - q_0 q_x) \\ 2(q_x q_z - q_0 q_y) & 2(q_0 q_x + q_y q_z) & q_0^2 - q_x^2 - q_y^2 + q_z^2 \end{pmatrix} \tag{4.4}$$

For further lecture on quaternions the reader can refer to [Horn, 1987], [Howell and Lafon, 1975], [Salamin, 1979].

The pose estimation process combines the acquisition geometry and the sensing device intrinsics in order to decrease the poses' solution space as described hereafter:

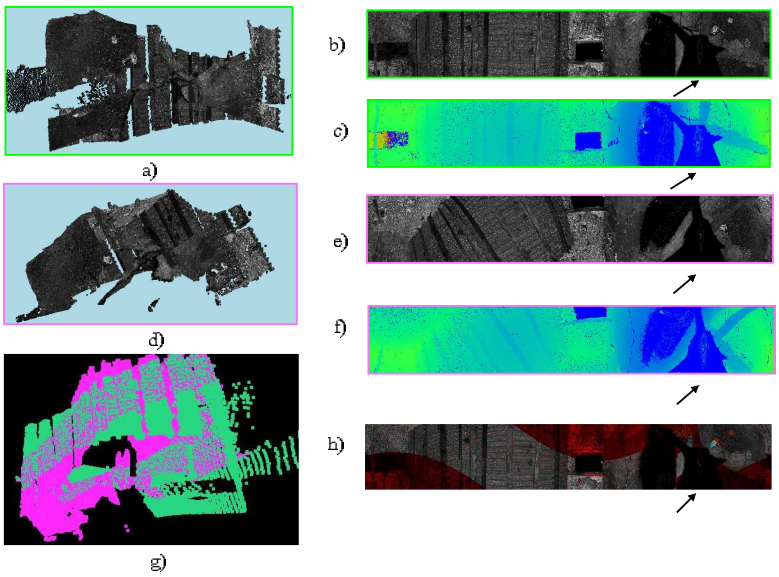


Figure 4.6: Example of 2D panoramic views obtained from 3D point clouds acquired in the Moulin de Languenay prehistoric cave, France: a) the VBP corresponding to the scan S_1, b) the 2D intensity panoramic image obtained from S_1 noted I_1, c) the 2D depth panoramic image corresponding to S_1 noted D_1, d) the VBP corresponding to the scan S_2 which overlaps partially S_1, e) the 2D intensity panoramic image obtained from S_2 noted I_2, f) the 2D depth panoramic image corresponding to S_2 noted D_2. In order to illustrate the surface shared by S_1 and S_2, figures g) and h) depicts the registered scans in 3D and 2D image space, respectively: h) shows the aligned scans with S_1 - green and S_2 - magenta, g) a 2-channel 2D panoramic image containing I_1 - red channel, I_2 - green channel. Their superposition produces the grey-level area corresponding to the overlap region. The area indicated by the arrow corresponds to the area situated right underneath the sensor which cannot be digitized due to the system's montage. This leads to a considerable amount of missing data for which a robust scans alignment method must be designed.

- following the *acquisition geometry* illustrated in Figure 4.3 c), the sensing device undertakes rotations of angle ψ around the L-mount laser rotation axis given by $\mathbf{n}(\theta_{\mathbf{n}}, \varphi_{\mathbf{n}})$ in order to acquire several partially overlapped scans, as shown in Figure 4.7 a).

- following the *laser construction*, the L-mount laser rotation axis $\mathbf{n}(\theta_{\mathbf{n}}, \varphi_{\mathbf{n}})$ is contained by the laser's equator plane given by (θ, φ_E), as shown in Figure 4.7 b). In addition, by construction, the scanning device has the equator fixed at $\varphi_E = \frac{\pi}{2}$.

The two aforementioned relations can be summarized by the following expression:

$$\mathbf{n}(\theta_{\mathbf{n}}, \varphi_{\mathbf{n}}) \subset (\theta, \varphi_E) \Rightarrow \varphi_{\mathbf{n}} \equiv \varphi_E \tag{4.5}$$

giving rise to the following calibration constraint:

$$\mathbf{n}(\theta_{\mathbf{n}}, \varphi_{\mathbf{n}}) \equiv \mathbf{n}(\theta_{\mathbf{n}}, \varphi_E) \tag{4.6}$$

which reduces the 3D solution space of the rotation $\dot{\mathbf{q}}(\psi, \mathbf{n}(\theta_{\mathbf{n}}, \varphi_{\mathbf{n}}))$ to a 2D one, focused on the estimation of the remaining unknown angles $(\psi, \varphi_{\mathbf{n}})$.

Given the acquisition geometry (Figure 4.3 c)) and the 2D spherical projection of a VPB shown in Figure 4.3 d), we solve for the unit vector $\dot{\mathbf{q}}(\psi, \mathbf{n})$ in the panoramic image space. We search for the 2D image location $(u_{\theta_\mathbf{n}}, v_{\varphi_\mathbf{n}})$ corresponding to the spherical coordinates $(\theta_\mathbf{n}, \varphi_\mathbf{n})$ of the L-mounted-laser rotation axis $\mathbf{n}$. Following the calibration constraint given in Equation (4.6), we can deduce the equator's row location v_{φ_E} in the 2D panoramic image space using the internal parameters $\mathcal{P}_\mathcal{I}$ of the spherical acquisition φ_{min} and $\delta\varphi$ (previously recovered in Section 4.5.1), as follows:

$$v_{\varphi_E} = \frac{\varphi_E - \varphi_{min}}{\delta\varphi} \tag{4.7}$$

The remaining unknowns, $(\psi, \theta_\mathbf{n})$ are computed by varying the parameters of ψ and $\theta_\mathbf{n}$ within an homogeneous searching area, $\mathcal{P}_{SA}$, which is recursively updated at each pyramidal level. $\mathcal{P}_{SA}$ is defined by the following parameters: the ψ range: $[\psi_{min}, \psi_{max}]$, the $\theta_\mathbf{n}$ range: $[\theta_{\mathbf{n},min}, \theta_{\mathbf{n},max}]$, the ψ step $\delta\psi$ and the $\theta_\mathbf{n}$ step $\delta\theta_\mathbf{n}$. We search for the column $u_{\theta_\mathbf{n}}$ by varying $\theta_\mathbf{n}$ in the searching area within a range of $[0°, 180°]$, as shown in Figure 4.3 d). The rotation angle is computed by applying rotations of angle $\psi \in [0°, 180°]$ around $\mathbf{n}(\theta_\mathbf{n}, \varphi_E)$ axis to the 3D points of S_2 and matching the corresponding transformed pixels with pixels from I_1, which was previously obtained by from S_1.

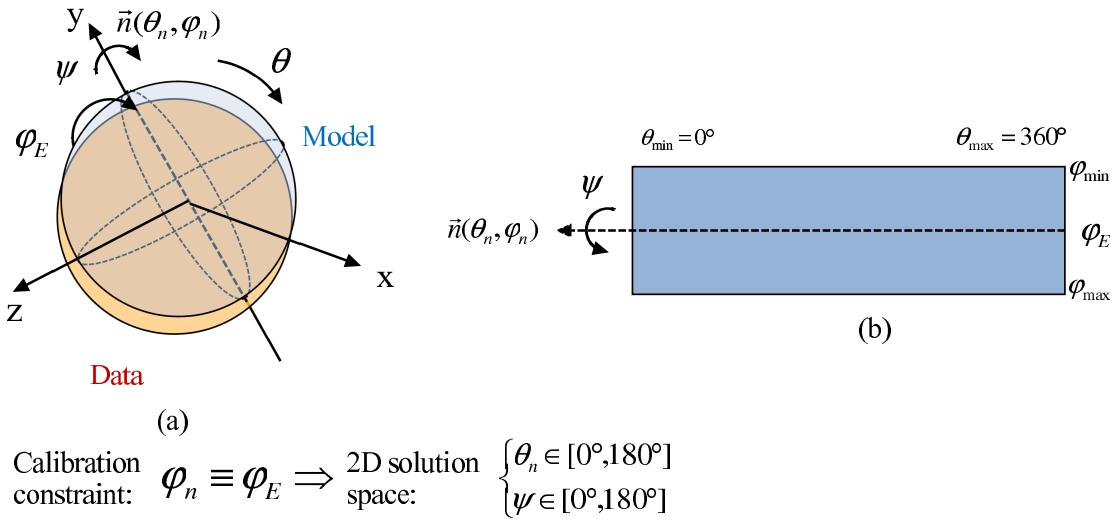


Figure 4.7: Calibration constraints induced by the acquisition geometry of the L-mount laser prototype. (a) the L-mount laser prototype acquisition geometry, (b) the intrinsics of a 2D spherical projection of a VPB acquired by the L-mount laser prototype.

Translation computation. In the 3D mosaicing acquisition scenario presented in Section 4.3 translations are negligible, i.e. less than $\frac{1}{4}$ pixel at levels $l > 0$, shadowing the system's instability during the sensing device rotations which introduces small amounts of parallax. Nevertheless, the pyramidal framework allows to handle larger amounts of translations. Further details on the translation estimation are provided in Section 4.5.4.

Let us now describe the core of the pair-wise scan matcher in more mathematical details. At every pyramidal level $l = 0, .., L_{max}$, where L_{max} defines the height of the pyramid, the goal is to find the 3D rigid transform $\mathbf{T}^l = [\mathbf{R}, \mathbf{t}]^l$. Since the same type of operation is performed at each level l, let us drop the superscript l through the following description.

4.5.3 2D-Panoramic-based Rotation Estimation

For a given 3D rigid transformation $\mathbf{T}_{12}$ we can directly obtain the 2D image coordinates $\hat{\mathbf{m}}_1$ of a 3D point in scan S_1 from its pixel coordinates $\mathbf{m}_2$ in I_2 using the following composed projection:

$$\hat{\mathbf{m}}_1 = \mathbf{S}_1 \circ \mathbf{T}_{12} \circ \mathbf{S}_2^{-1}(\mathbf{m}_2) \tag{4.8}$$

The rotation computation is performed by matching 2D panoramic views within a pyramidal dense correlation framework using quaternions. The algorithm solves for the rotation estimation by exploiting either intensity or depth 2D panoramic views, following the input data provided by the capturing device. The following subsection describe the intensity mode of the rotation estimation process. The depth mode employs an similar procedure whose description can be found in Appendix B.1.

4.5.3.1 Intensity Mode

For a given quaternion $\dot{\mathbf{q}}(\psi, \mathbf{n}[\theta_\mathbf{n}, \varphi_E]), (\psi, \theta_\mathbf{n}) \in \mathcal{P}_{SA}$, we map pixels $\mathbf{m}_2^j$ from I_2 in the I_1's space using the spherical projection expressed in Equation (4.8). The optimal rotation is obtained by minimizing the difference in brightness between the two panoramic images I_1 and I_2 in the overlapping region. The dissimilarity in brightness is measured by Mean Absolute Differences (MAD) for all corresponding pixels belonging to the overlap, being defined on the interval $[0, 1]$.

$$\mathbf{E}^I(\psi, \mathbf{n}) = \frac{1}{255 N_{12}} \sum_{j=0}^{j=N_2-1} \Phi_j^I |I_2(\mathbf{m}^j) - I_1(\hat{\mathbf{m}}_{\psi,\mathbf{n}}^j)| \tag{4.9}$$

Φ_k^I defines a characteristic function which takes care of "lost" (i.e. the pixel falls outside of the rectangular support of I_1) and "zero" pixels (i.e. missing data either in $I_1(\hat{\mathbf{m}}_{\psi,\mathbf{n}}^j)$ or $I_2(\hat{\mathbf{m}}_{\psi,\mathbf{n}}^j)$), which may occur when mapping pixels $\hat{\mathbf{m}}_{\psi,\mathbf{n}}^j$ in the I_1's space. Thus, we penalize "lost" and "zero" pixels using the following weighting function:

$$\Phi_j^I = \begin{cases} 0, & \text{if } I_1(\hat{u}^j, \hat{v}^j) = 0 \text{ or } I_2(\hat{u}^j, \hat{v}^j) = 0 \\ 0, & \text{if } \hat{u}^j, \hat{v}^j < 0 \text{ or } \hat{u}^j > n_u - 1 \text{ or } \hat{v}^j > n_v - 1 \\ 1, & \text{otherwise} \end{cases} \tag{4.10}$$

The number of N_{12} denotes the number of pixel matches found between I_1 and I_2 for which $\Phi_j^I = 1$, and defines the overlapping area $\mathbf{O}^I[\psi, \mathbf{n}]$ of the corresponding rotation $\mathbf{R}^I[\psi, \mathbf{n}]$ between I_1 and I_2. The overlap is evaluated with respect to the reference image, I_1. Hence,

$$\mathbf{O}^I[\psi, \mathbf{n}] = \frac{N_{12}}{N_1} \tag{4.11}$$

being essentially a subunit value.

The optimal rotation $\hat{\mathbf{R}}^I[\dot{\mathbf{q}}(\hat{\psi}, \hat{\mathbf{n}})]$ is given by the minimal dissimilarity measure, thereby maximizing the overlap $\hat{\mathbf{O}}^I[\hat{\psi}, \hat{\mathbf{n}}]$:

$$\hat{\mathbf{R}}^I[\hat{\psi}, \hat{\mathbf{n}}] = \arg \min_{(\psi,\mathbf{n}) \in \mathcal{P}_{SA}} \mathbf{E}^I(\psi, \mathbf{n}) \tag{4.12}$$

When referring to the GS200 laser range finder data sheet, we can find that the scanning device delivers 3D measurements with an accuracy of 1.5mm which decreases starting with

50m. In the research work presented in this dissertation, we employ the scanning device in underground environments with depth values up to 30m. In this sense, the Φ function can be extended to weight laser measurements wrt their accuracy.

4.5.4 Translation Estimation

Let us now drop the superscripts I and D through the following description and consider that an optimal rotation $\hat{\mathbf{R}}$ was computed using either the intensity or depth mode. The corresponding overlap $\hat{\mathbf{O}}$ and dissimilarity score $\hat{\mathbf{E}}$ are used for further computations.

We use the rotationally aligned scans to eliminate bias with respect to the median error and to compute the optimal translation vector $\hat{\mathbf{t}}$, which is given by the difference between the two barycenters of the 3D coordinates points belonging to the overlapping areas of each scan. Since the rotation is computed using the non-centered 3D coordinates, an additional translation appears when applying $\hat{\mathbf{R}}^{-\mathbf{1}}$ to the 3D coordinates of S_2. The rotationally aligned scans are then obtained by compensating the total translation, $\tilde{\mathbf{t}} = \hat{\mathbf{t}} + \mathbf{t}_{res}$.

When computing $\tilde{\mathbf{t}}$ based on correspondences between 3D coordinates points, we have to deal with outliers. We minimize a weighted residual error metric defined by the euclidian norm in order to discard false matches between the 3D points correspondences with respect to the optimal rotation model $\hat{\mathbf{R}}$. The goal is to minimize the criterion $\mathbf{Q}_{\hat{\mathbf{R}}} = \sum_{k=0}^{k=N_{12}-1} r_k$, where r_k are the weighted residual errors defined by:

$$r_k = \Phi_k \|(\mathbf{p}_1^k - \hat{\mathbf{R}}^{-\mathbf{1}}\mathbf{p}_2^k\| \tag{4.13}$$

The discarding procedure eliminates 3D points correspondences with a residual error grater than a threshold $\xi = f(\bar{\mathbf{r}}, \sigma_{\bar{\mathbf{r}}})$, where $\bar{\mathbf{r}}$ and $\sigma_{\bar{\mathbf{r}}}$ are the mean and the standard deviation of the residual errors r_k obtained with respect to the rotation model. Hence, the outliers are penalized by updating the function Φ_k as follows:

$$\Phi_{k,\xi} = \begin{cases} 0, & \text{if } r_k > \xi \\ 1, & \text{if } r_k < \xi \end{cases} \tag{4.14}$$

with $f(\bar{\mathbf{r}}, \sigma_{\bar{\mathbf{r}}}) = \bar{\mathbf{r}} + \alpha\sigma_{\bar{\mathbf{r}}}$. In the presented work, the threshold corresponds to a rotation residual error of 0.01 ± 0.01, while α tunes the scanning device standard deviation (3mm for GS100 and 1.5mm for GS200). We estimate the optimal translation vector $\hat{\mathbf{t}}$ by double thresholding the residual errors, r_k. We first discard outliers in order to compute the total translation $\tilde{\mathbf{t}}$:

$$\tilde{\mathbf{t}} = \mathbf{C}_{1,\xi} - \hat{\mathbf{R}}^{-1}\mathbf{C}_{2,\xi} \tag{4.15}$$

where $\mathbf{C}_{b,\xi}$, $b = 1, 2$ denotes the barycenters coordinates after the first outlier rejection using the rotation model. We compensate the total translation on scan S_2,

$$\bar{\mathbf{p}}_2^k = \mathbf{p}_2^k - \tilde{\mathbf{t}} \tag{4.16}$$

and we discard a second time the residual errors between the rotationally aligned scans. In practice, up to 15% of 3D correspondences are rejected through the double-thresholding process. The translation vector is computed using the difference of the rotationally aligned centroids corresponding to the remaining 3D points matches:

$$\hat{\mathbf{t}} = \mathbf{C}_{1,\xi\xi} - \hat{\mathbf{R}}^{-1}\bar{\mathbf{C}}_{2,\xi\xi} \tag{4.17}$$

The total residual error with respect to the global model is expressed by:

$$\mathbf{Q}_{[\hat{\mathbf{R}},\hat{\mathbf{t}}]} = \sum_{k=0}^{k=N_{12}-1} \Phi_{k,\xi\xi} \| (\mathbf{p}_1^k - \hat{\mathbf{R}}^{-1}\bar{\mathbf{p}}_2^k - \hat{\mathbf{t}} \| \tag{4.18}$$

4.5.5 Pyramidal Matching Strategy and Incremental Pose Refinement

Additionally to the use of calibration constraints for reducing the solution space of $\mathbf{T}$, the algorithm performs pose estimation in a pyramidal searching fashion to cut down the combinatory.

After computing intensity and depth 2D panoramic images for each scan, we build pyramidal structures for both, the intensity and depth panoramic images of S_n noted I_n^l and D_n^l, respectively, where $l = 0, .., L_{max}$ and L_{max} is the height of the pyramid [Bouguet, 2000]. We build 2D panoramic images for 3D Cartesian coordinates noted I_{xyz}, and generate their corresponding pyramidal structures I_{xyz}^l in order to have direct access to the true 3D coordinates at each level, when projecting back in the 3D space each matched pixel. The rotation computation starts at the lowest resolution level L_{max}, where an exhaustive searching is performed in order to provide a coarse value of the global minimum. For $l = L_{max}$ the $\mathcal{P}_{SA}^{L_{max}}$ parameters are initialized using Equation (4.19):

$$\mathcal{P}_{SA}^{L_{max}} = \begin{cases} [\psi_{min}^{L_{max}}, \psi_{max}^{L_{max}}] = [0°, 180°] \\ [\theta_{\mathbf{n},min}^{L_{max}}, \theta_{\mathbf{n},max}^{L_{max}}] = [0°, 180°] \\ \delta\psi^{L_{max}} = \frac{\delta\theta_{\mathbf{n}}^{L_{max}}}{2} = 1° \end{cases} \tag{4.19}$$

Let $(\hat{\psi}, \hat{\theta}_{\mathbf{n}}, \varphi_E)^{L_{max}}$ be the coarse global minimum computed at the lowest resolution level L_{max}. The coarse rotation estimation is refined at higher resolution levels $l = L_{max}-1, .., 0$, increasing the accuracy and reducing the searching space around the global minimum $(\hat{\psi}, \hat{\theta}_{\mathbf{n}}, \varphi_E)^l$ computed at each previous level. The solution space for each level is delimited by $\Delta\psi^l$ and $\Delta\theta_{\mathbf{n}}^l$ which are defined as the corresponding range values for ψ^l and $\theta_{\mathbf{n}}^l$, respectively. For each new level $l = L_{max} - 1, .., 0$, the $\mathcal{P}_{SA}$ parameters are updated following the recursive scheme described in Equation (4.20):

$$\mathcal{P}_{SA}^l \begin{cases} \Delta\psi^l = \frac{\Delta\psi^{l+1}}{2} \\ \Delta\theta_{\mathbf{n}}^l = \frac{\Delta\theta_{\mathbf{n}}^{l+1}}{2} \\ [\psi_{min}^l, \psi_{max}^l] = [\hat{\psi}^{l+1} - \Delta\psi^l, \hat{\psi}^{l+1} + \Delta\psi^l] \\ [\theta_{\mathbf{n},min}^l, \theta_{\mathbf{n},max}^l] = [\hat{\theta}_{\mathbf{n}}^{l+1} - \Delta\theta_{\mathbf{n}}^l, \hat{\theta}_{\mathbf{n}}^{l+1} + \Delta\theta_{\mathbf{n}}^l] \\ \delta\psi^l = \frac{\Delta\psi^l}{4} \\ \delta\theta_{\mathbf{n}}^l = \frac{\Delta\theta_{\mathbf{n}}^l}{4} \end{cases} \tag{4.20}$$

For $l = L_{max}$ an exhaustive searching has been processed in a $(\psi, \theta_{\mathbf{n}})$ window of size 180° by 180°. For $l = L_{max} - 1$ we choose a $\mathcal{P}_{SA}^{L_{max}-1}$ of size 2° by 2° around the global minimum $(\hat{\psi}, \hat{\theta}_{\mathbf{n}})$ found at L_{max}, (choosing $\Delta\psi^{L_{max}-1} = \Delta\theta_{\mathbf{n}}^{L_{max}-1} = 1°$ with $\Delta\theta_{\mathbf{n}}^{L_{max}-1}$ corresponding to 1 pixel in the 2D panoramic image space), which is sampled in a 9 by 9 pixel window.

Since in our 3D mosaicing scenario translations are negligible, i.e. less than $\frac{1}{4}$ pixel at levels $l > 0$, we can speed up the matching process by introducing and computing the translation vector directly at the highest resolution level, $l = 0$. Accordingly to the height of the pyramidal structure given by $L_{max} = 3$, the algorithm can handle maximal translation values of $2^{Lmax} = 8$ pixels corresponding to 10 cm.

4.6 Pair-wise Rigid Scans Alignment Experiments

Data input. We applied the 3D mosaicing scenario described in Section 4.3 in two prehistoric caves from France: Moulin de Languenay - trial 1 and Tautavel - trials 2, 3 and 4. Each trial is composed by sequence of 4-VPBs acquired nearly from the same 3D position. In order to evaluate the robustness of the proposed method wrt different scanning devices and different scans resolutions, we performed several tests on data acquired with different acquisition setups.

Moulin de Languenay - trial 1: time and in-situ access constraints were not noticed and therefore the Trimble® GS100 laser was set to deliver multi-shot and high resolution scans.

Tautavel - trials 2, 3, 4: the experiments were run in a large-scale and "difficult-to-access" underground site. Therefore, the acquisition setup was designed to handle large-scale scenes while dealing with time and in-situ constraints. In particular, Trimble® GS200 was employed to supply accurate measurements at long ranges. In addition, during experiments we focused to limit as much as possible the acquisition time by setting the sensing device to acquire one-shot and low resolution scans, emphasizing the robustness of our algorithm with respect to sparse large scale data sets caused by depth discontinuities.

Figures 4.8 and 4.9 depict the pair-wise scan matching process for trial 1 and trial 2, respectively, using both modes of the rotation estimation, i.e. intensity and depth. The basic inputs are directly exploited in order to automatically solve for the intrinsic parameters of the spherical acquisition $\mathcal{P}_{\mathcal{I}}$, which are used to generate intensity and depth 2D spherical panoramic views (step described in Section 4.5.1) and their associated 3-level pyramidal structures corresponding to each scan. In the panoramic image space 1° corresponds to $4, 2, 1, 0.5$ pixels at level $l = 0, .., 3$, respectively. We illustrate the dissimilarity maps (ψ-row, $\theta_{\mathbf{n}}$-column) for each level l and the superposed images obtained for the optimal rotation, i.e. I_1 and I_2 warped in I_1's space for the intensity mode, and D_1 and D_2 warped in D_1's space, for the depth mode. The dissimilarity maps for level L_{max} represent a full search, i.e. $(\psi \in [0^\circ, 360^\circ])$ and $(\theta \in [0^\circ, 360^\circ])$, in order to emphasize the score maps symmetry. Nevertheless, in practice for rapidity purposes, only the half of the space is explored. The warping is performed by resampling the image with subpixel accuracy and computing image brightness via bilinear interpolation. The maximal accuracy is obtained at the highest resolution level $l = 0$, where steps of $(\delta\psi^0, \delta\theta^0_{\mathbf{n}}) = (0.0625^\circ, 0.0625^\circ)$ were used to explore the SA solution space, where $\delta\theta^0_{\mathbf{n}}$ corresponds to $\frac{1}{4}$ pixel. The final outputs of the pair-wise scan matcher are given by the highest resolution level, $l = 0$: the pose $\mathbf{T}[\hat{\mathbf{R}}, \hat{\mathbf{t}}]$, the overlap $\mathbf{O}[\hat{\mathbf{R}}, \hat{\mathbf{t}}]$, the dissimilarity score $\mathbf{E}[\hat{\mathbf{R}}, \hat{\mathbf{t}}]$ and the global criterion $\mathbf{Q}_{[\hat{\mathbf{R}}, \hat{\mathbf{t}}]}$. Figure B.1 from Appendix B.2 summarizes the pair-wise scan matcher processing pipeline.

Figure B.2 from Appendix B.2 and Figure 4.10 illustrate the aligned scans obtained by running the pair-wise scan matcher on one pair of partially overlapping scans belonging to trials 1 and 2, using depth and intensity modes, respectively. Table 4.6 shows the numerical results associated to Figures B.2 and 4.10.

Operating mode influence. When analyzing the results illustrated in Table 4.6 for trial 2 with respect to the mode employed for the rotation estimation (intensity or depth), we observe that for large motions resulting in low overlap values the laser rotation angle estimates are ambiguous. These pair-wise uncertainties are eliminated by the optimal absolute poses computed during the multi-view fine alignment process for which a description is provided in the next section.

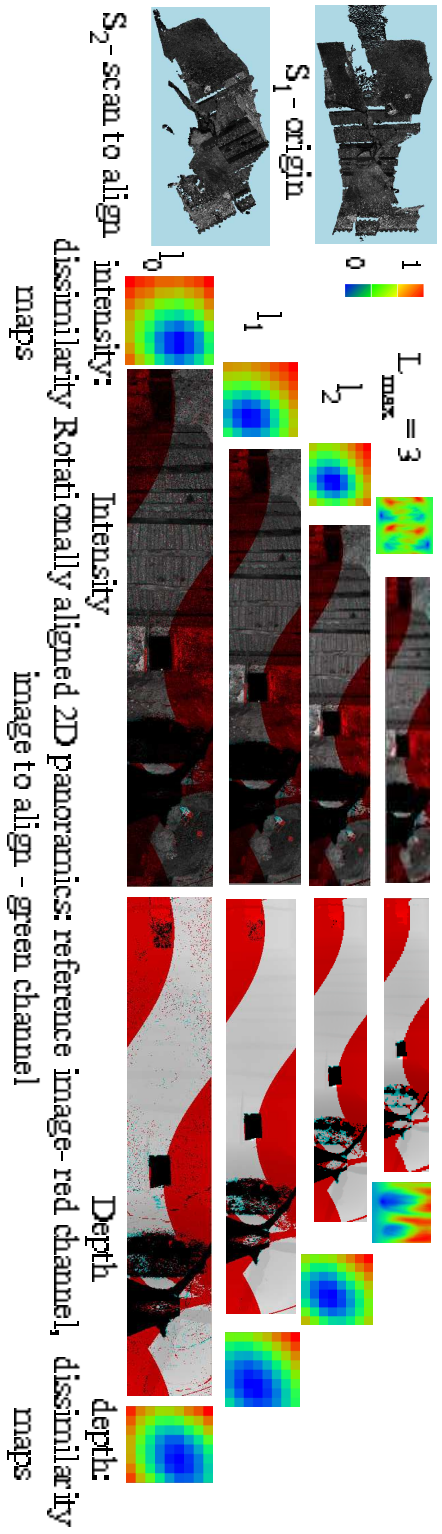


Figure 4.8: Pair-wise scan matching procedure on data sets acquired in Moulin de Languenay prehistoric cave (France) using a Trimble^{circledR} GS100 scanning device. Image size I_1^0: $n_u^0 = 1502$, $n_v^0 = 252$.

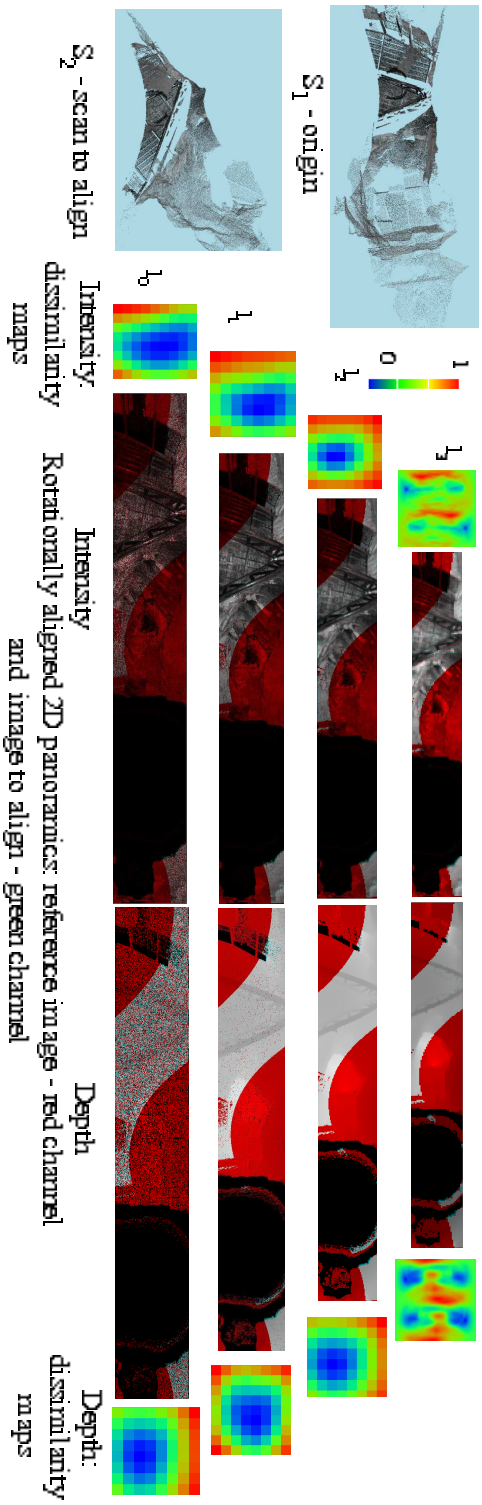


Figure 4.9: Pair-wise scan matching procedure on data sets acquired in Tautavel prehistoric cave (France) using a Trimble® GS200 scanning device. Image size I_1^0: $n_u^0 = 2083$, $n_v^0 = 310$.

Overlap influence. Comparing to trial 2, for trial 1 a large overlap region was provided. By having a closer look at Table 4.6 we observe that the size of the provided overlap

Trial	Mode	$\hat{\psi}(^\circ)$	$\mathbf{n}(\hat{u}_{\theta_\mathbf{n}}, \hat{v}_{\varphi_\mathbf{n}})$ (pixels)	$\|\hat{\mathbf{t}}\|$ (mm)	**O**	**E**
Trial 1	Intensity	33.875	$[483.1875, 158.6604]^T$	2.59	0.51	0.074
M-shot	Depth	33.875	$[483.1875, 158.6604]^T$	2.19	0.51	0.003
Trial 2	Intensity	47.875	$[521.75, 196.0377]^T$	0.85	0.19	0.054
1-shot	Depth	48	$[521.75, 196.0377]^T$	0.88	0.19	0.116

Table 4.1: The outputs of the automatic pair-wise scan matching procedure for Trial 1 and Trial 2.

influence the robustness of the rotation estimates. However, the translation estimates are still ambiguous, varying with an order of 10^{-3}m even when high overlaps are provided. This is explained by the fact that the accuracy of the algorithm is subject to the accuracy of the measurement delivered by the sensing device, which is our case is α $\alpha = 3$mm for $GS100$ - trial 1 and 1.5mm for $GS200$ - trial 2.

The multi-view alignment process described in the next section is designed to compute the optimal absolute poses which consists in minimizing and compensating eventual misregistration errors encountered during the pair-wise process.

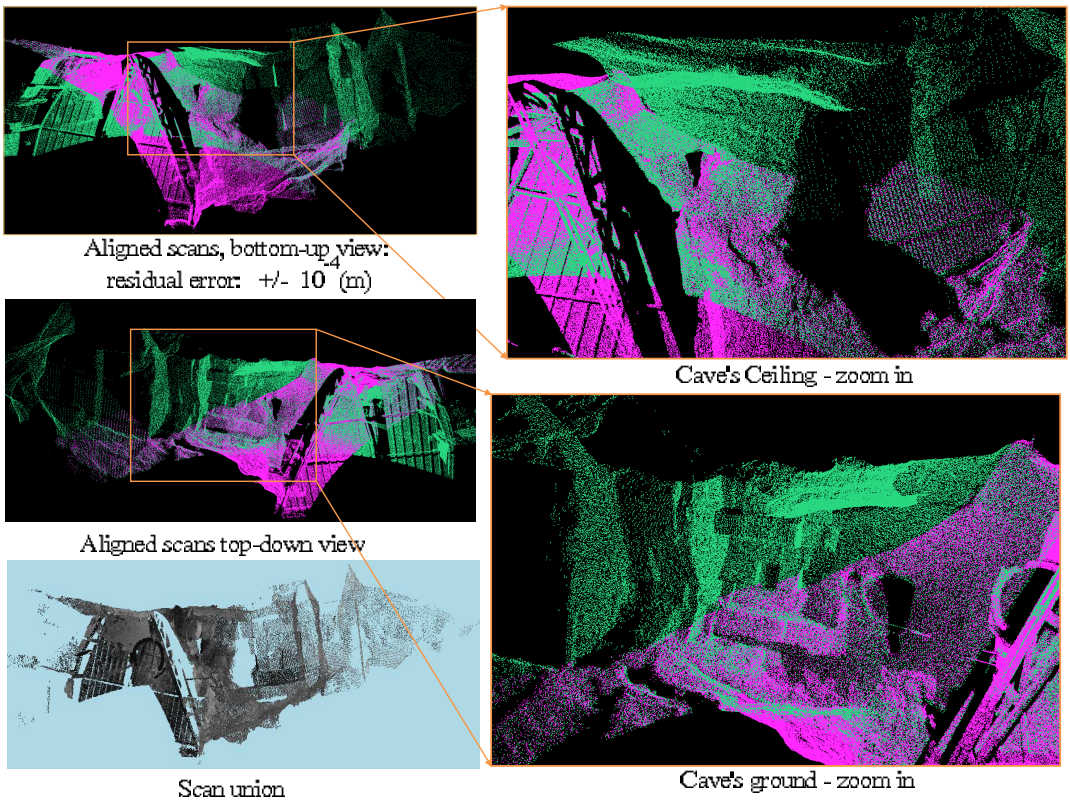


Figure 4.10: Pair-wise matching results on data sets acquired in Tautavel prehistoric cave (France). Operating mode: depth, total number of points: 1.312×10^6, runtime: 7 min 32 s on a 1.66 GHz Linux machine equipped with 2 Gb of RAM memory.

4.7 Multi-view Scan Matching via Topological Inference

This section focuses on the multi-view alignment process corresponding to phases (3) and (4) in Figure 4.5. The basic input is a sequence of N overlapping scans and no additional information. For an arbitrary sequence, no knowledge is given about the reference scan neither about the ordering between scans within the sequence.

Being given the symmetric relationship between the direct and indirect pair-wise poses, i.e. $\mathbf{T}_{ij} = \mathbf{T}_{ji}^{-1}$, we save computation time by running the pyramidal pair-wise scan matcher on pairs of scans (S_i, S_j) defined on $\Omega = \{(S_i, S_j)|i, j \in \{0, .., N-1\}, i \neq j, i < j\}$, i.e. C_N^2 pairs.

The pair-wise scan matcher outputs are used to assign a model graph introduced by Huber, [Huber, 2002], defined as $\mathcal{G} = \{\mathcal{N}, \mathcal{A}, \mathcal{B}, \mathcal{E}\}$ which encodes the topological relationship between all views within a sequence. $\mathcal{G}$ contains a node $\mathcal{N}_i$ for each view V_i associated to S_i. The attribute $\mathcal{A}_i$ for node $\mathcal{N}_i$ denotes the absolute pose $\mathbf{T}_i$ for V_i. $\mathcal{B}_{ij}$ includes the relative pose $\mathbf{T}_{ij}$, the overlap $\mathbf{O}_{ij}$, the dissimilarity score $\mathbf{E}_{ij}$, and the edge $\mathcal{E}_{ij}$, indicating that V_i transformed by $\mathbf{T}_{ij}$ overlaps V_j. We assign to each edge $\mathcal{E}_{ij}$ a weight, noted e_{ij}, which provides information about the scans' *adjacency*.

We define two topological relations over the graph $\mathcal{G}$. First, *a scans adjacency* notion is associated to the 3D mosaic acquisition scenario and second, *the pose consistency* is verified via a topological inference criterion.

Scans adjacency. In our proposed 3D mosaicing acquisition scenario, the *scans' adjacency* is defined as the apparentness of an arbitrary scan S_i to the currently processed 3D mosaic. Accordingly to the acquisition scenario and taking an extreme case when rotations larger than ψ_{max} are encountered, the adjacent scans composing the mosaic are always related by a minimum overlap situated at the top (ceiling) and bottom (ground) of the spherical mosaic which in theory correspond to an overlap of $\cong$ 30% of a VPB. In practice, the system's montage leads to an area which is not digitized, situated right underneath the scanner, as shown in Figure 4.6, which corresponds to the south pole of the spherical view, reducing therefore the global overlap to the half. Nevertheless, the missing data does not affect the quality of the alignment thanks to the dense correlation procedure which yields robustness to low overlap areas.

Therefore, we estimate that the minimum overlap which relates two VPB belonging to the same mosaic must be superior to $\mathbf{O}_{VBP}^{min} = 15\%$. Consequently, an inferior value to $\mathbf{O}_{VBP}^{min} = 15\%$ would signify that the tested scan S_i does not belong to the currently processed 3D mosaic, being an integrating part of a different 4-VBP sequence acquired from a different spatial 3D position of the system. This defines the scans adjacency notion associated to the 3D mosaicing acquisition scenario which is encoded within the graph $\mathcal{G}$ using the edges e_{ij} as follows:

$$e_{ij} = \begin{cases} 1, & \text{if } S_i \cap S_j \geq \mathbf{O}_{VBP}^{min} \\ 0, & \text{if } S_i \cap S_j < \mathbf{O}_{VBP}^{min} \end{cases} \tag{4.21}$$

Since no knowledge is given about the scan's adjacency, the weights values e_{ij} corresponding to each edge $\mathcal{E}_{ij}$ are initialized to 1.

Topological inference criterion [Sawhney et al., 1998]. As stated in [Huber, 2002], a model graph is said to be *pose consistent* if the relative pose between any two views is independent on the path used in calculation. When dealing with pair-wise scan matching it is very often observed that the estimated poses are locally consistent but

globally they are not. In order to ensure the consistency of the global 3D scene model, the multi-view scan matching process verifies the pair-wise matches' consistency via the topological inference (TI) procedure introduced by [Sawhney et al., 1998]. The TI process verifies the relative poses $\mathbf{T}_{ij}$ by composing the relative poses of adjacent views along the path:

$$\mathbf{T}_{ij}^{TI} = \mathbf{T}_{i,i+1} \circ ... \circ \mathbf{T}_{k,k+1} \circ ... \circ \mathbf{T}_{j-1,j} \tag{4.22}$$

where, $e_{k,k+1} = 1$.

Up to now we have formalized two relations defined on the graph $\mathcal{G}$: the *scans' adjacency* and the *topological inference criterion*, which are employed to perform the global alignment within two steps. First, the algorithm detects whether the input scan sequence contains *alien scans*, i.e. scans which do not belong to the currently processed 3D mosaic. Second, the pair-wise poses are used to detect the best reference scan which will optimally register all the adjacent scans in a global reference coordinate system. The multi-view scans matching procedure ends by integrating all views into a single 3D entity, hence providing a complete 4π steradians field of view for a given station. Figure B.3 from Appendix B.3 illustrates the global pipeline of the multi-view scans alignment process. The following two subsections focus on the description of the two aforementioned steps of the multi-view alignment process.

4.7.1 Alien Scans' Detection

The sensing device is set to acquire $N_{scenario}$ scans from a single 3D position in space. However, during the mission, multiple scans acquired from various view-points may be mixed up and incorrectly assigned to the currently processed 3D mosaic, being by default linked through edges $e_{ij} = 1$. Detecting scans which do not belong to the currently processed 3D mosaic requires to verify the adjacency condition expressed in Equation (4.21) and update the edge values into $e_{ij} = 0$ when the adjacency criterion is not fulfilled.

Alien scans' detection is performed within two steps. First, the pair-wise procedures outputs are exploited by a voting procedure to detect scans susceptible of being incorrectly assigned to the currently processed 3D mosaic. The currently processed 3D mosaic is given by the maximum VPBs fulfilling the adjacency condition expressed in Equation (4.21). As stated in the acquisition scenario, $N_{scenario} = 4$ VPBs are generally acquired to form a complete spherical mosaic. Second, a consistency test is performed on scans for which the susceptibility assumption was sustained by a maximum number of votes in order to decide whether the candidates may be classified as alien scans or not.

In our acquisition scenario $N_{scenario}$ is known. Nevertheless, we generalize our approach, providing two solutions, taking into account whether $N_{scenario}$ is provided or not.

Case (a) $N_{scenario}$ known. The algorithm performs alien scan detection if the input sequence contains more scans than the acquisition scenario needs, i.e. when $N > N_{scenario}$.

Each scan belonging to the scan sequence is tested in order to detect those which are susceptible of being incorrectly assigned to the current sequence. Let $S_k, k \in \{0, .., N_{sequence} - 1\}$ be a scan for which we want to find whether it is susceptible to be an incorrectly assigned one or not. For each scan, $S_i, i \in \{0, .., N_{sequence} - 1 | i \neq k\}$ the algorithm performs a voting procedure using the dissimilarity measure.

The dissimilarity scores computed wrt S_k, $\mathbf{E}_{ik}$, are compared against those obtained between the same scan, S_i and all the other scans, $S_j, j \in \{0, .., N_{sequence} - 1 | j \neq i, j \neq k\}$, $\mathbf{E}_{ij}$. For each scan, S_i the algorithm votes for the scan S_k susceptibility assumption as follows:

$$v_{i,k,j} = \begin{cases} 1, & \text{if } \mathbf{E}_{ik} > \mathbf{E}_{ij} \\ 0, & \text{if } \mathbf{E}_{ik} \leq \mathbf{E}_{ij} \end{cases} \tag{4.23}$$

Being given that the current sequence must contain a number of $N_{scenario}$ scans, for each scan S_i at least $N_{scenario} - 1$ votes must sustain the susceptibility assumption in order to declare S_k susceptible of being an incorrectly assigned scan. More precisely, the susceptibility assumption must be sustained in unanimity by scans $S_i, i \in \{0, .., N_{scenario} - 1 | i \neq k\}$. For an arbitrary scan S_i the total number of votes is given by $v_{ik} = \sum_j v_{i,k,j}$ and the susceptibility condition is expressed as:

$$v_{ik} \geq N_{scenario} - 1 \tag{4.24}$$

If for at least one scan S_i the above susceptibility condition is fulfilled, then the scan S_k is a candidate for the pose consistency test performed in order to decide whether the scan is incorrectly assigned or not.

Since the pair-wise scans matching procedure outputs the dissimilarity score and the overlap, we can express the susceptibility condition by making use of the minimum overlap value imposed by the acquisition scenario O^{min}_{VPB}.

$$\text{if } \exists S_i \text{ such as } v_{ik} \geq N_{scenario} - 1 \Rightarrow S_k \in \mathcal{C}_{TI} = \{S_k | S_k \cap S_i < O^{min}_{VPB}, k \neq i\} \tag{4.25}$$

The topological inference procedure is applied on the graph $\mathcal{G}$ in order to verify the consistency of the estimated pair-wise poses $\mathbf{T}_{ik}, i \neq k$ corresponding to each candidate scan $S_k \in \mathcal{C}_{TI}$.

Being given that all views within the graph $\mathcal{G}$ are initially linked through edges $e_{ij} = 1$, there are several paths possible for composing the pose $\mathbf{T}^{TI}_{ik}$. Following the topological inference criterion, if for at least one arbitrary path $\mathbf{T}^{TI}_{ik}$ the pose $\mathbf{T}_{ik}$ is found inconsistent, then the scan S_k is detected as an *alien scan* and consequently is discarded from the multi-view alignment procedure by updating the corresponding edge values e_{ik}. Using the topological criterion from Equation (4.22), the discarding condition can be formalized as follows:

$$\text{if } \exists \mathbf{T}^{TI}_{ik} \notin \{\mathbf{T}_{ik} \pm \xi^{TI}\} \Rightarrow e_{ik} = 0 \tag{4.26}$$

where $\xi^{TI} = |\bar{\mathbf{r}} + \sigma_{\mathbf{r}}|$ encoding the tolerance error attached to the pair-wise scan matching procedure.

Case (b) $N_{scenario}$ **unknown**. If the number $N_{scenario}$ is not known, the algorithm employs the median value of the dissimilarity score in order to detect susceptible alien scans. For each scan S_i, we compute the median score $\bar{\mathbf{E}}_i$ over all the dissimilarity measures obtained by matching pairs of scans $(S_i, S_j), j \in \{0, .., N-1 | i < j, \}$, $\mathbf{E}_{ij}$, which is defined by the following expression:

$$\bar{\mathbf{E}}_i = med_{j \in \{0,..,N-1 | i<j\}} \{\mathbf{E}_{ij}\} \tag{4.27}$$

For each scan S_i the algorithm performs the voting procedure as follows:

$$v_{ij} = \begin{cases} 1, & \text{if } \mathbf{E}_{ij} > \bar{\mathbf{E}}_i \\ 0, & \text{if } \mathbf{E}_{ij} \leq \bar{\mathbf{E}}_i \end{cases} \tag{4.28}$$

For each scan S_j the total number of votes is given by:

$$v_i = \sum_{j \in \{0,..,N-1 | i<j,\}} v_{ij} \tag{4.29}$$

The scans S_k whose susceptibility assumption is sustained by a maximal number of votes are considered candidates for the consistency test. The susceptibility condition form Equation (4.24) becomes:

$$S_k = \arg \max_{i \in \{0,..,N-1\}} v_i \tag{4.30}$$

Analogue to case (**a**), the pose consistency $\mathbf{T}_{ik}$ is verified for each scans S_k via the TI procedure. If at least one pose is found inconsistent, then the algorithm decides that S_k is an *alien scan* and consequently discards it from the multi-view scans alignment process.

Discarding alien scans. Let K be the total number of the discarded scans for both cases (a) and (b). Each detected alien scan S_k is discarded from the multi-view alignment process by updating the weights e_{ik} corresponding to each edge $\mathcal{E}_{ik}$ relating the scans S_k to all other scans S_i within the graph $\mathcal{G}$ as follows:

$$\mathcal{E}_{ik} \leftarrow e_{ik} = 0 \tag{4.31}$$

Discarded scans $S_k, k \in \{0,..,K-1\}$ with the same number of votes v_m are grouped into clusters $\mathcal{C}_m$:

$$\mathcal{C}_m = \{S_k, k = 0,..,M-1 | v_k = v_m\} \tag{4.32}$$

where $card(\mathcal{C}_m) = M < K < N$ and $K = \sum_m card(\mathcal{C}_m)$.

The alien scan detection procedure is applied iteratively to each cluster $\mathcal{C}_m$ until the topological inference procedure results in a completely pose consistent graph.

The multi-view scans alignment process goes on by exploiting the updated graph $\mathcal{G}'$ to compute the optimal absolute poses which register all the remaining $N-K$ adjacent scans in a common reference coordinate system to produce a 3D mosaic.

4.7.2 Find Optimal Absolute Poses

Absolute poses are required in order to register all scans wrt a global coordinate system. Since one view may overlap several others, it is necessary to detect which one will optimally register all scans wrt a global coordinate system.

Since a high overlap privileges the computation of highly accurate absolute poses, it is undoubtable that the optimal absolute poses maximize implicitly the global overlap while minimizing the dissimilarity score over the entire 3D scene model.

Generally, the first scan is chosen as the reference so that the global coordinate system is locked in the reference frame of that scan. However, this approach requires a specific acquisition scenario to provide a maximum overlap between the first scan and all the other views in order to produce accurate absolute poses estimates.

We propose a 3D mosaicing acquisition scenario which allows to freely rotate the scan laser, leading to imprecise overlapping areas. Since high overlap values ensures accurate pose estimates, the multi-view scan alignment process detects the absolute poses which will optimally register all views wrt the optimal reference scan, noted $\hat{S}_0$.

As shown in Table 4.6, very large motions (i.e. superior to the maximal rotation angle ψ_{max} imposed by the acquisition scenario in order to guarantee the minimum overlap O_{VPB}^{min}) corresponds to low overlap values, resulting in quite noisy pose estimates. In order

to handle such critical cases, the global multi-view scans alignment process compensates the misregistration errors of the pair-wise matching step by registering all scans wrt the one which maximizes the global overlap over the entire sphere between all scans, while minimizing the dissimilarity between them.

The basic idea based on which the algorithm searches for the optimal reference scan, $\hat{S}_0$ uses the relationship between the rotation angle ψ_{ij}, the overlap $\mathbf{O}_{ij}$ and the dissimilarity score $\mathbf{E}_{ij}$. Since high overlap values correspond to low dissimilarity scores, the optimal reference scan $\hat{S}_0$ maximizes the global overlapping area between all the registered scans, minimizing thereby the global dissimilarity score between all views. Therefore, the optimal reference scan is computed by maximizing the global overlap surface over the entire 3D scene model.

For each potential reference scan S_i, the absolute pose $\mathbf{T}_j$ which will register the view $V_j, j \in \{0, .., N-2 | j \neq i\}$ with respect to the absolute view V_i, is given by $\mathbf{T}_j = \mathbf{T}_{ji} = \mathbf{T}_{ij}^{-1}$, where $\mathbf{T}_{ij}$ has been already computed by the pair-wise matching procedure and its corresponding edge weight fulfills the adjacency condition, i.e. $e_{ij} = 1$.

The global overlap $G_O(i)$ is obtained by registering all scans S_j with respect to S_i, and summing all the overlaps $\mathbf{O}_j$ corresponding each absolute pose $\mathbf{T}_j$. The quantity $G_O(i)$ is evaluated with respect to the entire 3D scene model surface, which in our case is a fully covered spherical surface, noted G_s.

Since the overlap $\mathbf{O}_j$ is expressed with respect to the number of image points belonging to the 2D panoramic image associated to reference scan S_i, we compute first the overlap produced by the alignment process in pixel units, $\mathbf{O}_{n-view}(i)$ as follows:

$$\mathbf{O}_{n-view}(i) = n_u(i) n_v(i) \sum_{j=0, i \neq j}^{j=N-2} \mathbf{O}_j \tag{4.33}$$

where, $n_u(i)$ and $n_v(i)$ denote the rectangular support of the 2D panoramic image.

Using the angular steps corresponding to one pixel, $(\delta\theta_s, \delta\varphi_s)$, we express the global spherical surface in pixel units:

$$G_s = \frac{2\pi}{\delta\theta_s} . \frac{\varphi_s^{min} - \varphi_s^{max}}{\delta\varphi_s} \tag{4.34}$$

where, φ_s^{min} and φ_s^{max} denote the limits of the vertical field of view of the obtained 3D mosaic. The angular steps corresponding to one pixel are given by the median value over the angular steps of the VPBs composing the sequence, being expressed as:

$$(\delta\theta_s, \delta\varphi_s) = (med_{\{i \in 0,..,N-1\}}, \{\delta\theta(i)\}, med_{\{i \in 0,..,N-1\}}\{\delta\varphi(i)\}) \tag{4.35}$$

Therefore, the global overlap $G_O(i)$ may be expressed with respect to the sphere's surface as follows:

$$G_O(i) = \frac{\mathbf{O}_{n-view}(i)}{G_s} \tag{4.36}$$

The optimal reference scan is given by the maximum global overlap obtained over the entire 3D scene model.

$$\hat{S}_0 = \arg \max_{i \in \{0,..,N-1\}} G_O(i) \tag{4.37}$$

4.8 Experiments and Quality Assessment

This section presents the results obtained for trials 1 - 4. The algorithm starts by running the pair-wise scan matching procedure on each scan sequence, which outputs the pose estimation between each pair defined on Ω.

Alien scans detection. The multi-view global alignment process detects whether the sequence contains incorrectly assigned scans or not. Figures 4.11, 4.12 and 4.13 illustrates the case (**a**) of the alien scans' detection process described in Section 4.7.1 for a sequence containing $N_{scenario} = 6$ scans from which 2 scans were found as incorrectly assigned and consequently they were discarded from the scan sequence. Figure 4.11 shows that scans S_5 and S_6 are susceptible to be incorrectly assigned. Therefore, in order to decide if scan S_5 and S_6 are incorrectly assigned, the algorithm tests the pose consistency of their corresponding pair-wise poses via the TI procedure. Figures 4.12 and 4.13 illustrate the consistency tests using the laser rotation angle estimates $\hat{\psi}$. For both scans, the algorithm detects more than one inconsistent pair-wise rotation angle. Thus, the two candidates S_5 and S_6 were detected as incorrectly assigned and consequently, they are discarded from the multi-view scans alignment process.

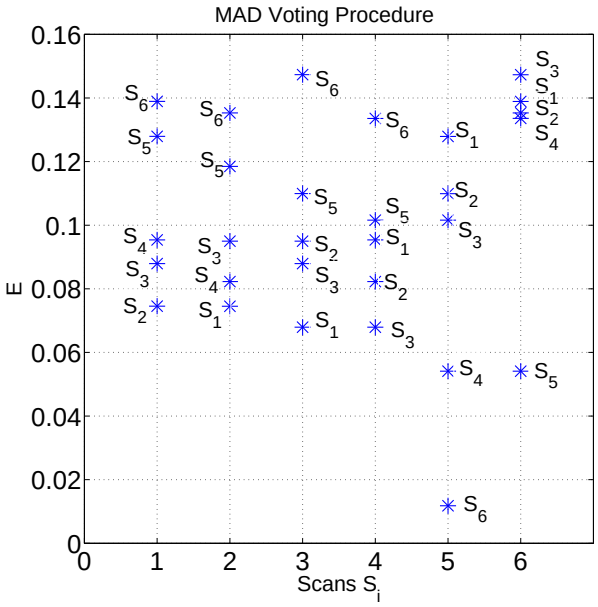


Figure 4.11: MAD voting procedure using the intensity mode. We observe that for scan S_1, $\mathbf{E}_{15} > \mathbf{E}_{12}$, $\mathbf{E}_{15} > \mathbf{E}_{13}$, $\mathbf{E}_{15} > \mathbf{E}_{14}$, which yield $v_{15} = 3$, meaning that the susceptibility condition in Equation (4.24) is fulfilled. A similar situation is obtained for S_2 and S_3, fulfilling also the susceptibility condition. However, one scan S_i suffice in order to declare the scan S_5 susceptible to be an incorrectly assigned scan. The same result is obtained also for scan S_6.

Compute optimal absolute poses. The remaining pair-wise poses are used to compute the optimal reference scan $\hat{S}_0$. Since each view V_i is a potential reference, the multi-view scan matching process evaluates the global overlap over the entire 3D scene model obtained by considering each view V_i as a reference. The optimal reference scan $\hat{S}_0$ is given by the absolutes poses which maximize the global overlap over the entire 3D scene model.

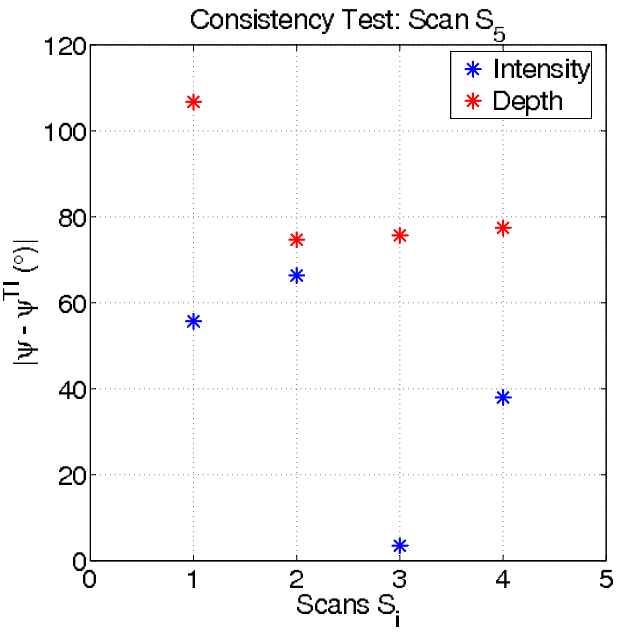


Figure 4.12: Consistency Test for Scan S_5 using the topological inference procedure. Each estimated rotation angle $\psi_{i5}, i \in \{1, .., 4\}$ does not coincide with the angle computed via TI procedure ψ_{i5}^{TI}. This yields the inconsistency of all relatives poses between scans S_i and S_5.

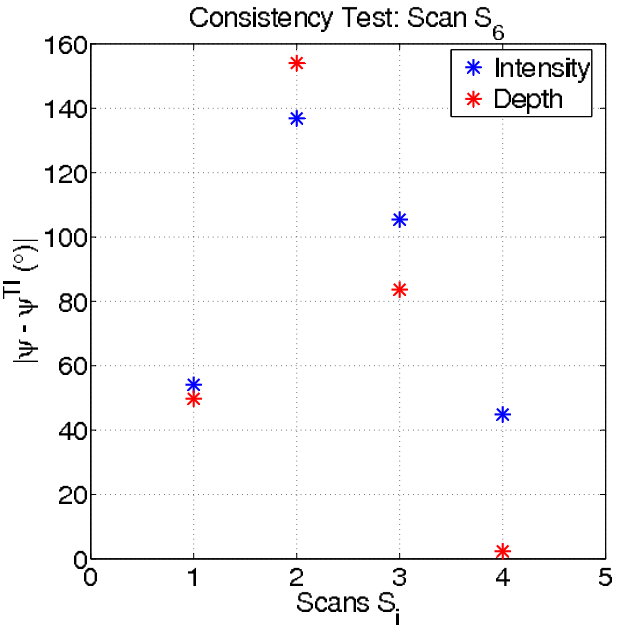


Figure 4.13: Consistency Test for Scan S_6 using the topological inference procedure. Each estimated rotation angle $\psi_{i6}, i \in \{1, .., 4\}$ does not coincide with the angle computed via TI procedure ψ_{i6}^{TI}. This yields the inconsistency of all relatives poses between scans S_i and S_6.

Figure 4.14 depicts the global overlap evaluation with respect to each potential reference scan for trials 1-4. Referring to trials 1, 2 and 3, we observe that for both modes, intensity

and depth, the global overlap over the entire 3D scene model is maximized by registering all views with respect to scan S_1. The multi-view scans alignment procedure ends by registering all views with respect to the reference scan $\hat{S}_0 = S_1$. We observe that for trial 4 different optimal reference scans were obtained with respect to each mode: $\hat{S}_0 = S_2$ for intensity mode and $\hat{S}_0 = S_1$ for depth mode. This is explained by the presence of missing data, which results in a low global overlap value over the entire sphere's surface ($G_O = 0.06$ %). Trial 4 illustrates an extreme case for which our method yields reliable alignment results in presence of critical low overlapping scans. More precisely, the laser was situated in a narrow area, close to the wall's cave. This experiment illustrates the robustness of the proposed method to large-scale data sets.

Figure 4.14 allows us to establish a repeatability rate with respect to each mode. When using the depth mode, the same scan S_1 was found as the one maximizing the global overlap for all trials (1-4), $\hat{S}_0 = S_1$, yielding an unitary repeatability rate, while for the intensity mode a repeatability rate of 0.75 is obtained.

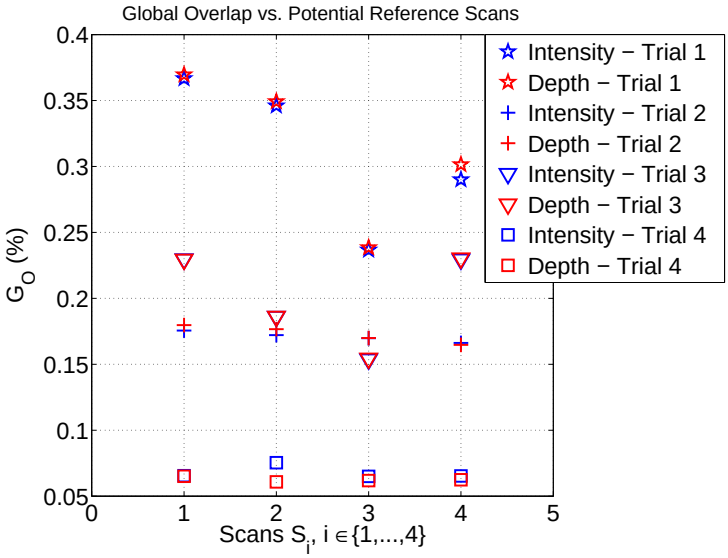


Figure 4.14: Global overlap evaluation with respect to the potential reference scans for trials 1 - 4.

Figures 4.15, 4.16, 4.17 and 4.18 depict the residual mean error and the corresponding standard deviation obtained for each pair-wise match, using both modes of the pair-wise scan matching procedure for all trials.

For trials 2, 3, and 4 the experiments were run in a very difficult-to-access underground environment, in presence of holes and sharp depth changes resulting in missing data. Despite data sparseness, more accurate results were obtained for trials 2, 3, and 4, comparing to trial 1. This is explained by the sensing device capacity: comparing to trial 1, for trials 2, 3 and 4 the number of points provided by the scanning device (GS200) is higher (see figures 4.8 and 4.9).

Figure B.4 from Appendix B.3 and Figures 4.19 and 4.20 illustrate the rendering results for trials 1 an 2, obtained by passing each 4-VPBs sequence to the automatic intensity-based multi-view scan matcher.

Table 4.2 provides the global residual errors obtained for all trials. When analyzing the residual mean errors, we observe the inter-dependency between the alignment accuracy

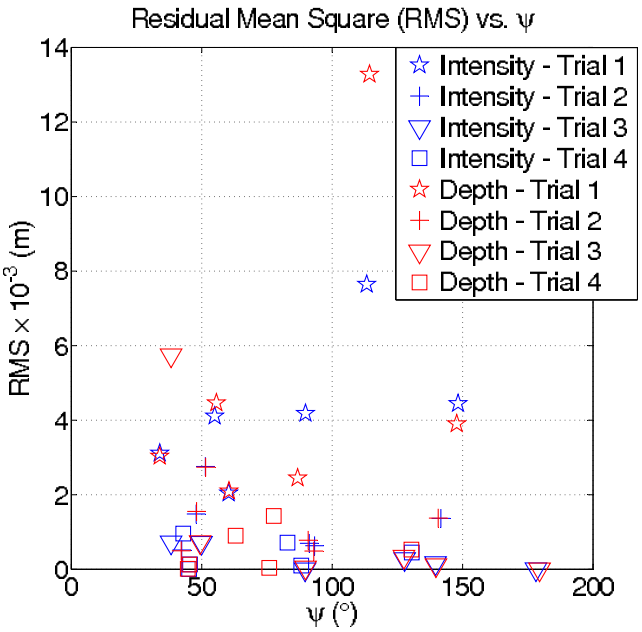


Figure 4.15: Residual Mean Square vs. Rotation Angle for trials 1 - 4.

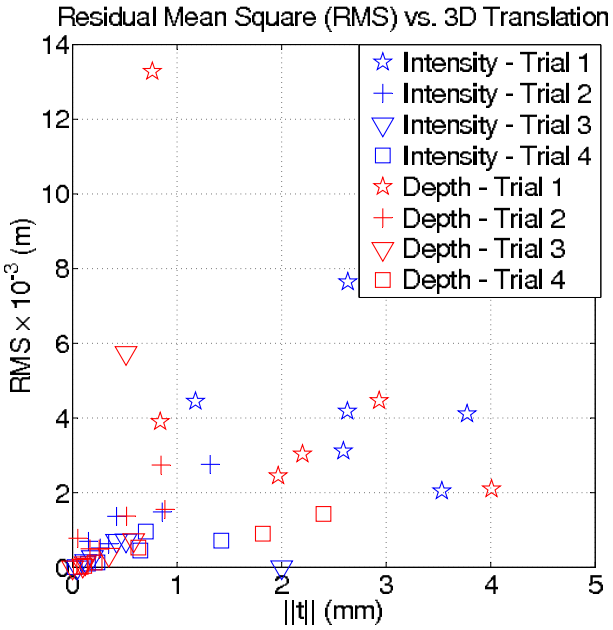


Figure 4.16: Residual Mean Square vs. 3D Translation for trials 1 - 4.

and the number of points provided by the capturing device for pose calculation. The experiments demonstrates the robustness and the reliability of our algorithm in complex environments where depth discontinuities lead to large scale sparse data sets. The fourth column of table 4.2 illustrates that following the scan matcher mode, the results' accuracy may vary between $[10^{-2}, 10^{-3}]$.

Operating mode selection. The accuracy between the two modes does not vary significantly. Consequently, when both data are available, i.e. intensity and depth, their jointly use provides robustness, completing each other drawbacks (depth mode is robust to

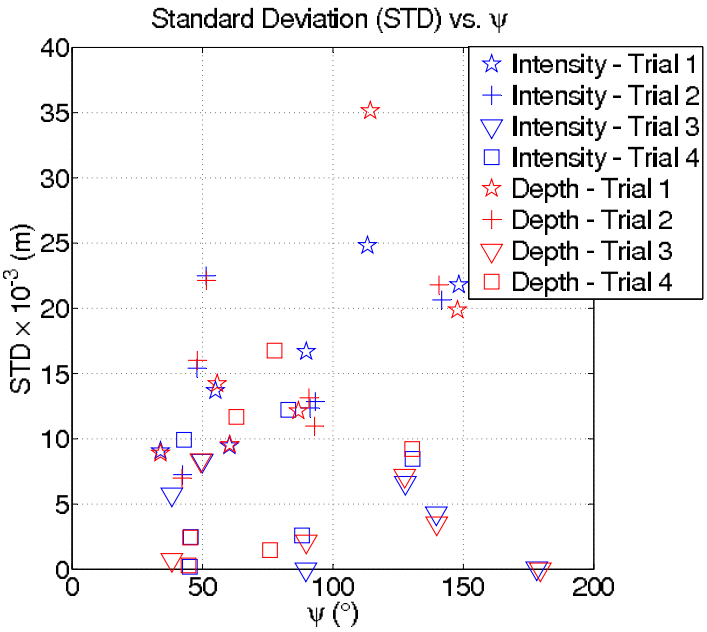


Figure 4.17: Standard Deviation vs. Rotation Angle for trials 1 - 4.

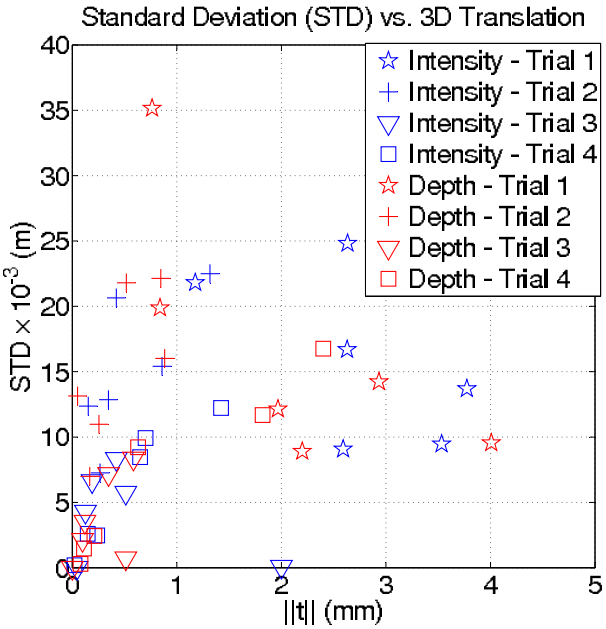


Figure 4.18: Standard Deviation vs. 3D Translation for trials 1 - 4.

poorly textured areas, while the intensity mode guarantees accurate results in geometrically symmetric man-made environments). When computation time must be reduced, one may choose between intensity and depth mode using two criterions.

- *for previously unknown environments, the scan matching output needed by the processing pipeline*: for instance, in our case, the intensity mode is used in order to recover the reflectance 2D spherical projection of the generated 3D mosaic which is exploited for further processing in Chapter 6 to align a texture map onto the 3D

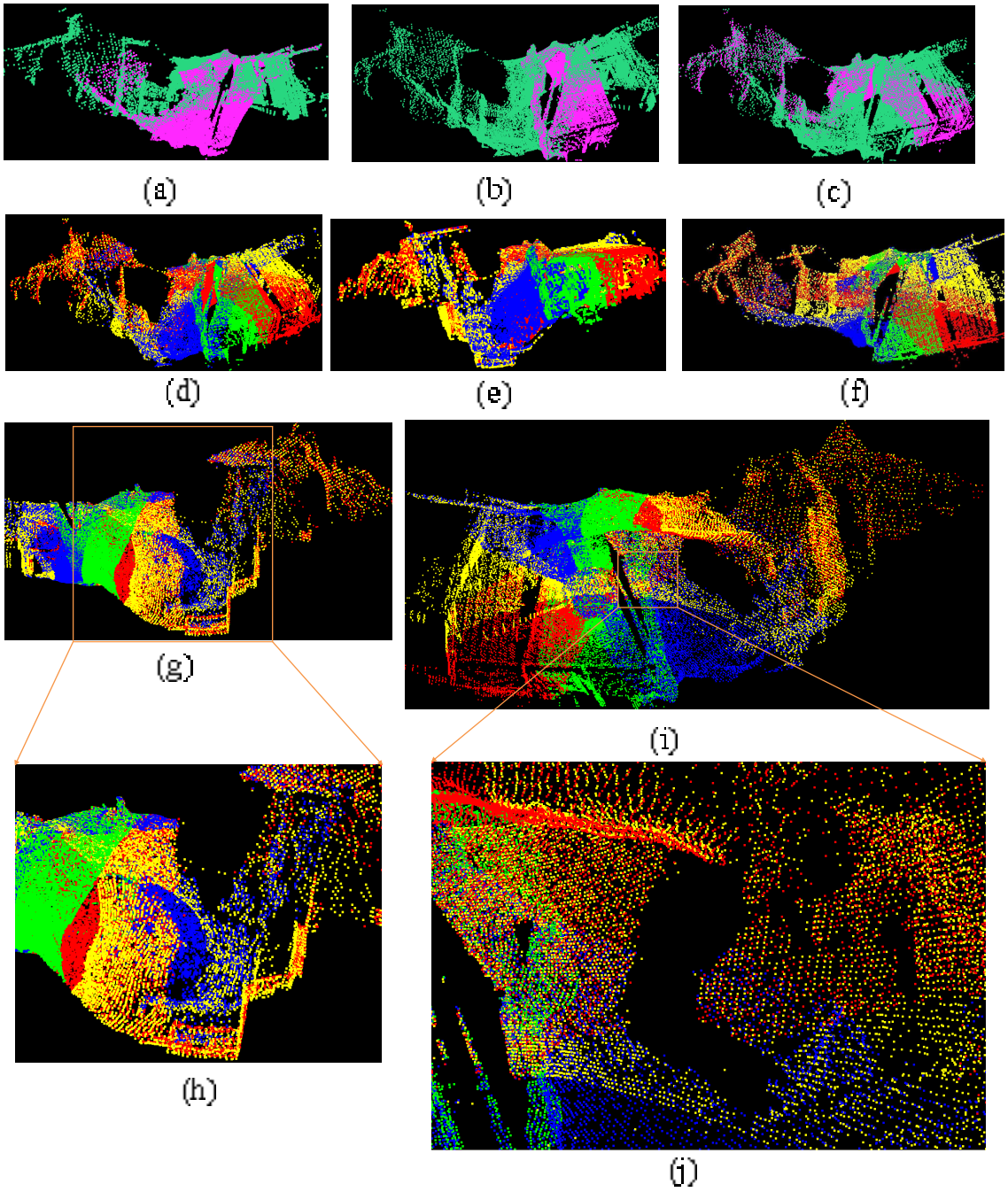


Figure 4.19: Multiview Scan Matching results on data sets acquired in Tautavel prehistoric cave, France - Trial 2. (a) S_1 - green, S_2 - magenta, (b) S_{12} - green, S_3 - magenta, (c) S_{123} - green, S_4 - magenta, (d) Multiview scan alignment - Top-down view, S_1 - yellow, S_2 - blue, S_3 - green, S_4 - red, (e) Front-left view, (f) Top view, (g) Front-right view, (h) Zoom-in outdoor front-right view, (i) Bottom-up view, (j) Zoom-in cave's ceiling.

mosaic using radiometric criterions for generating 4D-mosaics. Figure 4.21 illustrates the 2D spherical projection containing the reflectance obtained for trial 2.

- *environment type:* when knowledge about the environment is available, the algorithm

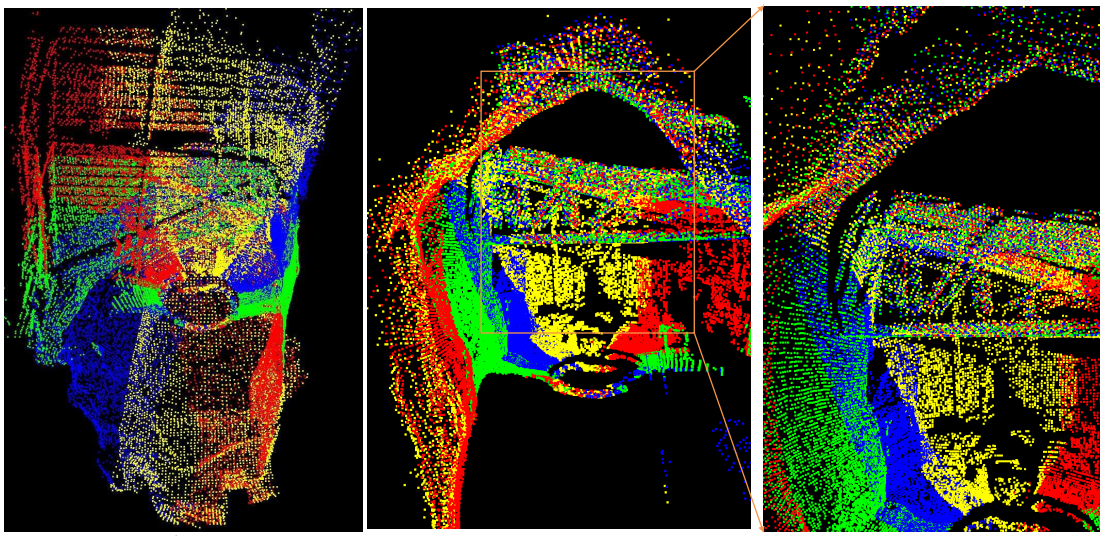


Figure 4.20: Multiview Scan Matching results on data sets acquired in the Tautavel prehistoric cave, France - Trial 2. (a) Cave's outdoor view, (b) Cave's indoor, (c) Zoom-in cave's indoor.

allows us to select the most suitable scan matching mode. For instance, the depth mode provides robustness in complex environments in presence of either too textured or too homogeneous areas. Although the proposed method does not rely or feature extraction and matching, it is suitable for structured environments as well. When applying the proposed method in structured man-made areas, with geometrically symmetric areas, such as rooms, the intensity mode yields more robust results.

Normal operating limits. We study the algorithm's operating limits for application in previously unknown environments, which implies either structured or unstructured for scenarios performed either indoor or outdoor. Two main factors may influence the scans alignment quality or even provoke the algorithm's failure. The first one is *the size of the overlapping region* which has the main impact on the pose computation process, and the second is the *inter-scans parallax* amount introduced by the system's instability during rotations.

When applied in underground or ceiling-covered environments, there are no limits imposed for the rotation of the sensing device. The acquisition scenario guarantees the minimum overlapping value $\mathbf{O}_{VBP}^{min}$ situated on the north pole of the spherical mosaic. Nevertheless, special attention must be given when applying the 3D mosaicing scenario in outdoors environments, since they lead to a non-digitized north pole of the spherical mosaic. This problem can be solved by rotating the L-mount-laser with smaller rotation values that ψ_{max}. A second issue which must be taken care of is the inter-VPBs parallax introduced by the system's instability during rotations. The pyramidal framework enables the algorithm to handle translations values up to 10 cm. In order to allow accurate poses estimation, it is necessary to fix the platform to avoid superior parallax values.

Runtime. The experiments were run on a 1.66 GHz Linux machine using a standard CPU implementation. The last column of Table 4.2 shows that the proposed approach exhibits robustness to registration errors with a reasonable computation time. Nevertheless, since the algorithm was originally designed in a multi-tasking fashion, it allows for both

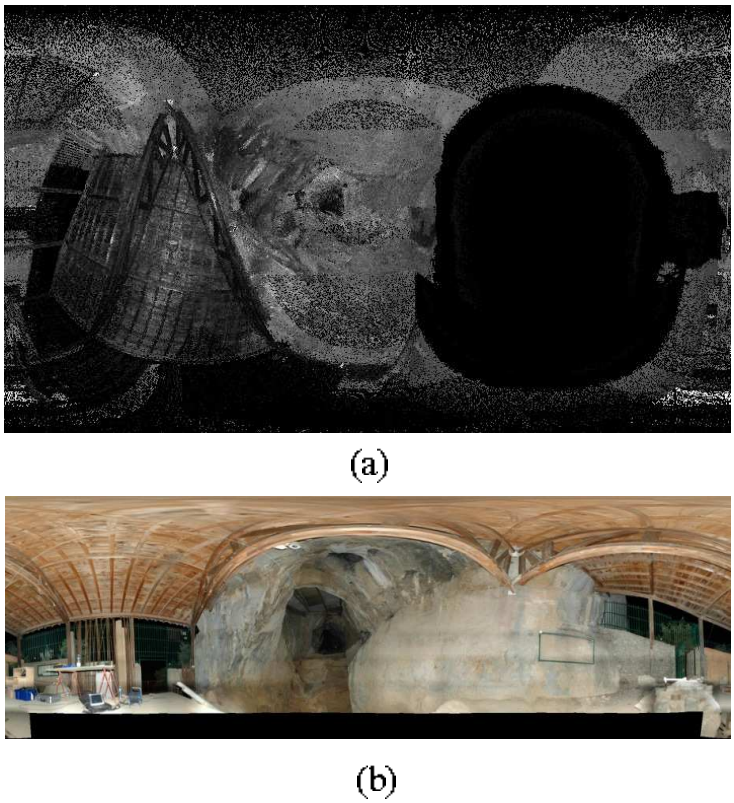


Figure 4.21: The 2D spherical projection of the global 3D mosaic obtained for trial 2. (a) The 2D reflectance mosaic: the resulted global mosaic contains a non-digitized area corresponding to the area situated underneath the scanner. (b) The 2D color mosaic generated from the same 3D pose of the system to facilitate the visualization of the 2D reflectance mosaic. The two mosaics (a) and (b) are separated by a 3D rotation and a negligible translation.

sequential and parallel processing on embedded platforms. In order to improve the speed of the algorithm, the next section provides the embedded design for parallel implementation on a multi-core embedded platform.

4.9 Embedded Design for onboard 3D Mosaicing

When porting an algorithm on an embedded system several aspects must be taken care of such as: which hardware fits the needs of the application, how fast the algorithm can be processed and how many power resources are required. Therefore, feasible schemes stand in the design of specific integrated solutions for a wisely chosen hardware accordingly to the application type.

This section proposes the embedded design for the multi-view scans alignment algorithm described in this chapter for performing onboard 3D mosaicing. In order to allow real time 3D mosaicing, the embedding design occurs at two levels: data acquisition and processing. Therefore, we present the embedded design for parallel processing and a hardware solution to allow real time acquisition.

Parallel processing. By having a closer look at Figures B.1 and B.3 from Appendices B.2 and B.3 respectively, one can notice that the 3D mosaicing process can be decomposed

Trial	Mode	$(\bar{\mathbf{r}} \pm \sigma_{\bar{\mathbf{r}}}) \times 10^{-2}$	$(\Delta\bar{\mathbf{r}} \pm \Delta\sigma_{\bar{\mathbf{r}}}) \times 10^{-2}$	♯points, CPU time (min)
Trial 1	Intensity	3.913 ± 15.86	0.793 ± 2.22	1.508×10^6
GS100	Depth	3.12 ± 13.64		16.44
Trial 2	Intensity	1.18 ± 16.14	1.94 ± 0.84	2.5829×10^6
GS200	Depth	3.12 ± 16.98		27.39
Trial 3	Intensity	0.332 ± 4.15	0.021 ± 0.154	2.6079×10^6
GS200	Depth	0.353 ± 4.304		27.66
Trial 4	Intensity	0.184 ± 1.249	0.007 ± 0.884	2.5321×10^6
GS200	Depth	0.191 ± 0.365		26.28

Table 4.2: Results of the global 3D scene models. The fourth column illustrates that the accuracy may vary following the mode used an order of 10^{-2}. of the pose estimates wrt the mode used. The last column illustrates the number of points and runtime obtained for each trial.

into smaller tasks that can be performed in parallel. We can divide the algorithm in the following main parts:

1. **Pre-processing:**
   - build 2D panoramic images of intensity/depth and 3D cartesian coordinates from VPB;
   - generate the corresponding L_{max}-level pyramidal structures;
2. **Core**: compute relative poses $\mathbf{T}_{ij}$ between scans;
3. **Post-processing:**
   - alien scans' detection;
   - optimal absolute poses computation;

The core represents the most computationally expensive part within the entire multi-view scan alignment process. Within this stage the pair-wise scan matching step is performed between all scans in order to provide the relative pose estimates. Moreover, this is a repetitive and independent task which can be processed simultaneously on a multi-core embedded platform.

We designed a multi-tasking software architecture suitable for parallel onboard processing on a multi-core embedded platform, capable to cope with the laser's architecture and to meet the embedded platform requirements.

The acquisition time for one single VPB containing 6×10^5 points (6.7Mb) is 15 min and 4 VPBs are at least needed to provide a fully 3D spherical mosaic. In order to minimize the execution time of the entire in-situ process (i.e. data acquisition and processing), data processing is performed on-the-fly, starting as soon as the each scan's acquisition is finished.

Figure 4.22 illustrates the algorithm's parallelization for an embedded platform equipped with 3 CPUs. After acquiring each scan, its pre-processing starts simultaneously with the next scan acquisition. As soon as the first scan pair is acquired, their relative pose computation is performed simultaneously with the third scan acquisition. This multi-tasking process continues until the last scan is acquired. Then, all its associated pair-wise poses are computed in parallel. The multi-view alignment process ends up with the post-processing

stage. Following the above principle and by parallelizing the algorithm on an embedded platform equipped with 3 cores, a factor of 3 can be gained over the entire runtime for the entire 3D mosaicing process. As shown in Table 4.3, once the scans' acquisition is finished, the system generates *in-situ* a 3D mosaic after 5 min - for GS100, and 9 min - for GS200 scanning device.

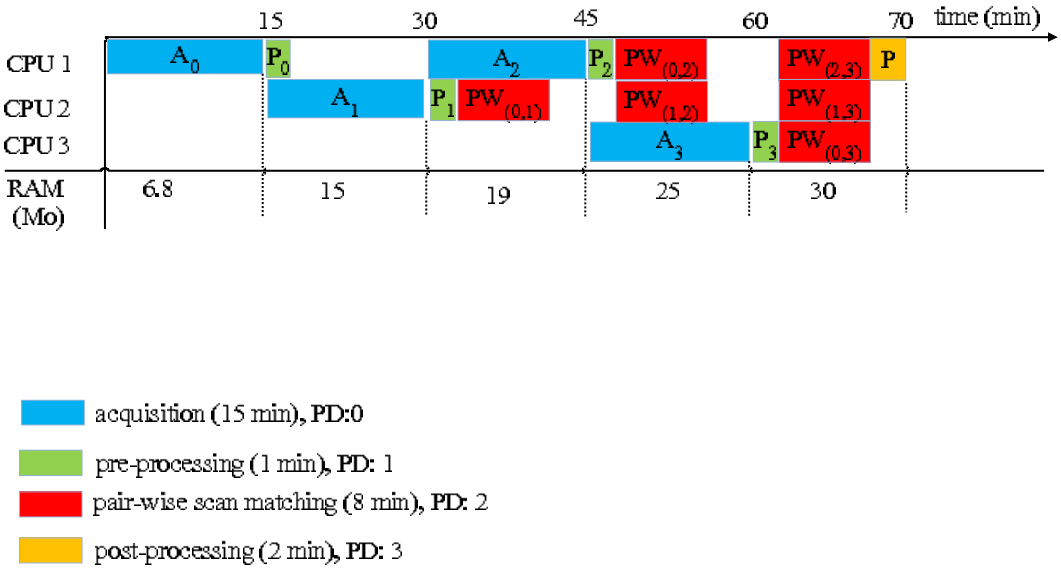


Figure 4.22: Parallelization on 3 CPUs.

runtime (min)	1 CPU	3 CPU
GS100	16	5
GS200	27	9

Table 4.3: CPU runtimes for sequential and parallel processing on two capturing devices from Trimble®.

Since the pre-processing stage waits for each scan acquisition (15 min), there are no speed constraints nor memory bandwidth limitations for data transmission or memory access. During the core and the post-processing stages all the 4 VPBs are exploited and a maximum amount of 35 Mb memory is needed. Therefore, the proposed algorithm can run on multi-core embedded platform with low amount of onboard memory. The power consumption of the scanning device reaches 150 W, while the embedded computer may require less than 65 W.

Improving runtime using silicon devices. With the new advances of Digital Signal Processing (DSP) and Field Programmable Gate Array (FPGA) researchers have reported implementations of commonly used computer vision algorithms - such as feature extraction and digital image warping [Giacon et al., 2005], [Baumgartner et al., 2007]. The recently developed cutting edge DSP provides enough performance when optimized computer vision algorithms are used. When choosing between DSP and FPGA, one has to pay attention to the execution mode of the algorithm, i.e. parallel or sequential. FPGA are suited for algorithms which benefit from parallel execution, which is the case in our research work. Therefore, more processing time could be gained by implementing the proposed algorithm on FPGA.

Real-time 3D mosaicing using 3D cameras. Although the multi-core implementation of the algorithm decreases the computational time for the processing stage, there is still room to improve the runtime of the entire 3D mosaicing process by reducing the acquisition time. A possible way to speed up the data acquisition step is to employ a real-time 3D camera, such as the one proposed by MESA®, mounted on a pan-tilt motorized platform instead of the L-mount-laser prototype. The 3D camera delivers a sequence of pose-annotated range images which can be further stitched into a 3D mosaic using the 2D Gigapixel mosaicing algorithm proposed in next chapter. Although the algorithm was designed and tested in 2D high-resolution color images, it is straight forward to apply it on 2D depth images.

Uses of the 3D mosaic output. The output of the proposed 3D mosaicing system can be exploited either *in-situ* or by a host wirelessly connected to the target.

- **in-situ:** in our research work, the 3D mosaicing process is integrated within the mosaic-driven 3D modeling process performed in-situ by the ARTVISYS system introduced in Chapter 3. The use of the 3D mosaic output occurs at two levels of the entire 3D modeling system. The first level is the main 3D modeling process. This is described further in Chapter 6 which illustrates how the 3D mosaic is exploited to generate in-situ fully spherical and photorealistically textured 3D models encoded as *4D-mosaics*. The second level occurs when looping to provide feedback control to the system by exploiting the 3D mosaic within visual servoing procedures. As we will see further in Chapter 7, at this level the 3D mosaic can be used to encode the geometric information into a volumetric global map which is further exploited to extrapolate semantic information about the system's surroundings, giving rise to *in-situ* perception needed to enable vision-based systems at performing *in-situ* site modeling and exploration.

- **off-line:** the acquired data together with the absolute poses can be transmitted wirelessly to a host, where the final 3D model rendering can be visualized and improved by an operator within in a controlled-laboratory rendering pipeline [Levoy et al., 2000], [Ikeuchi et al., 2007].

The aforementioned embedded design gives rise to a real-time 3D mosaicing sensor suitable to be integrated onboard unmanned mobile platforms for supplying site surveys missions in high-risk and difficult to access environments.

4.10 Conclusions

This chapter prototypes a 3D mosaicing sensor capable to supply automatically 3D modeling tasks in complex and difficult to access environments without human operator intervention. The contributions of this chapter range from hardware to software levels, along with theoretical and algorithmic concepts focused to solve for the automation of the 3D modeling pipeline which is actually the main scope of this dissertation. We list hereafter the main contributions of the research work presented in this chapter.

Hardware. We introduce a hardware architecture which enables the sensor to capture a fully spherical field of view of the system's surroundings for a single 3D pose. The capturing device delivers a sequence of partially overlapped scans acquired through imprecise pan rotations which are further aligned and merged automatically by the multi-view scans alignment method proposed in this chapter.

3D mosaicing scenario. An unsolved issue standing behind the automation of the 3D modeling process requires a data acquisition scenario capable to provide automatically the minimum overlap required in order to supply reliable and fast in-situ data matching. To deal with this aspect, the proposed scans alignment technique comes together with a 3D mosaicing acquisition scenario which provides automatically the minimum overlap required to ensure a reliable and fast data matching, while avoiding data redundancy. Furthermore, the 3D mosaicing scenario allows for an occlusion-free 3D modeling process by acquiring fully spherical mosaics and partially mosaics to sense the occluded areas, avoiding therefore data redundancy. Beside its stand alone use, the proposed acquisition scenario is an integrating part of the mosaic-driven acquisition scenario performed in-situ by ARTVISYS in order to achieve the 3D scene model completeness.

Theoretical and algorithmic contributions. Although prototyping the 3D mosaicing sensor is the global contribution of this chapter, we believe that the its essential contribution stands in the development of a rigid multi-view scans alignment algorithm capable to generate in-situ 3D mosaics in large-scale, complex and difficult to access environments without requiring operator's intervention. This is of main importance, since the proposed algorithm solves for one of the major issues standing behind the automation of the 3D modeling pipeline: the *automatic data alignment* problem. We summarize and justify hereafter several theoretical and algorithmic features integrated within the proposed scans alignment framework.

- **Global approach.** We propose a rigid multi-view scans alignment algorithm for generating in-situ 3D mosaics in large-scale and difficult to access environments. The basic input of our framework is a sequence of partially overlapped and unordered scans, without other additional information. The multi-view alignment framework falls in the category of *sequential* methods being powered by a pair-wise alignment technique which produces *precise poses estimates*. This overcomes the drawback of the *simultaneous* techniques which are generally followed by a global registration phase to distribute the accumulated errors resulted from the pair-wise coarse estimates. Moreover, the global registration step leads to a complexity with grows quadratically with the number of views, being unfeasible for large-scale environments. One may argue that the alien scans detection process introduced in this chapter may lead to a quadratic growth with the number of scans, too. This is not truth since in out case study we know that a 3D mosaic requires $N_{mosaic} = 4$ scans and therefore, the total number of scans acquired $N_{scenario}$ and integrated with the multi-view process can be limited by the user to $N_{scenario} = 10$ for instance.

- **Free-initial guess pair-wise alignment for precise rigid estimates.** Being given the existing pair-wise alignment approaches, feature-based and direct, for safety reasons, we choose a direct approach and save computation time using calibration constraints to reduce the solution space and by performing a pyramidal searching to cut down the combinatory. As for the core of the pair-wise scans alignment, it is powered by a matching procedure of 2D panoramic views using a pyramidal dense correlation framework via quaternions. Since the acquisition scenario ensures a minimum overlapping area in order to avoid data redundancy, the dense correlation procedure allows to exploit all the information presented in the overlap region. We designed a hybrid framework using a combination of radiometric and geometric measures. As a matter of fact, we solve for pose estimation within two steps: first, the rotation in computed in the 2D panoramic image space, while translation estimation

is performed using the rotationally aligned panoramic images back-projected in the 3D space. The 2D panoramic image provides spatial and appearance constraints, increasing therefore the reliability of the scan matching task. In addition, the pyramidal framework takes care of the two key aspects of any scan matcher: the accuracy and the robustness, allowing to compute large motions at low resolution levels and fine motions at the highest pyramidal level.

- **Multi-view rigid alignment.** We define a graph model over the entire scan sequence which exploits the pair-wise procedure outputs to encode the topological relationship between views. We associate a *scans adjacency* notion to the 3D acquisition scenario which is used jointly with a *topological inference criterion* [Sawhney et al., 1998] to discard scans which do not belong to the currently processed 3D mosaic, introduced as *alien scans*. In addition, since one view may overlap several others, the detection of the best reference view is seen as a dissimilarity minimization over the entire graph. Therefore, the optimal absolute poses are obtained by minimizing the dissimilarity score, thereby maximizing the global overlap over the entire 3D scene model.

Onboard capabilities:

- We designed a general multi-view scans alignment technique capable to receive different types of inputs: range images or 3D point clouds, and to process different type of data: intensity and depth, being able to automatically switch between the two modes for pose computation. Therefore, the proposed framework provides robustness wrt the scanning device, tacking into account whether the intensity is acquired or not.

- We provide a multi-tasking implementation of the algorithm and demonstrate its portability on either single or multi-core embedded platforms for on line 3D scan matching onboard mobile platforms. This software is likely be integrated onboard unmanned mobile platforms to provide them with 3D scene representation for supplying autonomy and decisional resources.

Experiments and quality assessment. We demonstrated the reliability of our method by automatically generating 3D mosaics in two challenging underground prehistoric caves situated in France. A quality assessment is addressed by illustrating and evaluating the results obtained wrt different scanning devices, using both modes of the scan matcher: intensity and depth. The presented experiments illustrate the robustness of the proposed method with respect to the available type of input available (range images or 3D point clouds), type of data to be processed (intensity or depth), feature-less areas and sparse large-scale data sets (caused by depth discontinuities which are inherent to complex and large-scale environments, such as natural underground sceneries).

Solved key issues. The proposed method solves explicitly for several key issues stated in the beginning of this chapter in Section 4.2.3, which need to be addressed when performing in-situ 3D modeling tasks in unstructured and underground environments, such as:

- the absence of detectable and trackable features via dense correlation methods;

- the non-reliability of navigation sensors;

- time and in-situ constraints, i.e. fast acquisition and processing.

While focusing to solve for the automation of the data alignment task, we have addressed the aforementioned issues by integrating the following features:

- a data acquisition scenario which guarantees the minimum overlap required to achieve accurate pose estimates, avoiding therefore data redundancy;
- since features' existence cannot be guaranteed in previously unknown environments, for safety reasons we employ a direct scans' alignment method for computing *precise rigid estimates* without requiring pre-alignment;
- a sequential multi-view scans alignment to avoid the high complexity of the global registration step employed by simultaneous multi-view alignment methods. Moreover, we reduce the number of views to be acquired to minimum 4 in order to generate a fully spherical field of view surrounding the system;
- in order to allow for in-situ processing, calibration constraints were used jointly with a pyramidal searching strategy to reduce the solution space and to cut down the combinatory.

Features and uses of the proposed 3D mosaicing system. Figure 4.23 shows the integration of the 3D mosaicing bloc within the ARTVISYS system and illustrates a top-down-zoom view of its composing processing blocs. On the left side of the figure are emphasized the features of each composing bloc wrt the currently existing scans alignment methods, while on the right side are stated their corresponding stand alone and integrated uses.

Perspectives. In [Craciun et al., 2008] we have investigated the possibility of introducing a *matching quality measure* for the pair-wise scan alignment by exploiting the inter-dependency between the dissimilarity scores and the overlap region. This is willing to reduce the computational time of the alien scans' detection process. An ongoing work is to verify whether the quality match is verified when extreme cases occur within a wide range of scenarios.

Since the nowadays 3D scanning devices are not equipped with an onboard multi-view scan matching technique, but with interactive post-processing software based on manual calibration, artificial landmarks or navigation sensors, they do not have enough capacity to perform autonomously data acquisition and processing to allow *in situ* verification and/or visualization of the 3D scene model completeness.

The need of such an algorithm is emphasized by the fact that during the post-processing step the operator observes that the final 3D scene model is incomplete. Therefore, without an in-situ pre-visualization of acquired data, the need to come back on site to complete data collection is unavoidable. Moreover, the algorithm proposed in this chapter represents one of the main processing blocs required by the vision-based 3D modeling systems currently under development aimed to be deployed in hostile environments for supplying site surveys missions [Nüchter et al., 2004], [Magnusson and Duckett, 2007], [Johnson et al., 2007], [Rekleitis et al., 2009]. Such systems aim to solve for the automation of the 3D modeling pipeline which is further exploited to power the system's autonomy based on visual perception.

We see the contribution of this chapter at two levels. From a system-oriented point of view, the main contribution of this chapter stands in the development of a 3D mosaicing sensor prototype embedding an automatic multi-view scans alignment algorithm capable

to cope with currently existing 3D scanning devices, to process data on line and to generate in situ a 3D mosaic. The presented software was designed to be embedded onboard 3D laser scanners to perform, guide and assist site survey missions in difficult-to-access environments. The second level refers to the automation of the 3D modeling pipeline it-self by introducing a fully automatic scans alignment algorithm without relying on navigation sensors nor nor feature matching, providing therefore an environment-independent solution.

We believe that the proposed approach has an interest of its own, since it can reliably generate *in-situ* fully 3D spherical mosaics, which nowadays represents an increasing need for multiple purposes requiring an accurate embedded 3D scene representation, such as 3D modeling, autonomous navigation, SLAM and path planning, being capable of providing active 3D vision in high-risk environments.

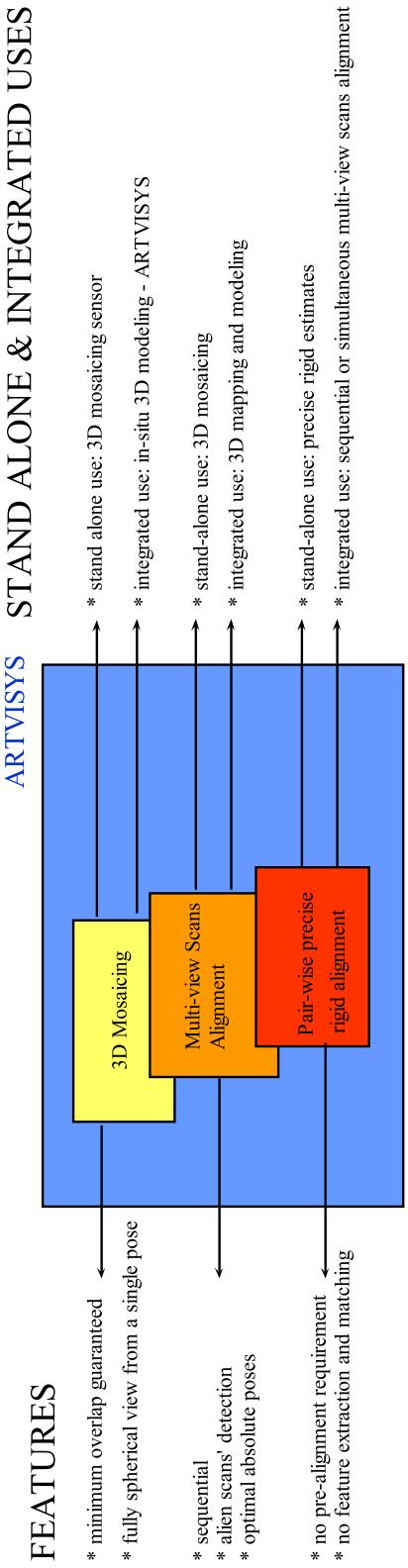


Figure 4.23: Global and local overview of the 3D mosaicing process.

Chapter 5

AGM: Automatic Gigapixel Mosaicing from Nodal Optical Images

This chapter addresses the multi-view image alignment problem for in-situ automatic generation of Giga-pixel spherical mosaics in complex and underground environments. An short version of the research work presented in this chapter can be found in [Craciun et al., 2009]. Beside the stand-alone use of the proposed Giga-pixel mosaicing algorithm, it constitutes the second main processing bloc of the ARTVISYS system introduced in Chapter 3.

The image mosaicing spreadness has its origin in the Digital Age, being increasingly supported by what one would call today ubiquitous digital computing or Information Era. We start this chapter by introducing a bit of history on image mosaicing, drawing its fast upgrade and wide applicability to which the last two decades of the Digital Age lead.

The next section presents an inside view of the mosaicing process and the available techniques composing it, pointing out their complementarity. Section 5.3 starts the description of our proposal by introducing the testbed employed for the Giga-mosaicing process. Section 5.4 studies the performances of a widely used mosaicing algorithm and points out its limitations when attempting to generate Giga-pixel mosaics in unstructured environments. Consequently, Section 5.5 lists the main requirements which must be integrated within a Giga-mosaicing scheme and provides an overview of the proposed framework fulfilling them.

Section 5.6 illustrates how we parameterize the camera motion in order to capture entirely the camera displacements and the image formation process. The proposed mosaicing algorithm is performed in two stages: first, the pair-wise image alignment establishes a global-to-local motion estimation and second, a multi-view fine refinement process is performed in order to compensate for the pair-wise possible mis-registrations. The mathematical design and the experimental results obtained for each step are presented in Sections 5.7 and 5.8, respectively. We close this chapter by summarizing the research proposal and by drawing future research directions in Section 5.9.

5.1 Once Upon a Time ... Image Mosaicing

Image stitching was pioneered back in the 1970s. Since then, it is an active research topic in Photogrammetry, Remote Sensing, Computer Graphics, Computer Vision and Robotics

research communities, promoting a wide range of applications.

Following the research community, different applications were motivating the development of image mosaicing techniques. Remote Sensing and Photogrammetry research community reported manual methods based on surveyed *ground control points* for aligning and merging aerial images into large-scale photo-mosaics [Slama, 1980]. Another traditional application is the construction of large aerial satellite photographs from collection of images [Moffit and Mikhail, 1980].

The next decade, Computer Vision research community was attacking the mosaicing problem for visual scene representation purposes [Anandan, 1995], while in Computer Graphics mosaic construction interferes with the IBR process [McMillan and Bishop, 1995], [Chen, 1995], previously discussed in Section 2.2.1. The image mosaicing process started to be widely employed for scene stabilization and change detection [Hansen et al., 1994], video compression [Irani et al., 1995a], [Irani et al., 1995b], [Lee et al., 1997], video indexing [Sawhney and Ayer, 1996], wide-angle FOV imagery [Heckbert, 1989], [Mann and Picard, 1994], [Szeliski, 1994] and high-resolution [Irani and Peleg, 1991], [Chaing and Boult, 1996], virtual environments [McMillan and Bishop, 1995] and virtual traveling applications [Chen, 1995]. Robotics research community employs mosaics as visual maps [Garcias and Santos-Victor, 2000], for localization [Ramisa et al., 2006], [Yazawa et al., 2009] and to enable SLAM capabilities [Lemaire and Lacroix, 2007] in order to supply autonomous navigation functionalities for site surveys, inspection and exploration purposes in hostile environments, where human presence is highly undesirable.

Means for mosaicing imagery. Several techniques have been used to supply panoramic imagery from real world scenes. Some of them employ special-purpose hardware acquisition, such as in [Meehan, 1990] in which authors capture directly a cylindrical panoramic image by recording an image onto a long film strip using a panoramic camera. A possible alternative is the use of lens with very large FOV, such as fisheye lens [Xiong and Turkowski, 1997], mirrored pyramid or parabolic mirrors [Nayar, 1997].

In this dissertation we focus on a low-cost technique which exploits a collection of partially overlapped images covering a desired FOV. Image stitching algorithms [Szeliski, 2006], [Brown and Lowe, 2007] were designed to align and merge a collection of partially overlapped images in order to deliver wide-angle imagery. Such composites are the result of the image mosaicing process resumed in Figure 5.1, which allows to increase significantly the image resolution and improves the SNR through the superresolution use.

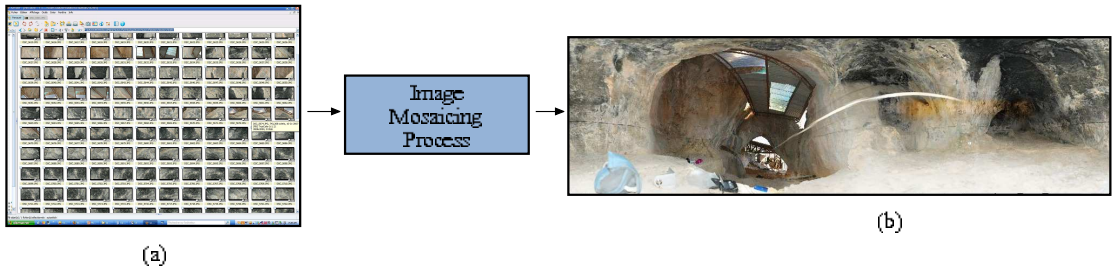


Figure 5.1: The purpose of the image mosaicing process exemplified on data acquired in the Tautavel prehistoric cave, France. (a)-input data consisting of several partially overlapped images injected in the mosaicing process which outputs a wide field of view image composite-(b).

A wide applicability of mosaic imagery was reported within the last three decades of the Digital Age. Figure 5.2 resumes the fast evolution of the mosaicing systems. Since

the reported techniques are influenced by the application type, we can place a virtual frontier between them, splitting them in two classes. Interactive and automatic mosaicing techniques are being introduced by Computer Graphics and Computer Vision research communities, promoting consumer-level mosaicing. More industrial-oriented applications have been reported in Remote Sensing and Robotics research communities.

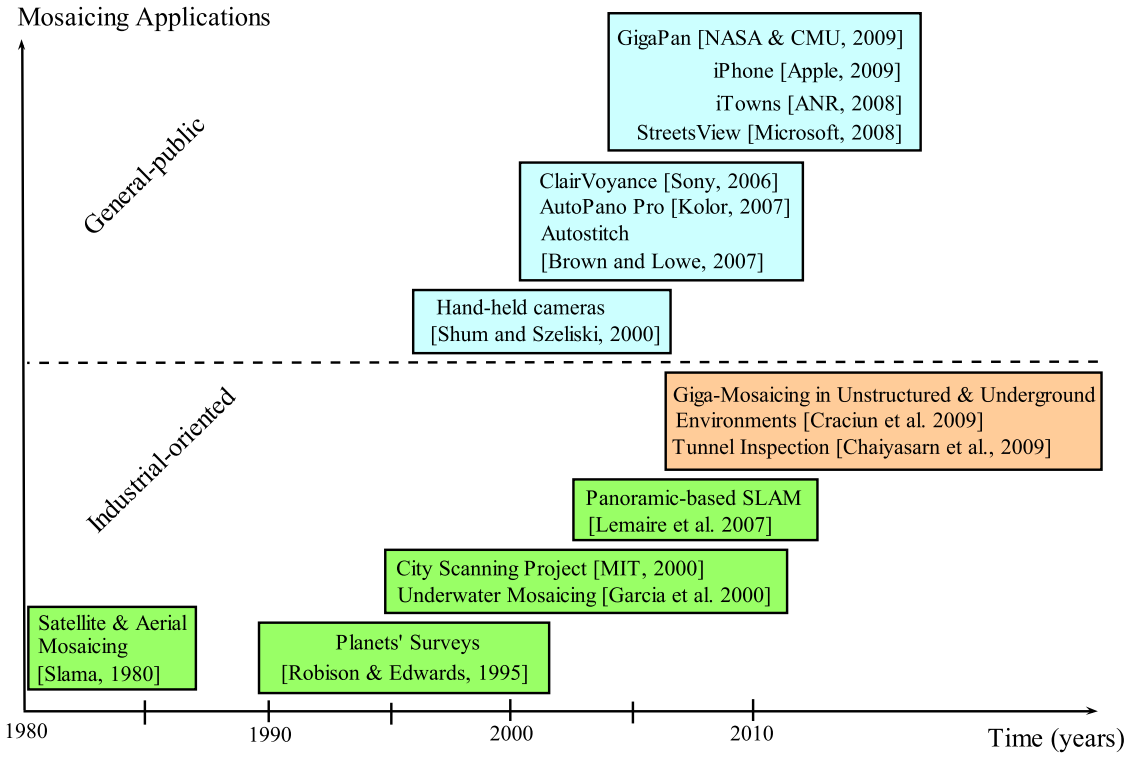


Figure 5.2: The evolution of the mosaicing applications wrt their capabilities. Our research proposal is highlighted in orange.

Consumer-level mosaicing. In mid 1990s, image mosaicing algorithms were limited to cylindrical panoramas acquired by cameras rotating on leveled tripods adjusted to minimize the parallax motion [McMillan and Bishop, 1995], [Chen, 1995], [Szeliski, 1996], limiting the range users to professional photographers dotted with such dedicated equipment. This bottleneck was overcome in the following decade when Shum and Szeliski introduced a mosaicing system designed for hand-held cameras [Shum and Szeliski, 2000], making mosaicing accessible to general-public.

In the late 2000s, commercial software releases are being made available for consumer-level mosaicing uses [Kolor, 2005], [Brown and Lowe, 2007]. Moreover, with help from fast computing and cheap disk storage devices, image mosaicing softwares have been ported onboard mobile phones [Lowe, 2007], [Lopez et al., 2009] and improved with Gigapixel capabilities [Kolor, 2009]. In [Nielsen and Yamashita, 2006], Sony© have recently reported the *ClairVoyance* mosaicing system for generating Gigapixel imagery. Gigapan^{TM} [GigaPan, 2009] is the latest release of the Global Connection Project developed by Carnegie Mellon University and NASA Ames Intelligent Robotics Group, with support from Google and BBC. The project aims at providing customers with a Gigapixel mosaicing package containing a robotic digital camera mount for capturing high resolution images, a software

for constructing Gigapixel panoramic images and a web site to allow hosting and sharing of panoramas through the world wide web. The project initiators are confident that this will bring together different communities across the globe having as main goal to explore and share each other's discovery experience. Nowadays web-based applications such as Google Streetview[1] are being powered by spherical panoramic views to enable virtual traveling. iTowns [iTowns, 2008] is an ongoing research project willing to provide virtual walkthroughts and high-level semantic data extracted from panoramic images obtained from a sequence of poorly overlapped images acquired on-the-fly by a 10-camera network mounted on a vehicle. A description of the mosaicing algorithm can be found in [Cannelle et al., 2009]. All the aforementioned methods yield good results for pure-rotating camera motion and small amount of parallax. One of their main drawbacks is that they cannot handle scene motion which introduces ghost effects within the final mosaic image. Recently, this problem has efficiently been addressed in [Qi and Cooperstock, 2007], [Qi and Cooperstock, 2008].

The above mosaicing systems were designed to enable general public applications or for hobbyists willing to produce manually wide-angle imagery. On the other side, Remote Sensing and Robotic research communities promote exclusively less accessible to general-public and more industrial-oriented applications of image mosaicing techniques.

Industrial-oriented applications. The Remote Sensing research community was one of the first areas to report an intensive use of mosaic imagery for site survey purposes [Slama, 1980], [Moffit and Mikhail, 1980]. Researchers from Lunar and Planetary Institute report the use of mosaics for investigating the color and albedo of planet Mercury [Robison and Edwards, 1995]. The use of mosaics for seabed mapping is reported in [Rzhanov et al., 2000] in order to allow to survey the ocean floor and to inspect underwater structures. Recently, in [Chaiyasarn et al., 2009] authors reported an ongoing research project aiming to employ mosaic imagery for tunnel inspection with an application to the underground infrastructure maintenance in London, United Kingdom.

Within all the aforementioned systems the mosaic processing is performed off-line which leads to missing data and misalignments, precluding therefore the in-situ interpretation of the acquired data. This is actually one of the main shortcomings of the current mosaicing methods which highlights the need for an automatic mosaicing framework in order to visualize and validate in-situ the mosaic correctness. This allows to perform accurate measurements for site inspection or exploration purposes undertaken either off-line or in-situ.

Robotics research community aims at integrating image mosaicing algorithms onboard unmanned platforms to supply active-vision, including autonomous navigation tasks. To this end, several research topics were attacked including panorama-based localization [Ramisa et al., 2006], [Yazawa et al., 2009], the use of mosaics as visual navigation maps [Garcias and Santos-Victor, 2000], seabed video mosaicing using a autonomous underwater vehicle (AUV) [Sakai et al., 2004], SLAM with panoramic vision [Lemaire and Lacroix, 2007]. The major challenge standing behind the development of the aforementioned systems is represented by the possibility to embed unmanned systems with autonomous capabilities to perform site surveys in hostile environments and to supply in-situ inspection, monitoring, maintenance and exploration of difficult-to-access environments without requiring human operator intervention.

Addressing key issues for automatic image mosaicing in feature-less areas.

[1] http://maps.google.com

Various mosaicing algorithms have reported successful frameworks based on feature extraction and matching. However, when attempting to deal with image alignment problem within an environment-independent framework, an important issue which needs to be addressed is represented by the absence of reliably extractable and trackable features. This chapter attacks the image alignment problem in order to enable reliable image matching in previously unknown environments.

Giga-pixel mosaicing in unstructured and underground environments. As highlighted in Figure 5.2, the research work related in this chapter falls in the second category, promoting an industrial-oriented application. In particular, we are mainly interested in generating in-situ Giga-mosaics in unstructured and difficult to access environments using specialized equipment, being therefore less consumer-level. The proposed mosaicing framework was designed and validated on real data acquired in two prehistorical caves situated in France. The utility of such environment digitization can be considered in different applications, ranging from data archiving to 3D modeling, passing through visual maps to supply active-vision capabilities.

In this dissertation, the image mosaicing process is used for 3D modeling purposes, being an integrating bloc of the ARTVISYS system introduced in Chapter 3. To this end, the Giga-mosaicing algorithm described in this chapter allows to generate texture maps to be mapped onto the 3D point cloud obtained through the 3D mosaicing process described in Chapter 4. Moreover, the proposed algorithm can be used as a stand-alone in-situ mosaicing system, being capable to perform in-situ acquisition, processing and visualization of the sensed area in order to ensure the mosaic correctness. Finally, the system can be used to provide onboard visual scene representation to supply in-situ autonomous site surveys in difficult to access environments.

5.2 The Image Mosaicing Pipeline

After reviewing the great potential leading to a wide applicability of the mosaic imagery, we will now take a look inside the mosaicing process and see which are the main issues standing behind the in-situ generation of Gigapixel mosaics in unstructured environments.

Since its early days, the image mosaicing problem has been extensively addressed and various mosaicing frameworks have been reported within the last decade [Shum and Szeliski, 2000], [Teller and Coorg, 2000], [Szeliski, 2006], [Brown and Lowe, 2007] and all them seem to converge to the same workflow processing illustrated in Figure 5.3 (a), showing the wide understanding of the image mosaicing problem. The mosaicing process can be split in two stages, the pair-wise and the multi-view fine alignment.

As shown in Figures 5.3 (b) and (c), following the input type, different approaches were employed for each stage, each of which having its advantages and its inconvenient highlighted in green and red. The third step is concerned with the mosaic compositing for which either interactive or automatic techniques were introduced in order to obtain artistic or basic mosaic rendering, respectively.

The next three subsections briefly review the available techniques for supplying each procedure of the image mosaicing pipeline and emphasize their main shortcomings wrt our research goal: *automatic and fast in-situ Giga-mosaicing in feature-less environments*, implying all scenery type.

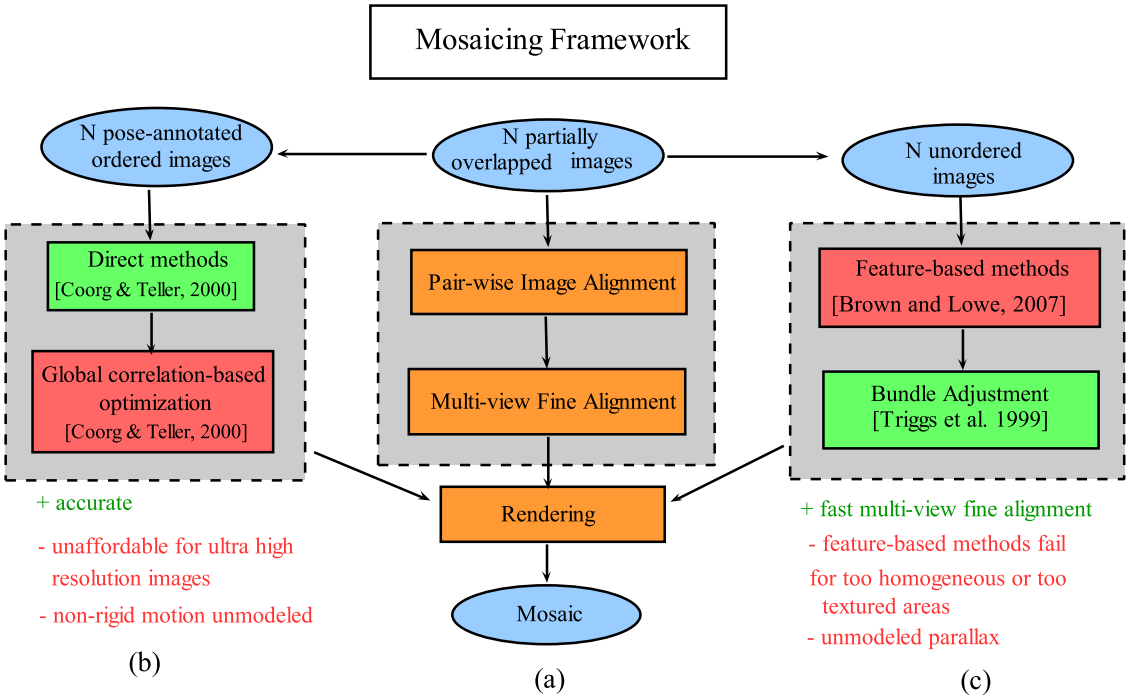


Figure 5.3: Mosaic workflow processing. (a) The general mosaicing framework; (b) - direct, (c) - feature-based approaches.

5.2.1 Pair-wise Image Alignment

When solving for the alignment of two partially overlapped images acquired from different 3D camera positions, as shown in Figure 5.4, it is necessary to back-track the camera motion encountered in-between the images' acquisition using nothing but the image data, which can be eventually calibrated beforehand.

In Figure 5.4 each camera has its referential coordinate system centered in O_{C1} and O_{C2} respectively, being related through a 3D rigid transformation. On the other side, the spatial camera motion is reflected in the 2D image space through a planar transformation for which a hierarchy is presented in Appendix C.1, Table C.1, each of which having its corresponding parametrization in the 3D world space.

Let us now set I_1 the template image and I_2 the target image. Each camera images different physical surfaces and the intersection of their FOV leads to an overlapping region in the 2D image space. In Figure 5.4 the spatial point $\mathbf{p}$ expressed in the world coordinate system $O_{(x,y,z)}$ belongs to the physical surface commonly imaged by the two cameras. Its corresponding pixel locations in I_1 and I_2 are given by $\mathbf{u}_1 = (u_x, u_y)_1^T$ and $\mathbf{u}_2 = (u_x, u_y)_2^T$, respectively.

The image alignment task requires to recover the 2D transformation which must be applied to pixels belonging to the target in order to minimize an error function measured in the overlapping area between the template and the target images. Finding such image points pairings $\mathbf{u}_1 \longleftrightarrow \mathbf{u}_2$ corresponding to the same 3D point $\mathbf{p}$ is the key element of the image alignment process [Hartley and Zisserman, 2004].

Applications requiring highly accurate geo-referencing are still making use of artificial landmarks whose 3D position are previously measured. Reliable means to perform this

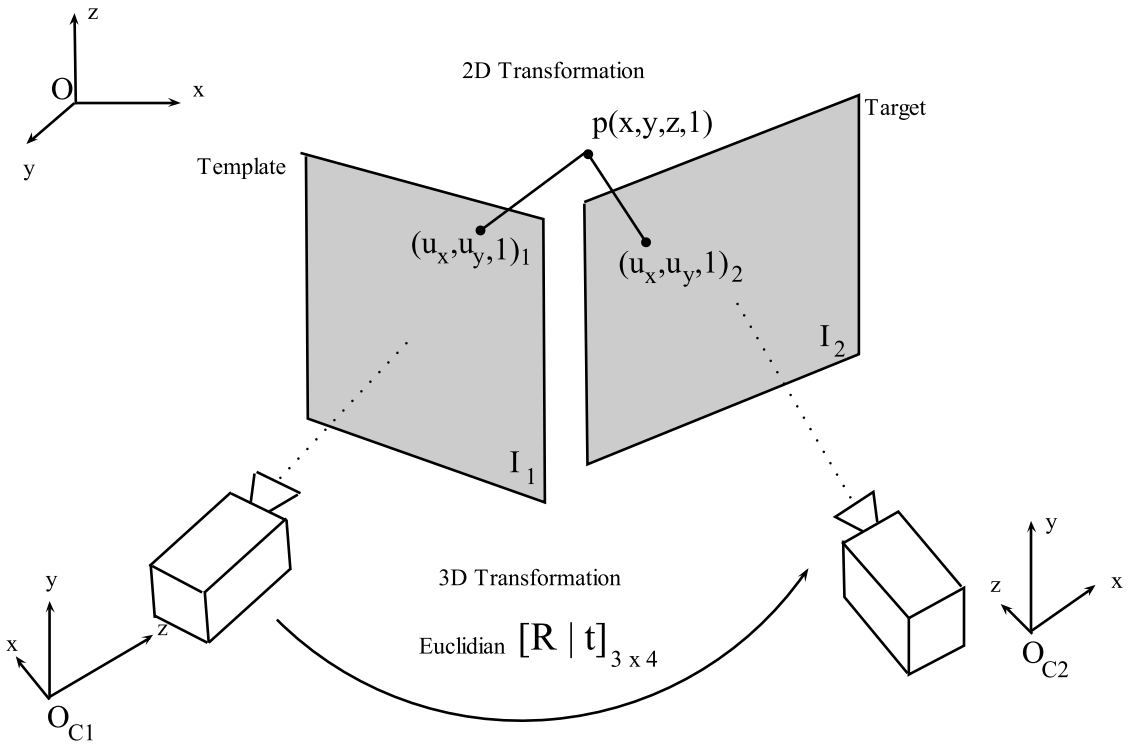


Figure 5.4: Camera motion encountered in-between two successively acquired images with a partially overlap region.

automatically are still an open issue, being particularly hard to solve in unstructured and textureless areas due to the high-ambiguity of the feature extraction and matching algorithms.

In order to come out with a reliable algorithm solving for the image alignment problem in feature-less areas, it was necessary to analyze first the available pose estimation strategies.

The next section reviews the existing techniques solving for the pair-wise image alignment. Readers not familiar with the image formation process can refer the Appendix C.2 which presents the technical background on perspective geometry and camera calibration.

5.2.1.1 Image Motion Estimation Strategies

When computing the relative poses, one must establish three main factors: (i) model the pixel relation by a parametric or non-parametric motion model; (ii) an error metric to quantify the quality of the alignment and (iii) a technical solution for exploring the available solution space. The possible parametric models are illustrated in Table C.1 from Appendix C.1.

As shown in Figures 5.3 (b) and (c), following the input data type, two main approaches are usually employed for the image alignment task. A detailed description of these methods can be found in Sezliski's tutorial on image mosaicing [Snavely et al., 2006]. This section recalls each one's characteristics willing to emphasize their complementarity which is later exploited by our research proposal.

(A)Pixel-to-pixel methods. When the input is a set of *ordered and pose-annotated*

images, direct approaches [Teller and Coorg, 2000] are usually employed for computing the relative pose estimates by minimizing a radiometric criterion measured in the overlapping area.

Technical solution. A possible least-squares solution is to minimize the *sum of squared differences* in brightness (SSD [Anandan, 1989]) expressed in Equation (5.1) which provides the optimal solution wrt to Gaussian noise.

$$E_{SSD}(\mathbf{\Delta u}) = \sum_i [I_2(\mathbf{u}_i + \mathbf{\Delta u}) - I_1(\mathbf{u}_i)]^2 \tag{5.1}$$

where $\mathbf{\Delta u}$ denotes the pixel displacement and E_{SSD} is the residual error. More robust metrics have been proposed by replacing the squared error with a robust norm [Huber, 1981], [Black and Rangarajan, 1996]. In video coding a widely employed function is the *sum of absolute differences* (SAD). However, since is not differentiable in the origin, it is not suitable for gradient-descend approaches [Szeliski, 2006]. Consequently, researchers have directed their studies toward cost functions characterized by a quadratic growth around the origin and slow convergence as getting far from the origin [Black and Rangarajan, 1996].

Illumination changes. When using a radiometric criterion, one must solve for the exposure difference problem which may rise when attempting to align two partially overlapped images acquired with different exposures. To this end, researchers introduced the *bias and gain* model [Lucas and Kanade, 1981], [Gennert, 1988], [Baker et al., 2003] which can be easily integrated within a least square estimation problem. Note that for color images such technique requires to estimate a different bias and gain for each channel in order to compensate for the automatic correction of digital cameras. In [Jia and Tang, 2003] authors established a non-parametric model to count for intensity variation within the registration process, providing robustness to lens *vignetting* caused by wide-angle cameras. In other cases authors pre-process images with a band-pass filter [Burt and Adelson, 1983], [Bergen et al., 1992] or maximize the mutual information [Viola and Wells-III, 1995], [Kim et al., 2003].

Instead of using SSD jointly with the bias and the gain model, another widely radiometric criterion is the *normalized cross-correlation* (NCC) criterion which is invariant to illumination changes. As stated in [Szeliski, 2006], in presence of low-contrast regions its performance decreases, being given that NCC is undefined for zero-variance patches.

Computational burden. Direct methods are accurate but unaffordable for high resolution images when using *full search* methods, even if a close initial estimation of $\mathbf{T}$ is given. To speed up the motion estimation process, coarse-to-fine solutions were introduced [Quam, 1984], [Anandan, 1989], [Bergen et al., 1992], [Craciun et al., 2009]. Note that for low-resolution images the pyramidal structure may coarsen the representation too much, causing blur of discriminative features. A solution to this problem is the use of Fourier-based alignment techniques [Nielsen and Yamashita, 2006] as an alternative for speeding up the image correlation process. Furthermore, it can also accelerate the sum of squared difference function and its variations [Szeliski, 2006].

Sub-pixel accuracy. Up to now the aforementioned techniques are capable of achieving motion estimation with pixel accuracy. The estimate can be further refined incrementally to achieve subpixel accuracy [Tian and Huhns, 1986]. One may choose to evaluate the cost function around the optimum and interpolate the matching score to find the analytic maximum. Another widely used approach is to perform gradient descent [Lucas and Kanade, 1981] on the cost function using a Taylor series expansion which leads to the cor-

rect solution only when the initial guess is close to a few pixels. In [Jurie and Dhome, 2002] the authors estimate the Jacobian using a least-square fit to a series of larger displacements in order to increase the range of convergence. Another technique combines the incremental approach with coarse-to-fine strategies [Bouguet, 2000]. The iterative estimation process is stopped when the magnitude of the correction displacement is less than an established threshold.

Lack of texture. Other metrics attempt to deal with the lack of texture in both directions which causes the rank-deficient Hessian matrix, resulting in wired guesses. In [Simoncelli et al., 1991], [Baker and Matthews, 2004b], [Govindu, 2006] the authors attempt to mitigate this problem by adding soft constraints on the expected motions. In [Triggs, 2004] authors showed that in practice the Gaussian model assumption [Simoncelli et al., 1991] is not verified due to the aliasing along strong edges.

Dealing with noise. Robust error metrics and weighting techniques can be used jointly within the Lucas-Kanade rule, leading to the *Iteratively Re-Weighted Least Squares* (IRLS) algorithm which alternates at each step the weights computation and the weighted least squares estimation process [Huber, 1981], [Stewart, 1999]. Other incremental least squares algorithms were reported in [Sawhney and Ayer, 1996], [Black and Rangarajan, 1996], [Baker et al., 2003].

Parametric motion model. Instead of estimating a translational motion model, gradient-decent methods can be used for estimating a parametric motion field $\mathbf{M}$, by minimizing the following criterion:

$$E_{12} = \sum_i (I_2(\mathbf{u}_i; \mathbf{T}_{2D}) - \mathbf{I}_1(\mathbf{u}_i))^2 \tag{5.2}$$

Since the Hessian and the residual vectors are computationally more expensive than for the translational case [Baker and Matthews, 2004a], researchers employed a patch-based approach [Shum and Szeliski, 2000]. For complex motion models, such as homographies the Jacobian computation becomes more elaborated, involving per-pixel division. In [Szeliski and Shum, 1997] authors proposed a simplified approach by first warping the target image using the initial estimate and by comparing the warped image against the template. In this context it is assumed that the images are similar and therefore an incremental parametric motion evaluated in the origin vicinity suffices. Several strategies were introduced to speed up the computation time: *forward compositional* [Baker and Matthews, 2004a], *forward additive* [Hager and Belhumeur, 1998], [Baker and Matthews, 2004a] and *inverse compositional* [Baker and Matthews, 2004a]. In [Baker and Matthews, 2004a] the authors compare the advantages of the Gauss-Newton iteration wrt to other techniques such as steepest descent and Levenberg-Marquardt. More advanced research topics on these approaches, including efficient strategies for pixel-weighting can be found in [Baker and Matthews, 2004a], [Baker et al., 2003].

We end reviewing direct methods by resuming their positives and negatives, which are also emphasized in Figure 5.3. Direct methods exploit all the information available in the overlapping area, allowing for accurate pose estimates. Nevertheless, special attention must be given to the initial guess quality when non-linear minimization techniques are used. In presence of noisy initial guess, *full search* performed in a coarse-to-fine fashion is the main technique which guarantees convergence.

(B)Feature-based methods. Feature-based methods are likely to be employed in the absence of an initial estimation of $\mathbf{T}$, being powered by feature pairings previously established via feature extraction and matching algorithms. This framework was originally

reported in the Robotic research community by Moravec [Moravec, 1983] and is being increasingly employed for image stitching purposes. Early feature-based methods were initialized by incremental refinement methods. Such a technique was firstly reported in [Shi and Tomasi, 1994a] which employs a translational and affine-based patch alignment to track Harris corners [Harris and Stephens, 1988] through an image sequence.

A suitable or a general keypoint detector? The first thing to do when designing a feature-based method is to choose a suitable keypoint detector, i.e. yielding invariance wrt the transformation encountered by the capturing device. In [Schmid et al., 2000] authors report a keypoint detector survey together with an improved version of the Harris operator [Harris and Stephens, 1988]. Recently, scale and affine transformation invariant descriptors were reported in [Lowe, 2004], [Mikolajczyk and Schmid, 2004] and in [Schaffalitzky and Zisserman, 2002], [Mikolajczyk and Schmid, 2005], respectively. A speeded-up variant of robust features called SURF was recently reported in [Bay et al., 2008]. As an alternative to keypoints features, line segments can be also used to for registering images. In [Zoghlami et al., 1997] authors exploits line segments in the same manner as the keypoints to estimate homographies between image pairs. Line segments were also used jointly with local edge correspondences to produce 3D structure and motion [Bartoli et al., 2004].

Feature matching strategies. Assuming that several features were found in the images, one must establish a set a pairings to feed the pose estimation process. This is a big ASSUMPTION, since up to now we are not aware of a general keypoint detector and since the existing ones are not guaranteed to be present in any environment.

Different matching strategies were introduced, following the amount of displacement encountered by the camera, i.e. either small or large motions.

Early works on feature matching techniques were reported for video sequences [Shi and Tomasi, 1994a]. These methods assume that the local motion around each feature is mainly translational making them suitable to compute SSD or NCC functions over small patches around each feature point. This is usually followed by a gradient-descent technique to obtain more accurate results but with expensive computational time [Brown et al., 2005]. When larger motions are encountered, a translational model can be first established to initialize an affine registration method [Shi and Tomasi, 1994a]. This matching approach is refereed to as *detect and track* in [Szeliski, 2006].

When matching is performed between several images separated by larger unknown motions, *detect and match* techniques [Schaffalitzky and Zisserman, 2002], [Brown and Lowe, 1983] are employed to firstly detect features in all images. This is a more complex case, since features are subjected to different orientations and scales, the feature recognition task requires for a view-invariant descriptor to be designed. An evaluation of such descriptors can be found in [Mikolajczyk and Schmid, 2005]. In the descriptor evaluation reported in [Mikolajczyk and Schmid, 2005] authors conclude that SIFT descriptors [Lowe, 2004] yield the best performances, followed by steerable filters [Freeman and Adelson, 1991] and the cross-correlation criterion. In [Brown et al., 2005] authors demonstrated that the NCC yields good results for small inter-image displacements with an application to image stitching. Recent research directions in the field of feature descriptors are oriented toward the use of PCA of SIFT [Ke and Sukthankar, 2004]. In [Weijer and Schmid, 2006] authors exploit color information to design image descriptors.

Speed issues. Assume that a set of features were extracted from two partially overlapped images. The most simple method for finding feature pairings is to compare all points in an image against all points in the second image. Since this approach leads to a quadratic growth wrt the number of features, several techniques for *rapid indexing* were introduced

based by finding the nearest neighbors in high-dimensional spaces including k-d trees spatial data structures [Samet, 1989] and modified versions of it: Best-Bin-First (BBF) [Beis and Lowe, 1997], local-sensitive hashing [Shakhnarovich et al., 2003], parameter-sensitive hashing [Brown et al., 2005] and metric tree [Nister and Stewenius, 2006]. Although the aforementioned techniques are a first step forward made for reducing the combinatory of the matching process, techniques for computational saving are still required in order to apply feature-methods on ultra-HR images for Giga-mosaicing purposes, which is the case in our research work.

Dealing with false matches (outliers). Considering that a set of initial features correspondences has been established, it is required to find the set which will produce an accurate alignment. Direct estimation methods use a coarse-to-fine approach to first lock onto a coarse estimate which is further refined on higher resolution levels [Bergen et al., 1992], [Craciun et al., 2009]. In exchange, a general approach for feature-based methods is to employ a least-square solution via IRLS [Szeliski, 2006]. A better technique is to start the estimation directly with a set of pairings points which are consistent with the estimated model within an error tolerance of a few pixels (*inliers*), previously established via a random sampling algorithm, such as RANSAC [Fischler and Bolles, 1981] or *least median of squares* (LMS) [Rousseeuw, 1984]. A random selection process of inliers set is repeated and the subset with the largest number of inliers (or with the smallest residual error) is kept as the final solution. A rapid variant of RANSAC, PROSAC (PROgressive SAmple Consensus) can be found in [Chum and Matas, 2005].

Geometric registration. In contrast to direct methods which minimize a radiometric measure, feature-based techniques minimize a geometric error measured between the corresponding coordinate points to estimate a parametric model motion relating two partially overlapped images. The criterion is written as follows:

$$E_{LS} = \sum_i \|\mathbf{r}_i\|^2 = \|\mathbf{u}_i^1 - (\mathbf{u}_i^2; \mathbf{T}_{2D})\|^2 \tag{5.3}$$

For motion models presented in Appendix C.1, Table C.1 having a linear relationship between the motion and the parameters, usually a simple linear regression using normal equations performs well. Some of LS solutions include uncertainty weighting to take into account the matching accuracy. Although RANSAC is supposed to produce consistent matches, the matching process is always subject to outliers and IRLS methods are used to robustly weight point matches.

For non-linear motion parameters, such as homography, an iterative solution is required to obtain accurate results. This solution is usually obtained via Gauss-Newton approximation which implies a first-order Taylor series expansion. Each iteration estimates an incremental motion which is used to iteratively update the parametric motion. However, this is highly subject to the initial estimation accuracy.

Dealing with arbitrary camera motion. When constructing a closed 360° panorama from multiple partially overlapped images by concatenating the pair-wise poses estimates, the accumulated error lead to either a gap or an excessive overlap between the two ends of the panorama. A possible solution to this problem is proposed in [Szeliski and Shum, 1997] which consists in distributing the error equally across the whole sequence. Related approaches reported solutions for the focal length estimation for the case of pure panning motion and cylindrical images [Hartley, 1994]. Such *gap closing* strategies are limited to one-dimensional panorama case and more sophisticated techniques were introduced in order to deal with arbitrary camera motions [Shum and Szeliski, 2000], [Brown and Lowe,

2007].

Nevertheless, when dealing with arbitrary camera motions, the pair-wise estimates are subject to mis-registration errors due to the 3D parallax, mis-calibration errors or scene motion. A traditional approach for dealing with this issue is to refine the relative poses within a global alignment step, which is actually the second main step of the mosaicing pipeline for which a description is provided in the following subsection.

5.2.2 Multi-view Global Alignment

Generally, a simple concatenation of the relative poses lead to a globally inconsistent alignment. This is generally due to the following geometric and radiometric factors, which cause ghosting and blurring effects in the final mosaic composite.

- *camera motion*: unmodeled parallax (failure to rotate the camera around its optical center), deviations from pinhole camera model [Hartley and Zisserman, 2004] and unmodeled camera distortions;
- *deal with various exposures*: when the camera is used in the automatic mode or when casually acquired images are stitched, the image alignment algorithm must be robust to illumination changes. A challenging ongoing project developed by Microsoft is the Photosynth® system [Snavely et al., 2006] in which authors attempt to match several images under a different lighting conditions, seasons, day time and so on;
- *scene motion*: dynamic scenes (at small and large scale, ranging from people to tree branches) affects the matching accuracy and introduce ghost effects in the final compositing.

The aforementioned items lead to several radiometric and geometric inconsistencies for which additional processing must be included within the 3D mosaicing pipeline. The geometrical correction is concerned with the poses' refinement, while the radiometric correction is usually integrated within the mosaic rendering process, being actually the last step of the mosaicing pipeline and for which a description is provided in the next section.

This section is dedicated to the geometrical correction which is seen as an optimization step and implies the computation of a set of globally consistent alignment parameters by minimizing the mis-registration errors measured over all image pairs [Szeliski and Shum, 1997], [Shum and Szeliski, 2000], [Sawhney and Kumar, 1999], [Teller and Coorg, 2000]. This can be accomplished by extending the pair-wise criteria (i.e. radiometric or geometric) to a global error function which evaluates the relative estimates over the entire image sequence composing the mosaic.

As shown in Figures 5.3 (b) and (c), following the pair-wise image alignment method preceding the multi-view stage, two main approaches are usually employed for the geometrical correction performed within the multi-view alignment step, which are briefly reviewed hereafter.

Geometrical correction through bundle adjustment (BA). A big step toward globally consistent solutions was the development of the *bundle adjustment* (BA) algorithm which allows to solve simultaneously for all the cameras poses. Pioneered in the photogrammetric community [Slama, 1980], it has increasingly been used by computer vision scientists to solve for the structure from motion problem [Szeliski, 1994] and later on for mosaicing purposes [Sawhney and Kumar, 1999], [Shum and Szeliski, 2000], [Brown and Lowe, 2007].

Generally, the BA algorithm is fed with a set of image point pairings previously established, being therefore likely to be employed within a feature-based framework. Nevertheless, as stated in [Szeliski, 2006], one can divide the image into patches and generate virtual corresponding features [Shum and Szeliski, 2000]. In its original form, the BA scheme solves simultaneously for the camera motion parameters and the 3D structure by minimizing the geometric error between the features belonging to the template image and the projections of the corresponding features in the target image obtained using the current motion model estimate. Note how the pose estimation is highly-dependent on the quality of the initial guess. In addition, this process has long iterations and slow convergence. One solution for reducing the computational complexity of the Gauss-Newton step is to use sparse matrix techniques [Szeliski, 1994], [Triggs et al., 1999], [Hartley and Zisserman, 2000], [Lourakis and Argyros, 2004].

BA-variants. Several variants of BA were reported, each of which employs different poses formulations, including a rotation matrix and a focal length, [Shum and Szeliski, 2000], but also in terms of homographies [Shum et al., 1997], [Sawhney and Kumar, 1999]. Estimating a 3D rotation matrix and optionally a focal length is intrinsically more stable than estimating a 8-D.O.F. homography, which justifies the use of this method for large-scale image stitching algorithms [Szeliski and Shum, 1997], [Shum and Szeliski, 2000], [Teller and Coorg, 2000], [Brown and Lowe, 1983]. Usually, when blending the images using the orientations and focal length estimates obtained through the BA process, it is often observed that the final composite contains artifacts, i.e. blur and ghost effects caused by a variety of factors including unmodeled radial distortion, 3D parallax and scene motion. When the BA process does not model the 3D parallax nor the camera distortions, the global alignment might be followed by a local alignment step via a patch-based optical flow technique, such as in [Shum and Szeliski, 2000]. However, since this approach does not model explicitly the error source, it can fail often or introduce unwanted distortions.

Dealing with distortions, parallax and dynamic scenes. Each of the aforementioned problems can be addressed individually: the radial distortion can be estimated by calibrating the camera beforehand and although more expensive, the parallax estimation can be integrated within the BA scheme. For dealing with objects which appear and/or disappear completely, a less robust solution is to select pixels only from one image source when computing the final composite [Milgram, 2006], [Davis, 1998], [Agarwala et al., 2004]. In [Qi and Cooperstock, 2007], [Qi and Cooperstock, 2008] authors deal with moving objects and parallax within the registration process.

Solutions for small or large parallax. Following the mosaicing acquisition scenario, either negligible or small parallax amounts can be introduced. The first case supposes that the images might be acquired from the same optical center or that the cameras slightly separated. Therefore, an optical flow approach suffices for capturing negligible up to small parallax motions. In exchange, the latter case requires to include the parallax estimation within the pair-wise estimation process in order to ensure that the multi-view fine alignment does not get stuck in a local minimum.

Solving optimally for distortions and parallax. An optimal approach to address this issue is to solve for the camera distortions before the first camera use and to compensate for parallax effects within the early stage of the pair-wise image alignment process in order to model and compensate different motion sources separately.

Spherical mosaicing via dense-correlation and quaternions [Teller and Coorg, 2000]. Direct methods [Teller and Coorg, 2000] employ a global correlation function defined for adjacent images wrt all orientations encoded as quaternions. In [Teller and

Coorg, 2000] authors refine the initial guess provided by the physical instrumentation using the LM non-linear optimization to produce an unique rotation for each image. In order to avoid gaps between the first and the last images, authors constraint the system such that the composed rotations along any cycle is the identity matrix.

The multi-view global alignment process solves for the geometric correction which improves significantly the mosaic quality. The remaining issue now is represented by the radiometric artifacts, which is generally solved within the rendering process. Nevertheless, in other research works authors attempt to solve for the remaining geometric errors within the rendering process, such as panorama straightening in [Brown and Lowe, 2007].

5.2.3 Mosaic Compositing

The last step of the image mosaicing pipeline exploits jointly all images and their associated transforms to warp the input images onto a parametric surface to form the mosaic image. Additionally, when one is concerned in providing an artistic rendering, it is within this stage that an artifact "make-up" is performed by focusing on how to combine pixels from different image sources in order to display overlapped areas when different exposures are encountered. This operation is refereed to as *feathering*.

Basic or artistic rendering. The mosaic image rendering step is highly dependent on the application type, not to mentioned that one can use a mosaic image without displaying it effectively, but by storing it as a combined topological-semantic map based on geometric and color attributes. However, most applications are interested in displaying mosaic imagery and they usually require fast rendering and visualization methods, willing to enable smooth navigation between panoramic-views.

Following the application type, different criterias are governing the rendering process. The first class employs mosaic imagery as visual maps, aiming to endow unmanned mobile platforms with active vision to supply autonomous navigation capabilities. These systems privilege fast in-situ rendering procedures, while the second group focuses on providing an artistic rendering for data archiving, virtual traveling and various general-public applications. More details on basic and artistic rendering can be found in Appendices C.3 and C.4, respectively.

5.3 Gigapixel Mosaicing Testbed

This section presents the experimental platform employed in this chapter for generating Giga-pixel color mosaic. The inputs of our algorithm are several hundreds of ordered HR images acquired from a common optical center by a NIKON® D70 digital camera fixed on a motorized pan-tilt head as shown in Figure 5.5. Although this chapter introduces a stand-alone Giga-mosaicing system, the entire system is an integrating part of the ARTVISYS hardware and acquisition scenario, as previously presented in Section 3.2.

Acquisition setup. The system's setup is parameterized through a laptop/pocket PC with the FOV to be covered by the mosaic and the desired overlap separating two images. The platform delivers a sequence of pose-annotated and high-resolution color images of size $n_u \times n_v = 3000 \times 2008$, where n_u and n_v denote the columns and the rows number, respectively. We employ a long-camera focal $f = 50$ mm to minimize lens vignetting effects, with a FOV of $H_{FOV} \times V_{FOV} = 26.34^\circ \times 17.73^\circ$ and pixel size of $7.8\mu m$. The camera is calibrated off-line and the radial distortion modeled with a 3-rd degree polynomial. As we will see further in this chapter, the use of a long focal allows to capture details over

(a) (b)

Figure 5.5: Mosaicing acquisition System: a NIKON® D70 digital camera (a) with its optical center fixed on a motorized pan-tilt head (Rodeon manufactured by Clauss®) attached to a tripod base (b).

long range sceneries, while images of low-depth scenes can be affected by blur if automatic focus fails.

The camera is used in the automatic mode and the shooting delay is fixed to 5s in order to provide enough time for synchronization with the Rodeon® platform. Generally, we aim at generating an image mosaic with a $360° \times 180°$ FOV and for the first experiment images with an overlapping area of 33% were acquired, leading to $N_{station} = 315$ images. Later in this chapter we show that our algorithm exhibits robustness for images with an overlap less than 1%.

Theoretical initial guess. The acquisition platform assigns to each image a relative pose $(\theta, \varphi)_{hard}$ corresponding to the theoretical rotations values imposed to guarantee the required overlap and the final mosaic's FOV.

The theoretical rotations delivered by the platform are related to the FOV to be covered by the mosaic given by FOV$=[\theta_{min}, \theta_{max}] \times [\varphi_{min}, \varphi_{max}]$ and to the number of images to be used for covering the entire field $N_{station}$, itself being related to the size of the overlapping areas. Therefore, the relative pose delivered by the motorized heading are given by $(\delta\theta, \delta\varphi)_{hard} = (\frac{\theta_{max}-\theta_{min}}{N_{station}}, \frac{\varphi_{max}-\varphi_{min}}{N_{station}})$. In practice, we observed that the difference between the theoretical and the physical rotation of the platform may vary within a range of $\Delta\epsilon_{hard} = \pm 5°$ in each direction.

Improving platform's motion control. In order to record the physical motion of the Rodeon® platform, it was therefore necessary to improve the quality of the initial guess provided by the physical instrumentation. To this end, our research lab developed a software to control and record the physical rotations of the capturing device. Although the platform's improvement does not influence the algorithm design, it was necessary for generating in-situ Gigapixel mosaicing with real time performances.

Capturing platform's motion. Although in theory the motorized platform undergoes 2D-rotations, the spherical acquisition geometry gives rise to rotations which can be better captured by adding a 3^{rd} angle - yaw denoted by ψ in order to count for rotations around the optical axis, corresponding to the Oz axis. It worths noting that the physical instrumentation does not provide an initial estimation for ψ_{hard}. Yaw effects are negligible for tilt values near the equator and increase as tilting either negatively or positively toward poles. As the motorized heading tilts upwards (or downwards) to reach north (or south)

pole, pan motions appears like rotations around the optical axis when projected in the 2D image plane. Figure 5.6 illustrates the spherical acquisition geometry performed by the experimental platform.

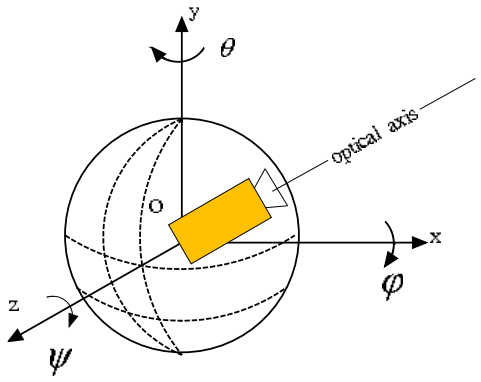


Figure 5.6: Rodeon® spherical acquisition geometry.

Being given the camera calibration and the initial guess $(\theta, \varphi)_{hard}$, it is necessary to refine the camera orientations in order to solve for the image alignment problem. To this end, in this dissertation we tackle the camera motion estimation within two phases. We first evaluate the behavior of a widely employed image mosaicing algorithm when attempting to generate Giga-mosaic imagery in unstructured and underground environments. Second, to solve for its limitation, we introduce a pair-wise image alignment which can be used in conjunction with a BA scheme to align and merge multiple partially overlapped images into a mosaic.

5.4 Existing Mosaicing Methods' Performances

It is very important to draw the limits of the currently existing mosaicing framework, and therefore, in this section we present an evaluation of the state-of-the-art mosaicing techniques on a data acquisition scenario performed in the Tautavel prehistoric cave situated in France. We employ the image stitching algorithm reported in [Brown and Lowe, 2007] which is being widely employed for commercial and industrial purposes.

The technique falls in the category of feature-based methods, being powered by a SIFT-based image matching algorithm. The image motion is parameterized by affine motions, justifying the use of SIFT features [Lowe, 2004] which are partially invariant to affine motions. The pose estimation is performed by utilizing the Direct Linear Transform (DLT) [Hartley and Zisserman, 2004] in conjunction with RANSAC [Fischler and Bolles, 1981] and verified using a probabilistic model. Under the assumption that the pair-wise alignment process yields geometrically consistent matches, the multi-view fine alignment follows a BA scheme, which beside the computation of the rotations parameters, re-estimates the camera internal parameters. The rendering process aims at providing artistic panoramic obtained from casually acquired images, including automatic straightening, gain compensation and multi-band blending.

Various implementations are available for both OSs, Linux and Windows. We tested a C version under Linux and on Windows a Matlab source, Autostitch [Lowe, 2007] and Autopano Pro [Kolor, 2005]. The evaluation is performed at different levels of the mosaicing

process:

- level (1): two-steps pair-wise matching: (i) feature extraction, (ii) feature matching;
- level (2): the global alignment step via BA.

At a first glance, we are mainly interested in evaluating the algorithm's capacity to find reliably SIFT matches, i.e. level (1). On several images pairs acquired in the Tautavel prehistoric cave the extraction of SIFT descriptors failed. In other cases, SIFT were detected but the algorithm failed to match them due to the high amount of blur effects introduced by the long focal, which is the case shown in Figures 5.7 (a), (b). As for the global alignment step, i.e. level (2) illustrated in Figure 5.7 (c), it highlights the hole corresponding to the unmatched images from Figures 5.7 (a) and (b). Figure 5.7 (d) exhibits an erroneous multi-view alignment due to false matches imposing erroneous spatial constraints.

As stated in [Labrosse, 2004] and [Nielsen and Yamashita, 2006], there are several issues with feature-based methods:

Major drawback for stitching fully textured images and ultra-high resolution images have potentially such fully textured areas.

The underlying assumption sustaining that such features exist and can be efficiently extracted and matched between sensor readings leads to the following issues:

- features that can be extracted do not necessary exist (natural environments);
- features of a particular type do exist in particular environments and algorithms developed for particular features are thus not portable;
- feature types used are more often a consequence of what can be extracted from the data provided by the sensor used than that of the environment itself;
- extracting and matching features is computationally expensive and is often not reliable.

Major advantage: when used along with a suitable motion model, the BA scheme allows for fast multi-view fine alignment, which is one of our main interests for generating in-situ Giga-pixel mosaicing.

Investigation and solutions. In our research work we attack the mosaicing problem from both sides: the pair-wise image matching - level (1), and the multi-view global alignment - level (2) processes.

Level (1). Since the main problem of the pair-wise process stands in the difficulty of identifying corresponding features points, we first tackle the pair-wise image alignment problem by proposing a global-to-local motion estimation methods which delivers a list of homologous points to make possible their exploitation along with a BA step at level (2). This stage helps to evaluate the BA process proposed in [Brown and Lowe, 2007] when consistent point matches are injected into it.

Level (2). Section 5.8.1 shows tests demonstrating that the BA process re-estimates the intrinsic parameters of the camera, precluding the estimation of the optimal pose. In addition, the use of a 2D criterion causes the rejection of consistent matches, while the multi-view alignment stage distributes the remaining errors across all poses causing artifacts in the final compositing. In order to solve for this problem, in Section 5.8.2 we propose a theoretical solution solving for the multi-view fine alignment problem.

The next section provides an overview of the proposed Giga-mosaicing algorithm and justifies its design comparing to the state-of-the-art algorithms.

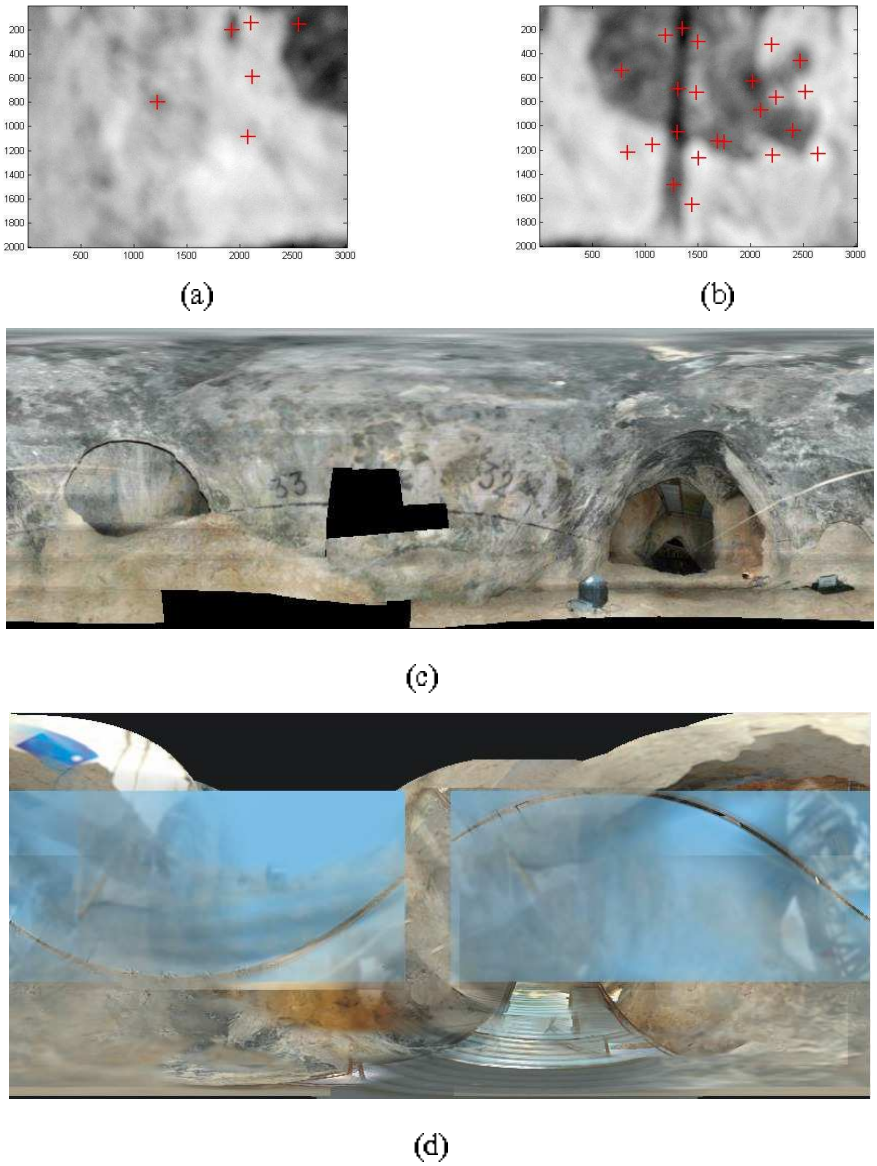


Figure 5.7: Evaluation of [Brown and Lowe, 2007] on data set acquired in the Tautavel prehistoric cave using Autostitch [Brown and Lowe, 2007] - default parameters. (a) - 8 SIFT extracted, (b)- 20 SIFT extracted, 0 matched. (c), (d) multi-view alignment and rendering on data sets acquired in Tautavel prehistoric cave - 5% of the total mosaic resolution. The two blue rectangles are the result of the bundle adjustment process in presence false matches (situated on a blue object - the laser's dust cover).

5.5 Proposed Giga-Mosaicing Algorithm

The image mosaicing problem has been widely understood, reaching a considerable state of maturity. Researchers attacked this problem from different sides, each one being interested in responding to several constraints imposed by the application context. As emphasized in Figures 5.3 (b) and (c), the main difference between the existing mosaicing approaches stands in the pair-wise image alignment step, and more important is the complementarity between the existing techniques solving for it, i.e. direct vs. feature-based.

Direct methods allow a reliable image matching in both, structured and feature-less areas, while feature-based methods are more likely to be employed in structured environments. Moreover, direct approaches privilege the *accuracy*, while feature-based the *rapidity*. Nevertheless, there is no doubt that these research works provide some idea about how the the fundamental skeleton of an environment-independent mosaicing pipeline should look like. When both approaches are strategically combined, their advantages can be grouped to yield an efficient multi-view alignment framework.

Our research work is concerned with the image matching problem in unstructured environments, including therefore the structured area case, and deals with ultra-high resolution images for generating in-situ spherical Giga-mosaics in underground environments. Consequently, we privilege the use of direct methods for reliable and accurate matching but nonetheless, since time and in-situ constraints are major concerns, we use them jointly with feature-based methods to enable fast global alignment via a BA step.

When tackling the image alignment problem for Giga-Mosaicing purposes, we prioritize the following factors:

- **Minimize every source of possible error.** Due to the error prone nature of the self-calibration process, camera off-line calibration is highly recommended when the same camera is employed during the entire mission.

- **Exploit every possible information given.** The proposed algorithm exploits the initial guess provided by the physical instrumentation and the camera calibration performed off-line. In-situ mosaicing requires fast processing which at its turn imposes the acquisition of low-overlapping images. Since the environment is previously unknown, none can predict nor guarantee salient features' existence in that particular area of overlap. Consequently, in order to ensure reliable matching, we choose to employ a dense correlation approach, exploiting therefore all the information presented in the overlapping region.

- **Model all physical motion sources.** Our main concern is to capture all the physical motions encountered by the capturing device within the early stage of the algorithm, i.e. the pair-wise image matching process.

  Rotation. In our research work, theoretically, the camera performs purely-rotations around it optical center, which are parameterized by unit quaternions [Horn, 1987].

  In practice, when modeling the camera motion using only a rotational model, we may notice visible seams due to images' misalignment. One of the main reasons is that the motorization of the capturing device yields some vibration noise which is further amplified by the tripod platform. Moreover, unmodeled distortions or the failure to rotate the camera around the optical center may result small amounts of parallax.

Small parallax. In order to handle deviations from pure parallax-free motion or from ideal pinhole camera model, we upgrade the camera motion with a non-rigid motion model estimated via a patch-based local matching procedure.

- **Robustness to noisy initial guess and accurate alignment of ultra-high resolution images.** In certain cases the initial guess may be very far for the true pose, for instance ranging between $\pm 5^\circ$ of error from the theoretical pose and $\pm 1^\circ$ for the ψ_{hard} delivered by the improved platform and consequently an incremental estimation, such as LM or gradient-descend would get stuck in a local minimum. Therefore for safety reasons, we perform a full search over the entire solution space, to ensure convergence. It is undoubtable that artifacts caused by mis-registration errors are more visible for the Giga-mosaics case and consequently, one of our main concerns is to produce accurate poses. This is another reason which enforces our choice toward the use of direct methods via dense correlation.

- **Fast processing of high-resolution images.** Since direct methods are unaffordable when applied on high-resolution images, we reduce the combinatory by correlating only border-patches (extracted in the overlapping region) along within a coarse-to-fine strategy.

  Beside the pair-wise process, a second factor related to the rapidity issue is the multi-view alignment process. When using high resolution images, direct approaches are computationally too expensive to be applied for the multi-view image alignment stage. Consequently, it is more likely to employ a BA approach for the global optimization step.

The technical solutions proposed to fulfill each of the aforementioned requirements form the ingredients of the proposed Giga-mosaicing framework proposed in this dissertation, whose global pipeline is illustrated in Figure 5.8 and detailed hereafter.

(A) Global-to-local pair-wise alignment. The initial step is first refined via a direct method which estimates a *global* 3D rotation motion which, at its turn initializes a patch-based local (non-rigid) motion estimation. The first stage estimation process is performed by exploring the solution space within a full searching strategy performed in a coarse to fine approach. The pair-wise procedure outputs a list of locally matched image points and a vector displacement for each patch, which gives the possibility to establish a global translational motion model.

More important is that the local alignment outputs a list of homologous patches which are perfectly exploitable by a BA scheme, enabling therefore a fast global alignment of high-resolution images. Since the matched points do not correspond to any corner-like features, we introduce them as *anonymous features* (AF).

(B) Multi-view fine alignment. The multi-view fine alignment is achieved by injecting the AF matches in a BA engine [Triggs et al., 1999]. We first employed the BA scheme proposed in [Brown and Lowe, 2007] for which we show in Section 5.8.1 that it does not solve efficiently for the multi-view fine alignment problem due to the re-estimation of the camera intrinsics parameters and due to the 2D criterion employed during the minimization process. Therefore, in Section 5.8.2 we propose an analytical solution for the BA step which minimizes an error measured in the 3D space and which could potentially improve the results considerably.

Exploit direct vs. feature-based methods complementarity and bridging in-between. The proposed Giga-mosaicing technique exploits the complementarity of

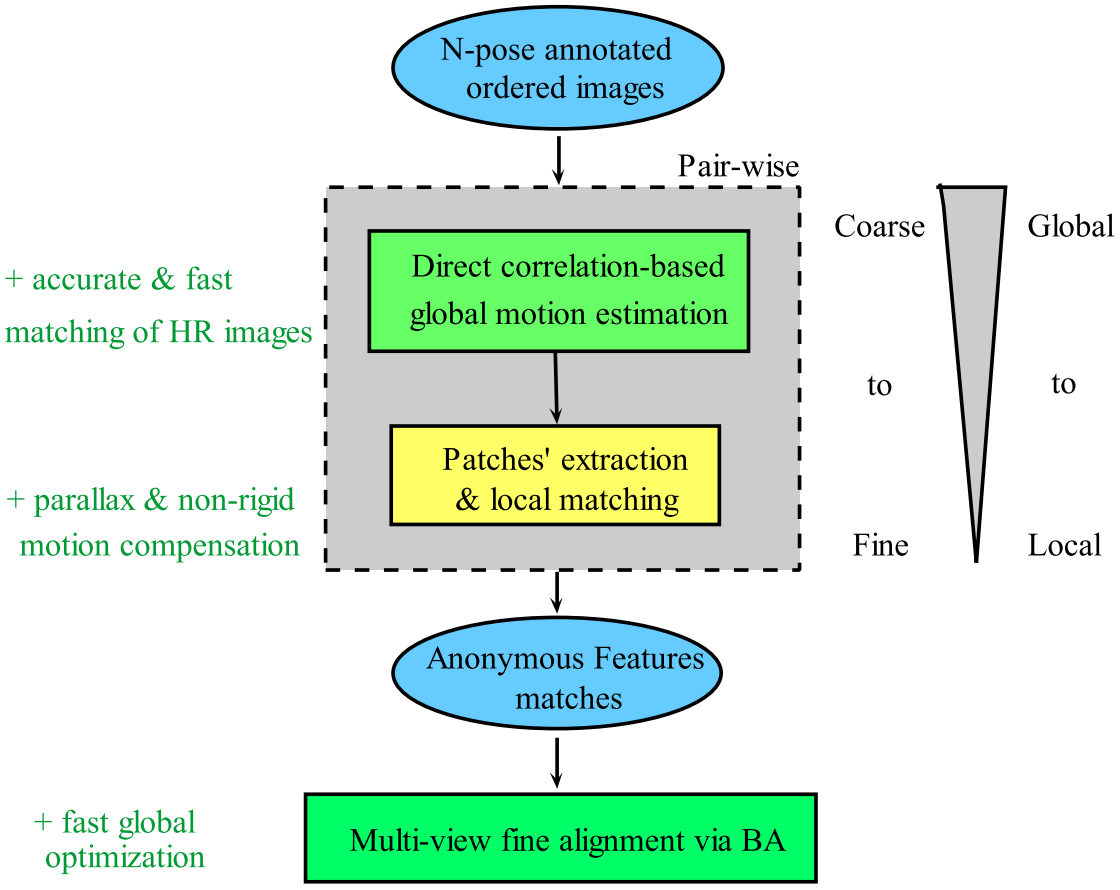


Figure 5.8: Overview of the proposed Automatic Giga-pixel Mosaicing (AGM) algorithm.

the existing image alignment techniques (direct vs. feature-based) and fuses their main advantages in an efficient fashion. It combines the *accuracy* of direct methods with the *rapidity* of the BA framework, usually employed with feature-based approached. Moreover, the *global-to-local* pair-wise motion estimation has a double advantage: first, it allows to capture small amounts of parallax and to compensate for deviations from the pinhole camera model or unmodeled distortions. Second, it provides a set of homologous point matches, providing therefore a bridge between the direct method and the BA step and enabling fast multi-view image alignment.

Comparing to other mosaicing frameworks. The proposed image mosaicing scheme detains several advantages over the existing methods.

Comparing to Teller's approach [Teller and Coorg, 2000], our method can handle very noisy initial guess and relatively small amounts of parallax. In addition, the pyramidal patch-based framework enables fast high resolution image matching which is a key aspect for the Giga-pixel mosaicing task. Furthermore, the BA scheme enables final optimization with real-time performances.

Comparing to Lowe's method [Brown and Lowe, 2007], the proposed algorithm can deal with feature-less areas, providing therefore an environment-independent method for the image alignment task.

In [Shum and Szeliski, 2000] authors deal with the image mosaicing problem for the hand-held camera case by performing a pair-wise alignment followed by a multi-view fine alignment stage. A third step is dedicated to the parallax removal and deghosting process, being performed via an optical flow approach to estimate a local motion estimation. In our case study, images are acquired from the same optical center and therefore small parallax motion or mis-calibration errors are susceptible of being introduced. In this situation, an optical flow approach can easily compensate for the residual motions which may appear. Since in [Shum and Szeliski, 2000] authors attempt to compensate for large amounts of parallax encountered by hand-held cameras, the optical flow may fail since the method does not model the real parallax sources.

The following subsections are dedicated to the camera motion parametrization and to the description of the overall flow of processing, including the global-to-local pairwise motion estimation and the multi-view fine alignment stages.

5.6 Camera Motion Parametrization

Let us now describe in more mathematical details how the camera motion is modeled and how we constraint its estimation process.

Rotational mosaics. When the camera undergoes purely rotations around it's optical center, it is assumed that all points are very far from the camera, i.e. on the plane at infinity, as shown in Figure 5.9. Since depth effects do not occur across two images acquired from the same optical center, the general perspective projection from Appendix C.2, Equation (C.4) can be simplified to a 3×3 rotation matrix $\mathbf{R}$ and the camera intrinsic matrix $\mathbf{K}$, yielding the following equation:

$$\begin{pmatrix} u_x \\ u_y \\ 1 \end{pmatrix} \cong \mathbf{KR} \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} \tag{5.4}$$

The inversion of Equation 5.4 yields a method to convert pixel position to 3D-ray. Therefore, using pixels from an image (I_2) we can obtain pixel coordinates in another image

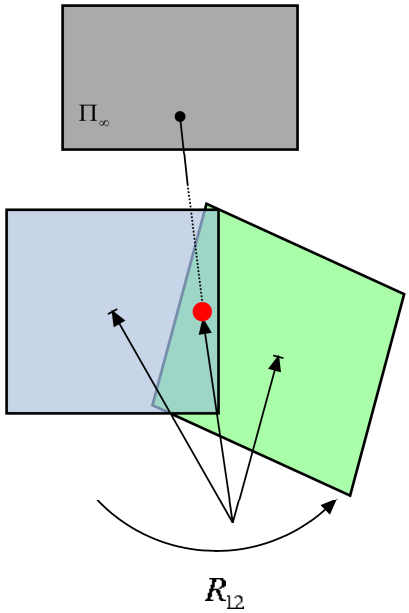


Figure 5.9: Camera encountering purely 3D rotations.

(I_1) by applying the corresponding 3D transform and by projecting the transformed points into the I_1's space using equation 5.4. This principle can be summarized by the warping equation which is expressed as:

$$\hat{\mathbf{u}}_1 \cong \mathbf{K}_1\mathbf{R}_1\mathbf{R}_2^{-1}\mathbf{K}_2^{-1}\mathbf{u}_2 \tag{5.5}$$

Assuming that all the intrinsic parameters are known and fixed for all n images composing the mosaic, i.e. $\mathbf{K}_i = \mathbf{K}, i = 1, .., n$, this simplifies the 8-parameter homography relating a pair of images to a 3-parameter 3D rotation.

$$\hat{\mathbf{u}}_1 \cong \mathbf{K}\mathbf{R}_{12}\mathbf{K}^{-1}\mathbf{u}_2 \tag{5.6}$$

Off-line camera calibration. The off-line camera calibration constraints the estimation process to 3-parameters, which is intrinsically more stable than estimating a 4- or 5-parameter model, encoding the 3 rotation parameters and a fixed or variable unknown focal length [Szeliski and Shum, 1997], [Shum and Szeliski, 2000], [Teller and Coorg, 2000], [Brown and Lowe, 1983].

Willing to minimize as much as possible distortions, we employ a long-focal camera (50 mm) introducing relative week distortions (i.e. about 1.5 pixels around the image corners). The radial distortion parameters were estimated using a 3rd degree polynomial function since in practice the use of higher order coefficients did not improve considerably the correction results.

Rotation parametrization. We employ unit quaternions $\mathbf{q}_\theta$, $\mathbf{q}_\varphi$, $\mathbf{q}_\psi$ for representing rotations around the tilt, pan and yaw axis which are denoted by their corresponding vectors $\mathbf{n}_\theta = (1, 0, 0)$, $\mathbf{n}_\varphi = (0, 1, 0)$, $\mathbf{n}_\psi = (0, 0, 1)$. The 4 components of an unit quaternion representing a rotation of angle θ around the n_θ axis are given by $q_\theta = (q_\theta^w, \mathbf{n}_\theta) = (q_\theta^w, q_\theta^x, q_\theta^y, q_\theta^z)^T$.

Consider that a rotation above the $\mathbf{n}_\theta$ axis is applied to a 3D point $\mathbf{p}$. By using the conjugate of the unit quaternion $\mathbf{q}_\theta^* = (q_\theta^w, -\mathbf{n}_\theta)$, this can be written as $\mathbf{p}_\theta = \mathbf{q}_\theta\mathbf{p}\mathbf{q}_\theta^*$. If we

apply a second rotation represented by the unit quaternion $\mathbf{q}_\varphi$ we obtain:

$$\mathbf{p}_{\theta\varphi} = \mathbf{q}_\varphi \mathbf{p}_\theta \mathbf{q}_\varphi^* = \mathbf{q}_\varphi(\mathbf{q}_\theta \mathbf{p} \mathbf{q}_\theta^*)\mathbf{q}_\varphi^* \tag{5.7}$$

Using quaternions properties is easy to verify that $\mathbf{q}_\theta^* \mathbf{q}_\varphi^* = (\mathbf{q}_\varphi \mathbf{q}_\theta)^*$, yielding:

$$\mathbf{p}_{\theta\varphi} = (\mathbf{q}_\varphi \mathbf{q}_\theta)\mathbf{p}(\mathbf{q}_\varphi \mathbf{q}_\theta)^* \tag{5.8}$$

meaning that the overall rotation is represented by the unit quaternion $(\mathbf{q}_\varphi \mathbf{q}_\theta)$. Therefore, rotation composition leads to quaternion multiplication. The product of two quaternions defined as $\mathbf{q}_\theta = (\cos\frac{\theta}{2}, \sin\frac{\theta}{2}\mathbf{n}_\theta)$ and $\mathbf{q}_\varphi = (\cos\frac{\varphi}{2}, \sin\frac{\varphi}{2}\mathbf{n}_\varphi)$ is given by :

$$\mathbf{q}_\theta * \mathbf{q}_\varphi = (q_\theta^w q_\varphi^w - \mathbf{n}_\theta \cdot \mathbf{n}_\varphi, q_\theta^w \mathbf{n}_\varphi + q_\varphi^w \mathbf{n}_\theta + \mathbf{n}_\theta \times \mathbf{n}_\varphi) \tag{5.9}$$

which can be conveniently expressed in terms of matrices multiplication between an orthogonal 4×4 matrix and a 4 dimensional vector:

$$\mathbf{q}_\theta * \mathbf{q}_\varphi = \mathbf{Q}(\mathbf{q}_\theta)\mathbf{q}_\varphi = \begin{bmatrix} q_\theta^w & -q_\theta^x & -q_\theta^y & -q_\theta^z \\ q_\theta^x & q_\theta^w & -q_\theta^z & q_\theta^y \\ q_\theta^y & q_\theta^z & q_\theta^w & -q_\theta^x \\ q_\theta^z & -q_\theta^y & q_\theta^x & q_\theta^w \end{bmatrix} \mathbf{q}_\varphi \tag{5.10}$$

As 3D rotations and matrix multiplications, quaternion multiplication is not commutative. The general form of the rotation matrix recovered from a unit quaternion noted $\mathbf{R}[\mathbf{q}]$ is given in Equation (4.4). As stated in [Horn, 1987], quaternions' multiplication is less expensive than 3×3 rotation matrices. In addition, since numerical computation has finite precision, the product of orthonormal matrices may no longer be orthonormal. The same problem may rise with unit quaternions. Nevertheless, finiding the closest unit quaternion is less difficult than finding the nearest orthonormal matrix.

Capture deviations from parallax-free or ideal pinhole camera model. In order to handle deviations from pure parallax-free motion of ideal pinhole camera model we improve the camera motion model by estimating a local motion estimation provided by a patch-based local matching procedure.

5.7 Global-to-Local Pair-wise Motion Estimation

In this section we describe the pair-wise motion estimation process. As seen in the previous section, the entire camera motion is parameterized by a global rotation and a local motion which allows to establish a list of patch matches and a global translational motion model over the entire image. This helps to deal with small amount of parallax and allows to integrate the pair-wise motion model within a BA process.

The entire global-to-local motion estimation is performed in a pyramidal fashion. Since the parallax is negligible, the motion process can be speeded up by performing the local alignment step only at the highest resolution level.

The motion estimation process follows four steps: (i) pyramid construction, (ii) patch extraction, (iii) motion estimation and (iv) coarse-to-fine refinement. At every level of the pyramid $l = 0, ..., L_{max}$ the goal is to find the 3D rotation $\mathbf{R}^l$. Since the same type of operation is performed at each level l, let us drop the superscript l through the following description.

5.7.1 Rigid rotation computation

Let $\mathbf{R}(\mathbf{q}_\theta, \mathbf{q}_\varphi, \mathbf{q}_\psi)^{init}$ be the initial guess provided by the pan-tilt head, where $(\theta, \varphi, \psi)_{hard}$ denote the pitch, roll and yaw angles, respectively expressed in the camera coordinate system. The optimal rotation is computed by varying the rotation parameters (θ, φ, ψ) within an homogeneous *pyramidal searching space*, $\mathcal{P}_{\mathcal{SS}}$, which is recursively updated at each pyramidal level. $\mathcal{P}_{\mathcal{SS}}$ is defined by the following parameters: θ range $\Delta\theta$, φ range $\Delta\varphi$, ψ range $\Delta\psi$ and their associated searching steps, $\delta\theta, \delta\phi, \delta\psi$.

The rotation angles are computed by applying rotations $\mathbf{R}_{(\theta,\varphi,\psi)}, (\theta, \varphi, \psi) \in \mathcal{P}_{\mathcal{SS}}$ to the 3D rays of recovered from pixels belonging to I_2 and matching the corresponding transformed pixels with pixels from I_1. For a given rotation $\mathbf{R}_{(\theta,\varphi,\psi)}, (\theta, \varphi, \psi) \in \mathcal{P}_{\mathcal{SS}}$ we can map pixels $\mathbf{u}_2$ from I_2 in the I_1's space using the warping equation expressed in Equation 5.6.

$$\hat{\mathbf{u}}_1 \cong \mathbf{K}\mathbf{R}_{(\theta,\varphi,\psi)\in\mathcal{P}_{SS}}\mathbf{K}^{-1}\mathbf{u}_2 \tag{5.11}$$

We obtain the rotated pixel from I_2 warped in the I_1's space which yields an estimate of I_1, noted $\hat{I}_1$. The goal is to find the optimal rotation which applied to pixels from I_2 and warped in the I_1's space minimizes the difference in brightness between the template image I_1 and its estimate, $\hat{I}_1(\mathbf{u}_2; \mathbf{R}_{(\theta,\varphi,\psi)})$.

For the first experimental test of the Rodeon® platform the camera was used in an automatic mode, meaning that the flash could fire automatically while focusing. However, blur effects were introduced due to the fact that the imaging device failed to focus low-depth scenes. Moreover, our research work deals with the mosaicing problem in complex environments where depth varies chaotically leading to considerable light variations which poses difficulties to the image matching process. In order to deal with this issue, the Rodeon® platform was improved by our research lab in order to provide the possibility to acquire the same image with different flash values and focal lengths.

Since images belonging to the same mosaic node are subject to different flash values, we employ the Zero Normalized Cross Correlation score to measure the similarity robustly wrt illumination changes. For each pixel, the score is computed over each pixel's neighborhood defined as $\mathcal{W} = [-w_x, w_x] \times [-w_y, w_y]$ centered around $\mathbf{u}_2$ and $\hat{\mathbf{u}}_1$ respectively, of size $(2w_x + 1) \times (2w_y + 1)$, where $w = w_x = w_y$ denote the neighborhood ray. The similarity score $\mathcal{Z}$ is given in Equation (5.12), being defined on the $[-1, 1]$ domain and for high correlated pixels is close to the unit value.

$$-1 \leq \mathcal{Z}(I_1(\mathbf{u}), I_2(\hat{\mathbf{u}})) = \frac{\sum_{\mathbf{d}\in\mathcal{W}}[I_1(\mathbf{u}+\mathbf{d}) - \bar{I}_1(\mathbf{u})][I_2(\hat{\mathbf{u}}+\mathbf{d}) - \bar{I}_2(\hat{\mathbf{u}})]}{\sqrt{\sum_{\mathbf{d}\in\mathcal{W}}[I_1(\mathbf{u}+\mathbf{d}) - \bar{I}_1(\mathbf{u})]^2 \sum_{\mathbf{d}\in\mathcal{W}}[I_2(\hat{\mathbf{u}}+\mathbf{d}) - \bar{I}_2(\hat{\mathbf{u}})]^2}} \leq 1 \tag{5.12}$$

The global similarity measure is given by the mean of all the similarity scores computed for all the patches belonging to the overlapping region. For rapidity reasons, we correlate only border patches extracted in the overlapping regions.

$$\mathbf{E}[\mathbf{R}_{(\theta,\varphi,\psi)}] = \frac{1}{N_w}\sum_{j=0}^{N_w-1}\Phi_j\mathcal{Z}(I_1(\mathbf{u}^j), I_2(\hat{\mathbf{u}}^j_{\mathbf{R}_{(\theta,\varphi,\psi)}})) \tag{5.13}$$

Φ_j defines a characteristic function which takes care of "lost" (i.e. the pixel falls outside of the rectangular support of I_2) ($[0, n_u - 1] \times [0, n_v - 1]$) and "zero" pixels (i.e. missing data

either in $I_1(\hat{\mathbf{u}}^j_{\mathbf{R}})$ or $I_2(\hat{\mathbf{u}}^j_{\mathbf{R}})$, which may occur when mapping pixels $\hat{\mathbf{u}}^j_{\mathbf{R}}$ in the I_2's space. Thus, we penalize "lost" and "zero" pixels using the following weighting function:

$$\Phi_j = \begin{cases} 0, & \text{if } I_1(\hat{u}^j, \hat{v}^j) = 0 \text{ or } I_2(\hat{u}^j, \hat{v}^j) = 0 \\ 0, & \text{if } \hat{u}^j, \hat{v}^j < 0 \text{ or } \hat{u}^j > n_u - 1 \text{ or } \hat{v}^j > n_v - 1 \\ 1, & \text{otherwise} \end{cases} \tag{5.14}$$

N_w denotes the number of valid pixel matches founded between I_1 and I_2 for which $\Phi_j = 1$ and defines the overlapping area $\mathbf{O}[\mathbf{R}_{(\theta,\varphi,\psi)}]$ of the corresponding rotation $\mathbf{R}_{(\theta,\varphi,\psi)}$. The global dissimilarity score $\mathbf{E}(\mathbf{R}_{(\theta,\varphi,\psi)})$ is defined on the interval $[0, 1]$.

The optimal rotation $\hat{\mathbf{R}}_{(\theta,\varphi,\psi)}$ is obtained by maximizing the global similarity score $\mathbf{E}[\mathbf{R}_{(\theta,\varphi,\psi)}]$ over the entire searching area $\mathcal{P}_{SS}$.

$$\hat{\mathbf{R}}_{(\theta,\varphi,\psi)} = \arg \max_{(\theta,\varphi,\psi)\in\mathcal{P}_{\mathcal{SS}}} \mathbf{E}[\mathbf{R}_{(\theta,\varphi,\psi)}] \tag{5.15}$$

5.7.2 Non-rigid Motion Estimation

In order to handle deviations from pure-parallax motions or from ideal pinhole camera, we use the rotationally aligned images to perform the local patch matching. Since deviations from parallax-pure motion are negligible (i.e. less than $\frac{1}{4}$ pixels at level $l < 0$) shadowing the failure to rotate the camera around its optical center, we can speed up the process by computing the local motion directly at the highest resolution level, $l = 0$. For this reason, the optimal rotation produced at the highest resolution $\mathbf{R}^0_{(\theta,\varphi,\psi)}$ level initializes local motion estimation procedure.

Let $\mathbf{P}_1 = \{\mathcal{P}(\mathbf{u}^k_1)|\mathbf{u}^k_1 \in I_1, k = 1, ..., N_1\}$ and $\mathbf{P}_2 = \{\mathcal{P}(\mathbf{u}^k_2)|\mathbf{u}^k_2 \in I_2, k = 1, ..., N_2\}$ be the patches extracted in image I_1 and I_2 respectively, which are defined by a neighborhood $\mathcal{W}$ centered around $\mathbf{u}^k_1$ and $\mathbf{u}^k_2$ respectively. For each patch $\mathcal{P}(\mathbf{u}^k_1) \in \mathbf{P}_1$ we search for its optimal match in I_2 by exploring a windowed area $\mathbf{W}^{\mathbf{SA}}(\mathbf{u}^k_2; \hat{\mathbf{R}})$ centered around $(\mathbf{u}^k_2; \hat{\mathbf{R}})$, where $\mathbf{SA}$ denotes the searching area ray.

Let $\mathbf{P}^{k,\mathbf{SA}}_2 = \{\mathcal{P}(\mathbf{u}^m_2)|\mathbf{u}^m_2 \in \mathbf{W}^{\mathbf{SA}}(\mathbf{u}^k_2; \hat{\mathbf{R}}) \subset I_2, m = 1, .., M\}$ be M patches extracted from the warped image's searching area centered around $(\mathbf{u}^k_2; \hat{\mathbf{R}})$, with 1-pixel steps. For each patch $\mathcal{P}(\mathbf{u}^m_2)$ we compute the similarity score $\mathcal{Z}(I_1(\mathbf{u}^k), I_2(\mathbf{u}^m))$ and we perform a bicubic fitting in order to produce the best match with a subpixel accuracy and real time performances. The best match is obtained by maximizing the similarity score $\mathcal{Z}$ over the entire searching area $\mathbf{W}^{\mathbf{SA}}$.

$$\mathcal{P}(\hat{\mathbf{u}}^k_2) = \arg \max_{\mathbf{u}^m_2 \in \mathbf{W}^{\mathbf{SA}}(\mathbf{u}^k_2; \hat{\mathbf{R}})} \mathcal{Z}(I_1(\mathbf{u}^k), I_2(\mathbf{u}^m_2)) \tag{5.16}$$

In order to handle "lost" or "zero" pixels, patch matches corresponding to uncomplete warped patches are discarded. This yields a list of matched patches $\mathcal{P}(\mathbf{u}^k_1)$ and $\mathcal{P}(\hat{\mathbf{u}}^k_2)$ which gives the possibility to compute a local translational model for each patch: $\mathbf{t}^k = \|\mathbf{u}^k_1 - \hat{\mathbf{u}}^k_2\|$ and compensates eventual parallax motions or deviations from the ideal pinhole camera model. Moreover, the local motion allows the possibility to establish a mean translational motion model over the entire image space, noted $\bar{\mathbf{t}}$. The list of the patch matches are further injected into a bundle adjustment engine for multi-view fine alignment and gap closure.

5.7.3 Pyramidal Refinement

Suppose that a global maximum similarity score was localized at the lowest resolution level L_{max}, producing a coarse global rotation estimation $\hat{\mathbf{R}}_{(\theta,\varphi,\psi)}^{Lmax}$. The motion refinement is performed at higher resolution levels $l = L_{max} - 1, .., 0$ by searching within a 4×4 pixels neighborhood around the global optimum center sampled with 1-pixel step. Although pixels are considered to be squared during the camera calibration process, in practice we employ the FOV dimension and the size image to compute the angular steps corresponding to a pixel in both directions, being given by $\delta u_x = \frac{H_{FOV}}{n_u}$ and $\delta u_y = \frac{V_{FOV}}{n_v}$. Therefore, the $\mathcal{P}_{SS}$ is recursively updated using the following scheme:

$$\mathcal{P}_{SS}^{l-1} = \begin{cases} \Delta\theta^{l-1} = 4\delta u_x^{l-1} = \delta u_x^l \\ \Delta\varphi^{l-1} = \Delta\varphi^{l-1} = 4\delta u_y^{l-1} = \delta u_y^l \\ \delta\theta^{l-1} = \delta u_x^{l-1} \\ \delta\varphi^{l-1} = \delta\psi^{l-1} = \delta u_y^{l-1} \end{cases} \tag{5.17}$$

5.7.4 Experimental Results & Performance Evaluation

We tested the pair-wise image alignment process wrt different applications. The first application is the main concern in this dissertation which aims at generating in-situ Giga-mosaics to produce texture maps for photorealist in-situ 3D modeling. The second application consists in performing mosaic-based city mapping using a 10-camera network mounted on a vehicle to allow virtual traveling on the world wide web and to extract high-level semantics from panoramic imagery [iTowns, 2008].

Being given each application's context, different constraints are imposed to the mosaicing process. In this section we show that the proposed pair-wise alignment algorithm, which is the main ingredient of the mosaicing process, copes with both applications' constraints. To this end, different acquisition scenarios include unstructured and structured areas, being undertaken in indoors and outdoors environments. For each acquisition scenario different testbeds were deployed, one of them being fixed and the other mobile. In addition, different camera sensors were employed, each one being calibrated beforehand.

5.7.4.1 Unstructured and Underground Environments

Since our research work focuses on unstructured and difficult to access environments, experimental tests were undertaken in two prehistoric caves situated in France. First experiments were performed in the Tautavel prehistoric cave using the motorized pan-tilt heading delivering the theoretical rotations. The Rodeon® platform was upgraded to deliver a closer initial estimate and deployed in Mayenne Science prehistoric cave. We present hereafter the results obtained for each trial.

Experiment performed in Tautavel prehistoric cave, France. Figures 5.10 and 5.11 illustrate the results obtained by running the global-to-local image motion estimation procedure on an image pair using the testbed presented in Section 5.3. In order to evaluate our technique with respect to a feature-based method, Figures 5.10 and 5.11 show the results obtained on an image pair for which the SIFT detection and matching failed due to blur effects introduced by the long focal length and low-depth scenery (i.e. the cave's wall in this example).

Global rotation estimation. The global rotation estimation illustrated in Figure 5.10 uses a 6-levels pyramidal structure in which 1 pixel goes from 0.00875° up to 0.28° for

levels $l = 0, ..., 5$ respectively. The rotation computation starts at the lowest resolution level $L_{max} = 5$, where a fast searching is performed by exploring a searching space of 5° with 1-pixel steps in order to localize the global maximum, as shown in Figure 5.10(c). The coarse estimation is refined at higher resolution levels $l = L_{max} - 1, .., 0$ by exploring a searching area of 4 pixels around the global maximum using 1-pixel steps.

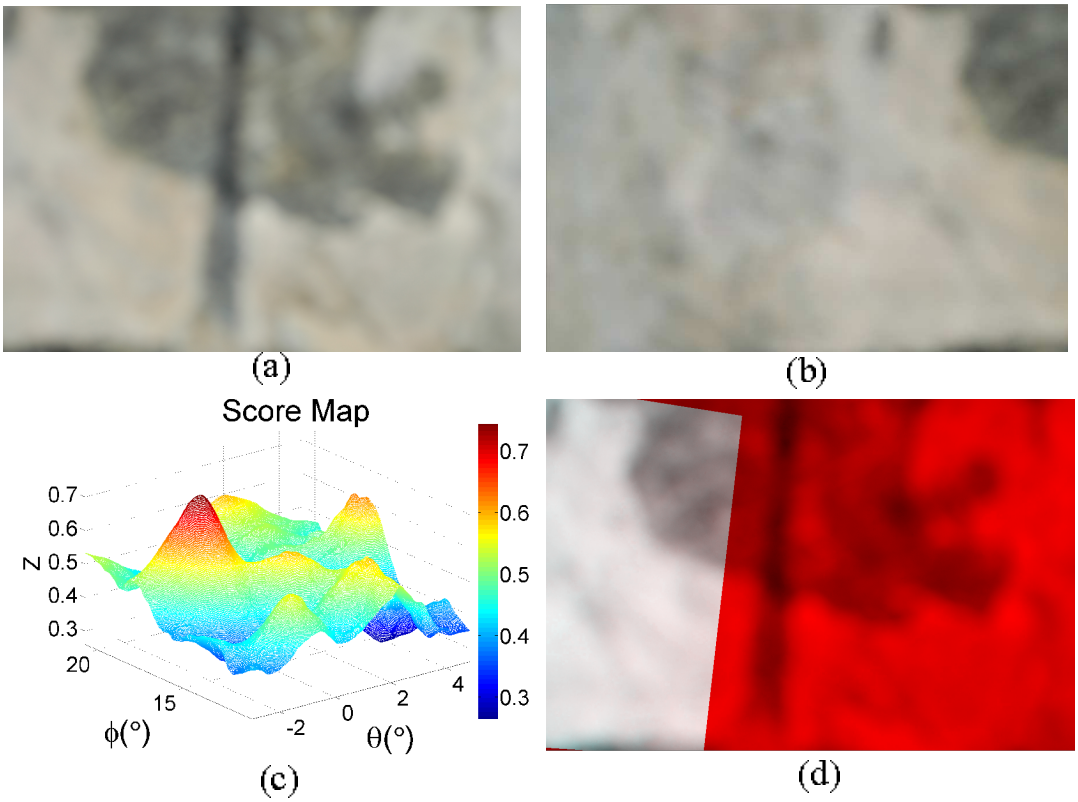


Figure 5.10: Test performed in Tautavel prehistoric cave. Rigid rotation estimation procedure. (a) I_1 - template image, (b) I_2 - target image, (c) global maximum localization at level $L_{max} = 5$, (d) superposed rotationally aligned images at level $l = 0$: I_1-red channel, the warped image $I_2(\mathbf{u}; \hat{\mathbf{R}})$-green channel, $\hat{\mathbf{R}}_{(\theta,\varphi,\psi)} = (17.005°, 0.083°, 0.006°)$.

Non-rigid motion estimation. Since deviations from parallax-pure motion are negligible we speed up the process by computing the local motion directly at the highest resolution level, $l = 0$. Figure 5.11 illustrates the results of the local estimation procedure which matches patches with a ray of $w = 15$ pixels by exploring a searching area of $\mathbf{SA} = 32$ pixels with 1-pixel steps. For rapidity, a bicubic fitting procedure is used to compute the maximum correlation score with subpixel accuracy. For illustrative purposes only, in this experiment the patch location yielding the maximum correlation was interpolated using 0.005 sampling step.

The local translational motion estimated for each patch $\mathbf{t}^k$ allows to compute a mean translation motion model to be used for compensating the parallax motion. After translation compensation, the camera motion consists of purely rotations and therefore, the optimal rotation minimizes the angle between the corresponding 3D rays of each match pair given by $\hat{\gamma} = (\mathbf{v}_1^k, \mathbf{v}_2^k)$, where $\mathbf{v} = \mathbf{K}^{-1}\tilde{\mathbf{u}}$, where $\tilde{\mathbf{u}}$ denotes the image point in homogeneous coordinates.

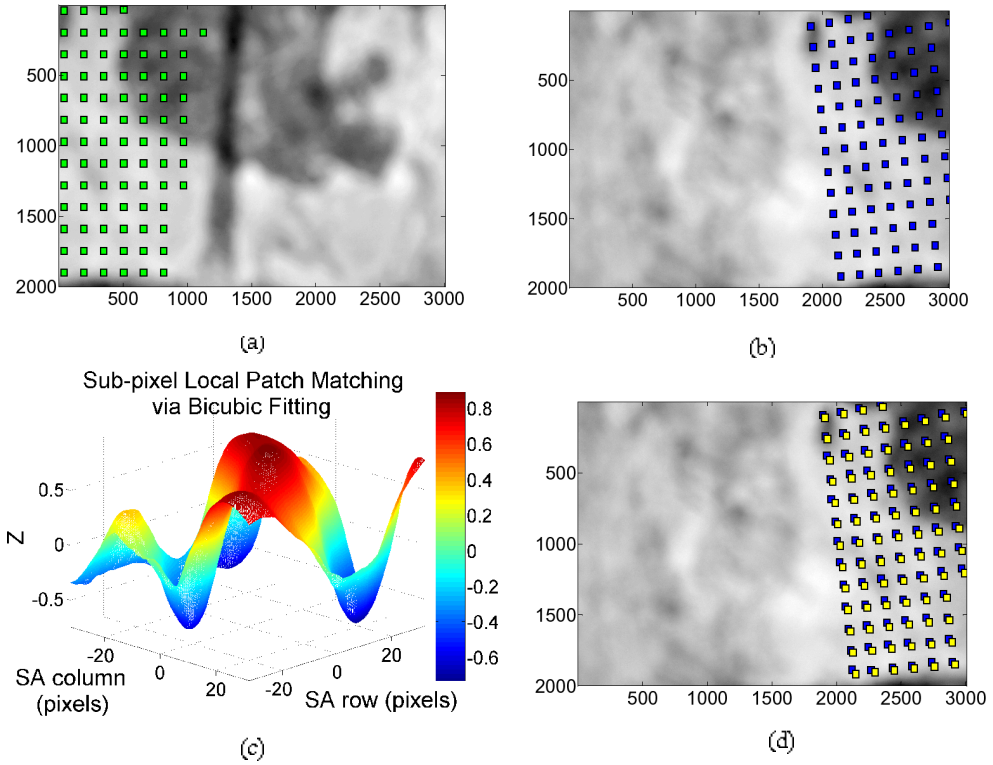


Figure 5.11: Test performed in Tautavel prehistoric cave. Anonymous Features extraction and matching procedure. $w = 15$ pixels, 85 AF matches. (a)$\mathcal{P}(\mathbf{u}_1^k)$, (b)$\mathcal{P}(\mathbf{u}_2^k)$ extraction in I_2 using the rotation initialization, (c)Bicubic fitting for an arbitrary patch: $\mathbf{SA} = 32$ pixels, sampling step: 0.005 pixels, (d)AF locally matched: $\mathcal{P}(\mathbf{u}_2^k)$ blue, $\mathcal{P}(\hat{\mathbf{u}}_2^k)$ yellow, $\overline{\mathbf{t}} = [1.6141, 1.0621]$ pixels.

Residual Errors. Table 5.1 illustrates the residual mean square error ($\bar{\mathbf{r}}$) and the standard deviation ($\sigma_{\mathbf{r}}$) of the pair-wise motion estimate $[\hat{\mathbf{R}}, \overline{\mathbf{t}}^k]$ computed with two different criteria: (a) the projection error in the 2D space given in Equation 5.18 and (b) the angle between the 3D-rays corresponding corresponding to AF matches given by their cross product expressed in Equation 5.19.

The second row of Table 5.1 verifies the rotation estimation correctness showing that the angular distance $RMS_{\times 3D}$ between the non-aligned images (first column) corresponds to the optimal rotation estimate, $\hat{\mathbf{R}}_{ij}$.

$$\bar{\mathbf{r}}_{2D} = \frac{1}{N} \sum_{k=0}^{k=N-1} \|\mathbf{u}_i^k - \mathbf{K}\hat{\mathbf{R}}_{ij}^T \mathbf{K}^{-1}(\hat{\mathbf{u}}_j^k - \mathbf{t}^k)\| \tag{5.18}$$

$$\bar{\mathbf{r}}_{\times 3D} = \frac{1}{N} \sum_{k=0}^{k=N-1} \|\mathbf{v}_i^k \times \hat{\mathbf{R}}_{ij}^T \mathbf{K}^{-1}(\hat{\mathbf{u}}_j^k - \mathbf{t}^k)\| \tag{5.19}$$

Rendering. Since Autopano Pro did not detect SIFT matches, we have injected the detected AF matches in order to obtain the rendering of the aligned images. Figure 5.12 illustrates the rendering of the merged images.

$\bar{\mathbf{r}} \pm \sigma_{\mathbf{r}}$	no model	$\bar{\mathbf{t}}$ compensation	$[\hat{\mathbf{R}}, \bar{\mathbf{t}}]$ model
RMS_{2D}(pixels)	1989.68 ± 62.83	1988.05 ± 62.74	0.08 ± 0.01
$RMS_{\times 3D}(^\circ)$	16.99 ± 0.51	16.98 ± 0.5	$(7 \pm 1) \times 10^{-4}$

Table 5.1: Trial Tautavel - Residual Error Measures. $\hat{\mathbf{R}}_{(\theta,\varphi,\psi)} = (17.005^\circ, 0.083^\circ, 0.006^\circ), \bar{\mathbf{t}} = [1.614, 1.062]$ pixels.

Figure 5.12: Test performed in Tautavel prehistoric cave. Rendering of the aligned images using the optimal rotation estimate $\hat{\mathbf{R}}$ and the AF features injected in the Autopano Pro 1.4.2. Mosaic size: 5658×2074 (pixels), mosaic FOV: $[50.82^\circ \times 18.63^\circ]$.

Tests using the improved platform in Mayenne Science prehistoric cave. Several problems were encountered during the first acquisition scenario performed in Tautavel prehistoric cave. The main one is concerned with the erroneous initial estimation delivered by the motorized platform. In addition, the synchronization between the platform's rotations and camera shooting required long shooting delay.

Upgraded-Rodeon®. The motorized platform was improved with a control software as a mean for parameterizing the acquisition with the number of views to be acquired in order to cover the entire sphere. This step establishes the rotation angles $(\theta, \varphi) \in [0^\circ, 360^\circ] \times [-90^\circ, 90^\circ]$ to be performed by the platform which are further exploited by the pair-wise alignment algorithm. For this trial, images with a lower overlapping area were acquired (15%), reducing the acquisition time with a factor of 4.

In complex environments depths scene changes suddenly, leading to light changes and thus one must employ different flash values in order to allow reliable image matching. For this reason, the control software allows to capture the same image with different flash values. Consequently, during the rendering step it is possible to select the most suitable one, taking into account changes light variations. Also, the new platform design allows to send data directly to the processing unit on-the-fly, allowing for on-line processing.

Pair-wise alignment. A new acquisition scenario was undertaken in the Mayenne Science prehistoric cave using the improved motorized platform. We test the pair-wise alignment procedure on the images pair illustrated in Figures 5.13 (a) and (b). Figures 5.13(c)-(e) illustrates the alignment obtained using the initial orientation provided by the physical instrumentation, exhibiting a much better estimate.

Figure 5.14 (a) depicts the aligned images using the global rotation estimate computed at level L_{max} initialized by the new estimate provided by the acquisition platform. Since

the initial estimation exhibits a good approximation of the global rotation, the pair-wise alignment process can skip the global motion estimation and jump directly to the local matching stage. This emphasizes the main impact of the improved platform which leads to considerable computational savings, enabling therefore the in-situ generation of Gigapixel mosaics. Figure 5.14 (b) shows the initialization and the local patch matching procedure using the global rotation from Figure 5.14 (a), while Figure 5.14 (c) illustrates the final result of the local patch matching procedure performed at highest resolution level $l = 0$.

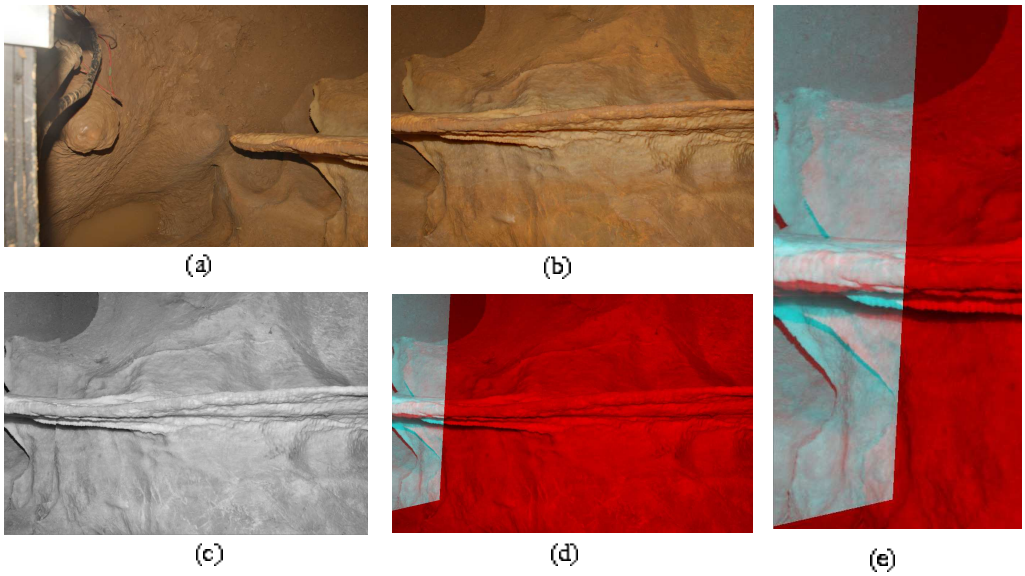


Figure 5.13: Test performed in Mayenne Science prehistoric cave. (a) I_1 - reference image, (b) I_2 - image to align, (c) I_2 transformed in the I_1's space using the initial guess provided by the upgraded-Rodeon ® $(\theta, \phi)_{hard} = (48°, -2°)$, (d) aligned images using $(\theta, \varphi)_{hard}$: I_1 red channel, $[I_2; \mathbf{R}_{hard}]$ transformed - greed channel, (e) zoom-in on the overlapping area illustrated in (d).

Residual Error. Table 5.2 illustrates the residual errors demonstrating that as for the Tautavel case the residual cross product agrees with the main rotation angle. The 2D translation exhibits a predominant motion in the vertical direction which can be visually verified by tacking a closer look at the patch extraction initialized by the optimal rotation illustrated in Figure 5.14 (b).

$\bar{\mathbf{r}} \pm \sigma_{\mathbf{r}}$	no model	$\bar{\mathbf{t}}$ compensation	$[\hat{\mathbf{R}}, \bar{\mathbf{t}}]$ model
RMS_{2D}(pixels)	2421.498 ± 24.558	2416.852 ± 24.529	$(18 \pm 0.22) \times 10^{-3}$
$RMS_{\times 3D}(°)$	47.942 ± 0.538	47.861 ± 0.532	$(1.19 \pm 2.86) \times 10^{-4}$

Table 5.2: Trial Mayenne - Residual Error Measures. $\hat{\mathbf{R}}_{(\theta,\varphi,\psi)} = (48.69°, -2.0443°, 3.274°), \bar{\mathbf{t}} = [3.198, 20.127]$ pixels.

AutopanoPro residual error. In contrast to the image pair exemplified in the Tautavel trial for which SIFT features were not detected, on the images shown in Figures 5.13 (a) and (b), SIFT features were detected and correctly matched. The first row of Table 5.3 illustrates the numerical results corresponding to the mosaic rendering depicted in Figure

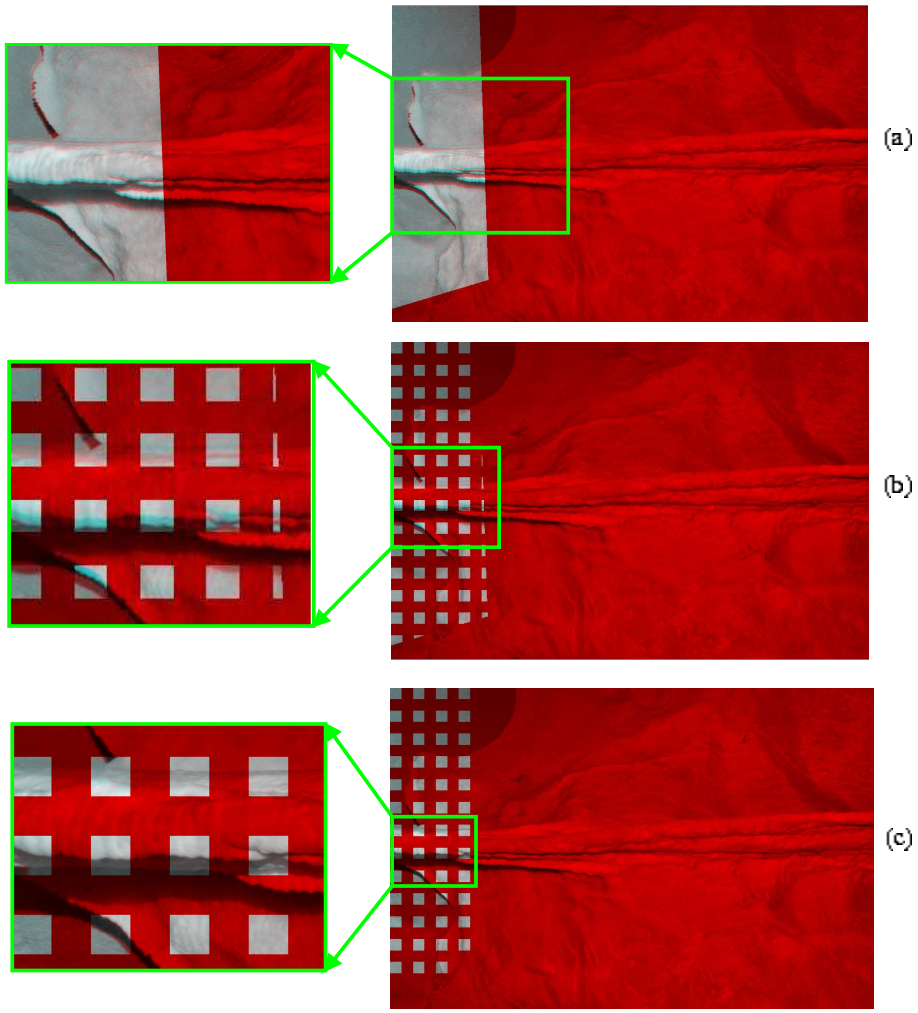


Figure 5.14: Test performed in Mayenne Science prehistoric cave. (a) the rotationally aligned images: I_1 - red channel superposed with $I_2[\mathbf{u};\hat{\mathbf{R}}]$ - green channel, $\hat{\mathbf{R}}_{(\theta,\varphi,\psi)} = [48.69°, -2.0443°, 3.324°]^T$, (b) patch extraction initialized by the optimal rotation $\hat{\mathbf{R}}$: patches extracted from $I_2(\mathbf{u};\hat{\mathbf{R}})$, $\mathcal{P}(\mathbf{u}_2^k)$ - green channel superposed with I_1 - red channel, (c) locally matched patches: $\mathcal{P}(\hat{\mathbf{u}}_\mathbf{2}^\mathbf{k})$ - green channel superposed with I_1 - red channel.

5.15. Although the rendering is visually coherent, the residual error has a relatively high magnitude. We evaluated the pose quality using different initial guesses in conjunction with either SIFT or AF features. The second row of Table 5.3 shows that the 2D residual error decreases when the matching process is initialized with the platform's pose, and more accurate results are obtained when AF are used instead SIFT features. Although the initial guess provided by the upgraded-Rodeon® combined with AF matches seems to increase the pose quality, the self-calibration process integrated within the pose estimation scheme leads to different intrinsic parameters, increasing the residual error.

Figure 5.15: Test performed in Mayenne Science prehistoric cave. Rendering of the aligned images using Autopano Pro 1.4.2.

Method	# features	RMS_{2D} (pixels)	Mosaic size (pixels)	Mosaic FOV(°)
AutoPano Pro	41	2.81	5541×2274	94.19×38.65
Rodeon® & SIFT	41	2.74	5523×2271	92.89×38.21
Rodeon® & AF	51	2.50	5634×2324	108.92×44.93

Table 5.3: Test performed in Mayenne Science prehistoric cave. AutopanoPro evaluation of the pair-wise matching results for image couple illustrated in Figures 5.13 (a), (b) using the initial guess provided by the upgraded-Rodeon® with SIFT and AF matches.

After testing the pair-wise method on two different prehistoric caves, using different experimental setups, we can conclude two main aspects:

- Trial Tautavel using Rodeon®: SIFT matches were not detected due mainly to blur caused by long focal length and low-depth scenery. We injected AF matches into Autopano Pro to refine pose and produce rendering;

- Trial Mayenne-Science using the upgraded-Rodeon®: each image was acquired with different flash values in order to deal with illumination changes, without using the camera in an automatic focus mode. This improved considerably the image matching task and SIFT matches detected. We test whether better results can be obtained using upgraded-Rodeon® pose and AF pairings. During this test a first doubt on Autopano Pro capabilities was raised. We conclude that there is a high probability for the self-calibration process to cause an erroneous multi-view alignment process.

5.7.4.2 Tests in Outdoor Structured Environments

In order to evaluate the algorithm's robustness wrt different mosaicing scenarios, several tests were undertaken in outdoor structured environments. The main purpose of the mosaicing system is to supply terrestrial city mapping applications using a vehicle equipped with multiple 2D and 3D imaging devices.

STEREOPOLIS$_{©IGN}$: a mobile system for terrestrial mapping. The mobile platform embeds both stereo- and panoramic-imaging montages, as well as 2D and 3D laser range finders. Figure 5.16 (a) illustrates the STEREOPOLIS$_{©IGN}$ prototype designed by MATIS research lab at French Mapping Agency[2] to supply terrestrial city mapping applications.

Onboard terrestrial mosaicing. In order to capture panoramic imagery, a 10-HD camera network is mounted onboard STEREOPOLIS$_{©IGN}$, as shown in Figure 5.16 (b). The image acquisition is synchronized with the vehicle's speed, i.e. 4 fps, in order to deliver one panoramic image to each 4 meters. This reduces considerably the overlapping area to 1%. In this case, it is possible for a feature-based method to fail in finding primitive matches in such a reduced overlapping area, emphasizing therefore the major advantage of the correlation-based approach employed within the proposed pair-wise alignment scheme.

When running Autopano Pro on image pairs acquired by the panoramic montage, SIFT's detection failed. Consequently, several tests were performed to analyze the behavior of the pair-wise matching procedure when running on images acquired by the different cameras composing the panoramic imagery montage. It is important to note that in this experiment images with a lower resolution are acquired, i.e. of size 1920×1080, in order generate panoramic imagery on-the-fly.

Small parallax. Comparing to trials presented in Section 5.7.4.1, in this case the cameras are separated by a few centimeters parallax, being therefore conveniently to either start the local matching from lower resolution levels or set a larger searching area ray $\mathbf{W^{SA}}$ at the highest resolution level.

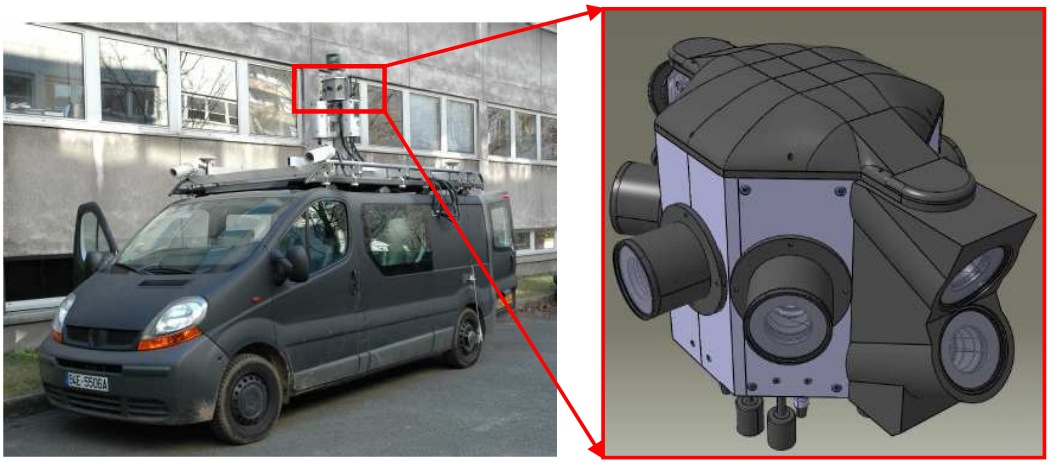
This section presents two tests performed on data acquired during two city mapping campaigns undertaken in Paris, France. The first trial is performed at Panthèon Square, while the second one in the 12^{th} district of Paris.

Trial Panthèon, Paris (France). The first trial is performed on two partially overlapped images acquired by two cameras with their positions highlighted in green in Figures 5.17 (a) and (b). The corresponding acquired images are illustrated in Figures 5.17 (c) and (d).

The initial pose provided by manual calibration initializes the global motion estimation stage of the pair-wise alignment. process. The initial guess is refined by exploring a pyramidal searching space initially set to $\mathcal{P}_{SS}^{L_{max}}$ which is reduced by searching for the main rotation angle θ within a range of $\Delta\theta = 6°$, while for φ and ψ a range of $\Delta\varphi = \Delta\psi = 0.5°$ is used. Figure 5.18 illustrates the results obtained using the estimated global rotation model $\hat{\mathbf{R}}_{(\theta,\varphi,\psi)}^{Lmax}$. The global motion initializes the local patch matching procedure for which the results are illustrated in Figure 5.19. Table 5.4 illustrates the residual errors quantified with two different criterions showing that the most sensible measure is the cross product measuring the angle between the 3D vector associated to the AF matches.

Trial Paris 12^{th} district. We present a second trial performed in the 12^{th} district of Paris. The image couple is illustrated in Appendix C.5, Figures C.2 (a) and (b). In Appendix C.5, Figures C.2 (c) and (d) illustrate the results obtained using the estimated

[2]Institut Geographique National www.ign.fr



(a) (b)

Figure 5.16: Tests performed in outdoor structured environments. (a) The STEREOPOLIS $_{©-IGN}$ mobile platform prototype designed by MATIS Research Lab, (b) the 10-HD camera network mounted onboard the vehicle designed to deliver panoramic imagery.

$\bar{\mathbf{r}} \pm \sigma_{\mathbf{r}}$	no model	$\bar{\mathbf{t}}$ compensation	$[\hat{\mathbf{R}}, \bar{\mathbf{t}}]$ model
RMS_{2D}(pixels)	1007.906 ± 11.354084	1006.14 ± 14.193	$(7.6 \pm 1.57) \times 10^{-3}$
$RMS_{\times 3D}(^\circ)$	39.004 ± 0.506	38.837 ± 0.498	$(0.368 \pm 0.09) \times 10^{-4}$

Table 5.4: Trial Panthèon - Residual Error Measures. $\hat{\mathbf{R}}_{(\theta,\varphi,\psi)} = (-10.871^\circ, 38.872^\circ, 1.734^\circ), \bar{\mathbf{t}} = [-6.382, 3.302]^T$ pixels.

global rotation model $\hat{\mathbf{R}}^{Lmax}_{(\theta,\varphi,\psi)}$. Note here the mirror effects visible in Figure C.2 (c) introduced by the cameras montage causing reflections. The global estimation initializes the local patch matching procedure for which the results are illustrated in Figure C.3. Since the local matching keeps only valid patches (without zero pixels), the patches extracted initially in the areas produced by the camera reflection are discarded.

5.7.4.3 Quality Assessment

This section analysis the performances of the above mosaicing system configurations wrt the main algorithms parameters. In order to emphasize the robustness of the pair-wise alignment process wrt different testbeds and acquisition scenarios, we summarize in Table 5.5 the characteristics of each trial.

Establishing the optimal patch value (w). In order to establish the optimal patch ray coping with all the aforementioned application contexts, we evaluate the residual error and the runtime of the local matching process wrt different patch ray values. Figure 5.20 shows that although the error does not vary significantly, it decreases slightly for all trials when a patch ray of $w = 15$ pixels is used. On the other hand, Figure 5.21 shows that the computation time increases as the patch ray w decreases. This is a normal behavior since for low values of w, the number of extracted patches increases and since for each patch, the algorithm searches for its optimal homologous patch within a searching area ray $\mathbf{W}^{\mathbf{SA}}$

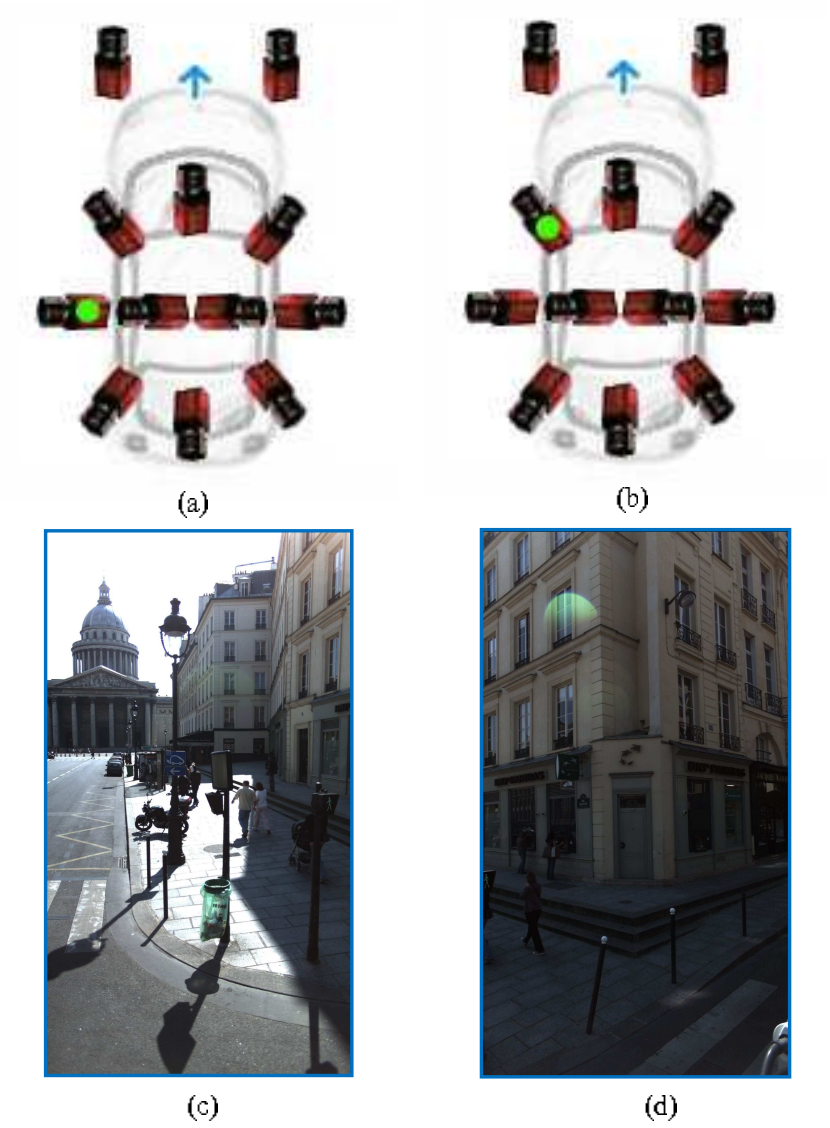


Figure 5.17: Test performed in outdoor structured environments. (a) camera no. 21, (b) camera no. 31, (c) I_1 - the reference image corresponding to camera no. 21, (d) I_2 - the image to align corresponding to the camera no. 31. Parallax $\approx$ 12 cm.

in the second image.

One might argue that a high patch ray value can yield a reliable matching. This is true when purely free-parallax rotation are encountered by the capturing device, which is not always the case in a real mosaicing scenario, since small parallax amount can be introduced when camera optical center is not superposed with the rotational center. In addition, in order to cope with poorly overlapped images, such as those acquired by the STEREOPOLIS$_{©IGN}$ platform, small patch rays need to be used in order to avoid the presence of patches laying out of the rectangular support of the image.

It can be observed in Figure 5.21 an offset runtime corresponding to each trial which is related to the overlapping values of each trial which are showed in Table 5.5. More precisely,

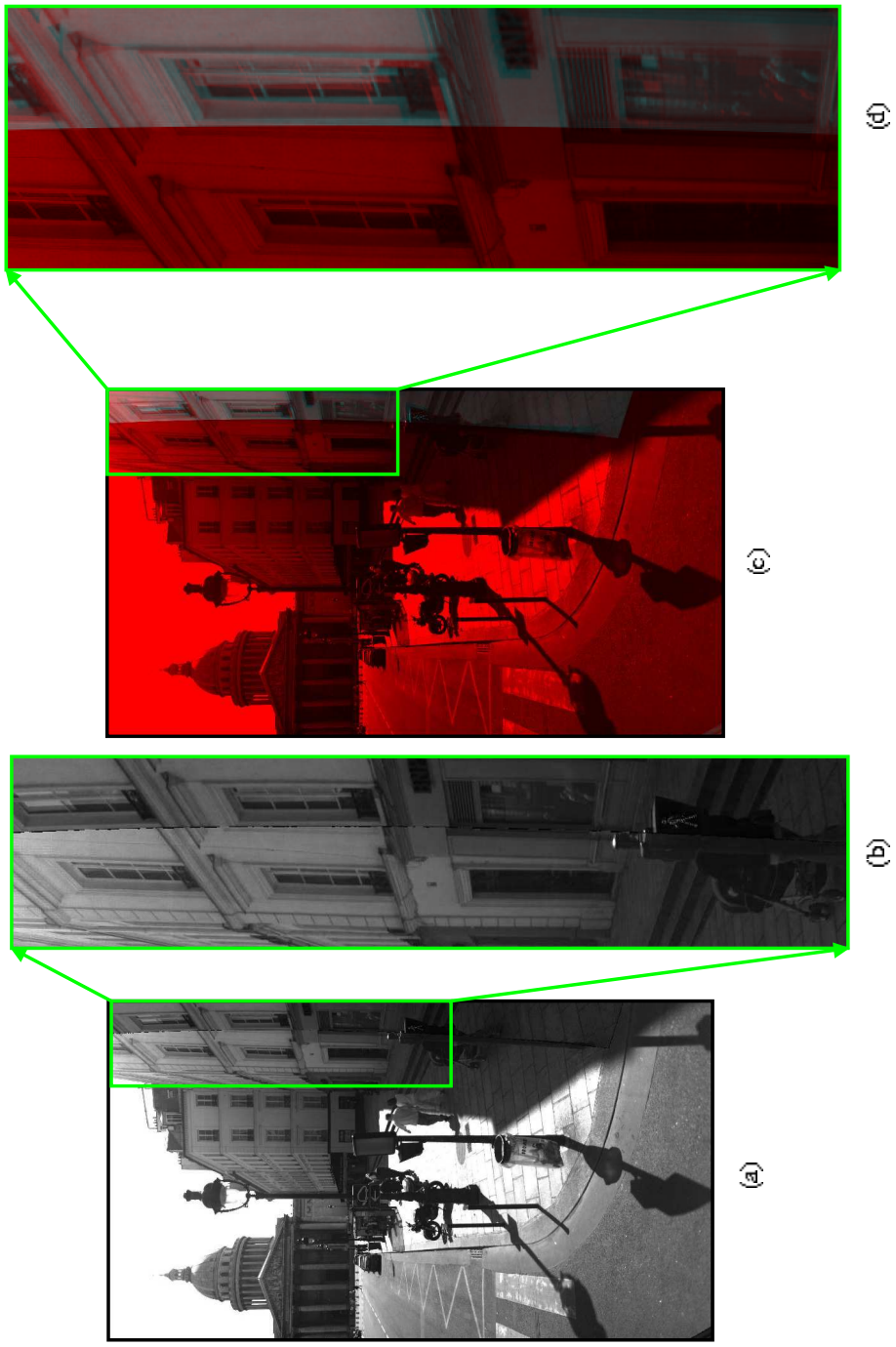


Figure 5.18: Trial Panthèon - Image alignment using the global rotation estimate - resolution L_{max}. (a) I_1 and $I_2(\mathbf{u};\hat{\mathbf{R}})$ superposed - grey levels, (b) zoom-in in the overlapped region of (a), (c) I_1-red channel superposed with the transformed image $I_2(\mathbf{u};\hat{\mathbf{R}})$ - green channel, (d) zoom-in in the overlapped region of (c).

for high overlap values, the number of the extracted patches increases and therefore, the runtime of the patch matching process increases too.

Establishing the optimal searching area ray ($\mathbf{W^{SA}}$). We set the patch ray value

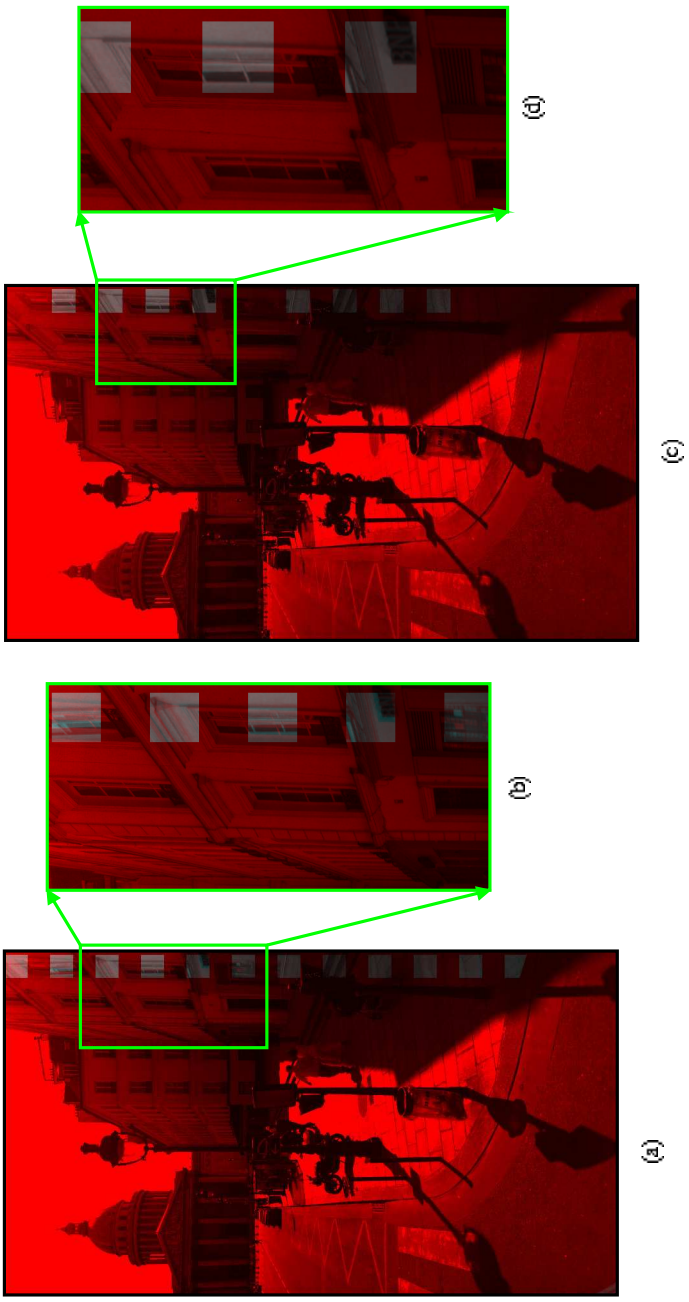


Figure 5.19: Trial Panthèon - local patch matching procedure. (a) patch extraction using the rotationally aligned images: I_1- red channel, patch extracted in $I_2[\mathbf{u};\hat{\mathbf{R}}]$ $\mathcal{P}(\mathbf{u}_2^k)$ - green channel, (b) zoom-in in the overlapping area of (a), (c) locally matched patches $\mathcal{P}(\hat{\mathbf{u}}_2^k)I_1$- green channel superposed with I_1 - red channel, (d) zoom-in in the overlapping area of (c).

to $w = 15$ and analyze the local matching process when different searching area rays $\mathbf{W}^{\mathbf{SA}}$ are employed. Figure 5.22 shows that for each trial, slightly different residual errors are obtained when the searching area ray varies. The RMS offsets corresponding to each trial are explained by the different numbers of patches matched for each trial, which increase

Trial vs. Features	Tautavel	Mayenne Science	Panthèon	Paris 12
Environment type	Unstructured underground		Outdoor structured	
Platform	fixed		mobile	
Testbed	Rodeon	Upgraded Rodeon	STEREOPOLIS	
Camera	NIKON D70		AVT MARLIN	
focal(mm)	50	20	variable	
Image size (pixels)	3000×2008		1920×1080	
Modeled distortion	R^3	R^3, R^5, R^7	R^3, R^5, R^7	
Overlap	33%	15%	$> 1\%$	
Parallax (mm)	negligible		> 32 mm	
Initial guess	NO	YES	YES	
SIFT detection	NO	YES	NO	

Table 5.5: Summary of different trials.

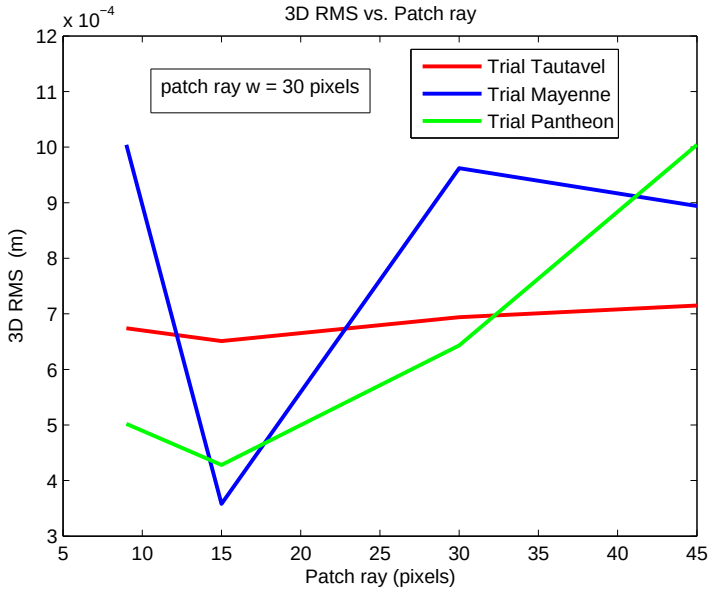


Figure 5.20: 3D RMS error vs. patch ray w.

with the overlapping area.

Figure 5.23 illustrates the computation time obtained when different searching areas ray are employed. In opposite to the patch ray study presented in Figure 5.21, runtime increases as large searching areas are used.

After this evaluation, one would choose a searching area ray of $\mathbf{W^{SA}} = 15$ pixels. Nevertheless, such a value cannot always cope with systems introducing large amounts of parallax, which is the case for STEREOPOLIS$_{©IGN}$. Consequently, we choose to use a patch ray of $w = 15$ pixels and a searching area ray of $\mathbf{W^{SA}} = 30$ pixels.

Computation time. Table 5.6 presents the computational time obtained for each trial. Note that these results are obtained when running an experimental version of the algorithm, without any optimization. We observe that the upgraded-Rodeon® improves the processing

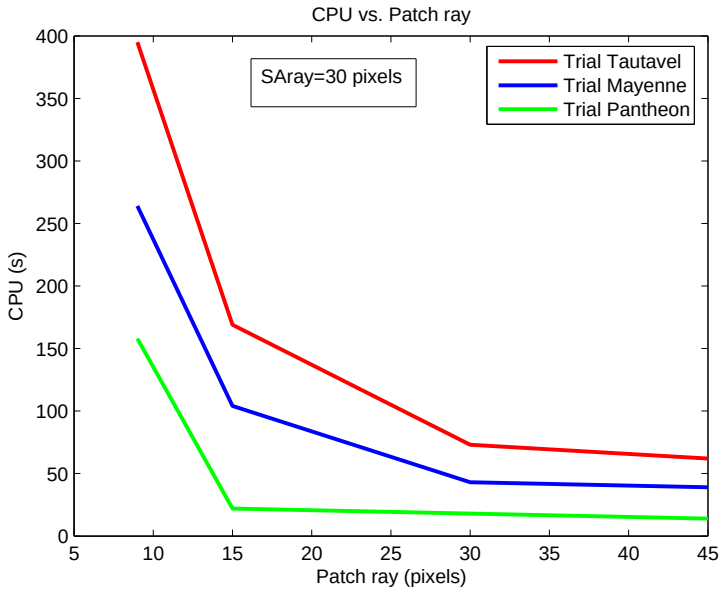


Figure 5.21: CPU vs. patch ray w.

time with a factor of 5.83.

Trial	Testbed	Global rotation estimation	Local patch matching	Total
Tautavel	Rodeon®	2 min 49 s	2 min 38 s	5 min 48 s
Mayenne Science	Upgraded Rodeon®	29 s	1 min 44	2 min 13 s
Panthèon	STEREOPOLIS$_{©-IGN}$	44 s	22 s	1 min 6 s

Table 5.6: Runtimes for global and local motion estimation steps for $w = 15$ pixels and $\mathbf{W^{SA}} = 30$ pixels.

Bounds on performances. The pair-wise image alignment algorithm is subject to two main factors: the quality of the initial estimation and the parallax amount. Table 5.7 illustrates several acquisition platform configurations taken in charge by the algorithm. Since prior knowledge on the acquisition setup is available, the algorithm can be adapted to each configuration in order to save computation time. For instance, when using the upgraded-Rodeon® platform, the close initial guess allows to skip the global alignment step for computational savings. Note that the proposed pair-wise alignment method is not adapted for hand-held mosaicing systems, since they introduce noisy initial guess and high parallax amounts.

5.8 Multi-view Fine Alignment

Given the pairwise motion estimates $\hat{\mathbf{R}}_{ij}$ and the associated set of AF matches $\mathbf{P}(i,j) = \{(\mathbf{u}_i^k \in I_i; \hat{\mathbf{u}}_j^k \in I_j) | j > i\}$, we refine the pose parameters jointly within a BA process [Triggs et al., 1999]. Two approaches are employed for the global optimization step, each

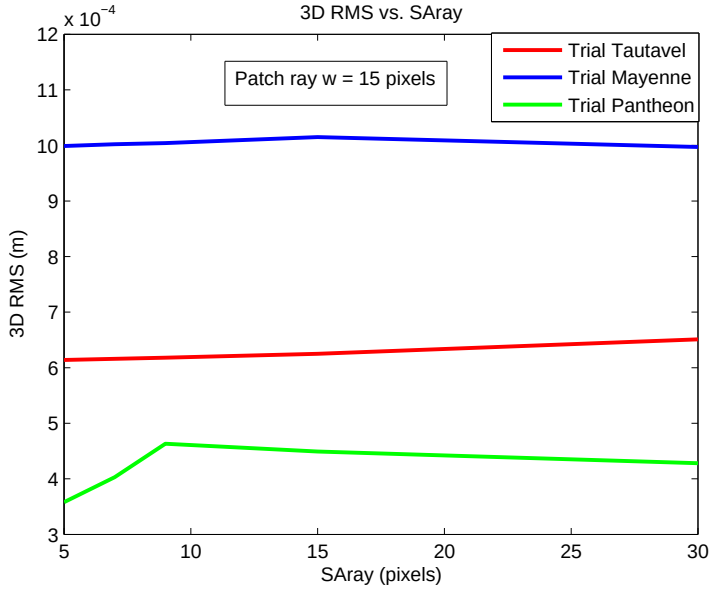


Figure 5.22: 3D RMS error vs. searching area ray $\mathbf{W^{SA}}$.

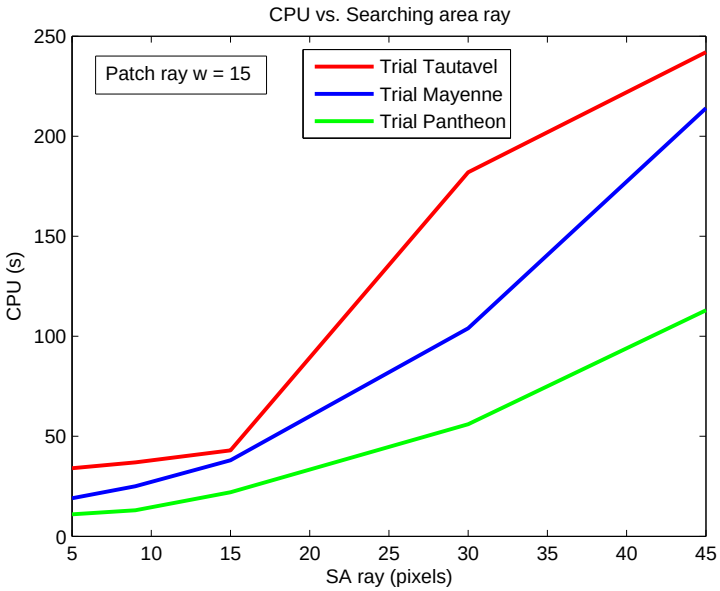


Figure 5.23: CPU time vs. searching area rays $\mathbf{W^{SA}}$.

Platform	Initial Guess		Parallax
	Source	Quality	
Rodeon® Upgraded-Rodeon®	physical instrumentation	noisy fine	negligible
STEREOPOLIS$_{©IGN}$	calibrated cameras montage	noisy	12 mm

Table 5.7: Different acquisition platforms configurations taken in charge by the global-to-local image alignment algorithm.

of which are making the object of the following two subsections.

5.8.1 Experimental Results using the Existent BA Solutions

In order to analyze the behavior of the existent BA schemes when consistent matches are injected into it, we run the BA step integrated within the Autopano Pro v1.4.2 [Kolor, 2005] by injecting AF pairings pre-computed by the proposed global-to-local pair-wise image alignment step described in Section 5.7.

As in [Brown and Lowe, 2007], the objective function is a robust sum squared projection error. Given a set of N AF correspondences $\mathbf{u}_i^k \longleftrightarrow \hat{\mathbf{u}}_j^k, k = 0, .., N-1$ the error function is obtained by summing the robust residual errors over all images:

$$e = \sum_{i=1}^{n} \sum_{j \in I(i)} \sum_{k \in \mathbf{P}(i,j)} h(\mathbf{u}_i^k - \mathbf{K}\hat{\mathbf{R}}_{ij}^T \mathbf{K}^{-1} \hat{\mathbf{u}}_j^k) \tag{5.20}$$

where n is the number of images, $I(i)$ is the set of adjacent images to image I_i and $h(\mathbf{x})$ denotes the Huber robust error function [Huber, 1981] which is used for outliers' rejection. This yields a non-linear least squares problem which is solved using the Levenberg-Marquardt algorithm. A detailed description of this approach may be found in [Brown and Lowe, 2007].

Trial Tautavel. Since our research work is focused on generating *in situ* complete and photorealistic 3D models of complex and unstructured large-scale environments, the Gigapixel mosaicing system was placed in different positions in order to generate mosaics covering the entire site. We illustrate in this section three examples of high-resolution mosaic views acquired from different spatial poses of the system corresponding to the cave's entrance, center and background.

Autopano Pro. When using Autopano Pro with SIFT features, a small number of errors were encountered for the cave's entrance and center. In opposite, more problems were observed when attempting to stitch the images acquired in the cave's background, as shown in Figure 5.25 (a), for which SIFT detection and matching failed. This is mainly due to blur effects introduced by the acquisition device which failed to focus areas too close to the optical center.

Autopano Pro and AF matches. Figures 5.24 (a), (b) and 5.25 (b) show the mosaic results obtained by injecting the AF pairings into the BA procedure integrated within the AutopanoPro v1.4.2 which took in charge the rendering process using a spherical projection and a multi-band blending technique. The mosaic's high photorealist level is emphasized by a high-performance viewer which allows for mosaic visualization using 4-level of detail (LOD), as shown in Figures 5.24 (c)-(f).

Ghost effects. Although Figure 5.25 (b) shows that better results are obtained when feeding the BA process with AF pairings, Figure 5.25 (f) illustrates important ghost effects which could be caused by two accumulated error sources. On one hand, they can be introduced by the erroneous initial estimation delivered by the Rodeon® platform, causing a real difficulty in establishing a general pyramidal searching space P_{SS}^{Lmax} valid for all image couples. On the other hand, the self-calibration process integrated within the Autopano Pro scheme leads to high residual errors causing the rejection of valid AF matches. Moreover, the minimization of an error measured in the 2D image space may not quantify correctly the camera rotations. Since they are performed in the 3D space, in the next section we suggest that a 3D error metric could be more sensible to camera motion and therefore, capture more accurately the camera motions.

Table 5.8 illustrates several characteristics of each mosaic generated by Autopano Pro v1.4.2 when AF matches are used. During this first use of the Rodeon platform a high number of images was acquired, i.e. $N_{im} = 310$, in order to cover a fully spherical FOV using the acquisition scenario presented in Section 5.3. As shown in Table 5.8, $N_{station} < N_{im}$. This is mainly due to the fact that the camera was used in an automatic mode, meaning that for low depth scenes the camera failed to focus and skip shooting. Scenes which are likely to be skipped by the capturing device are situated in the region right underneath the camera.

Memory limitations. Since the existing viewers are designed for lower size inputs, memory limitations were encountered when displaying the mosaics at their full size. For this reason, the mosaics are reduced by a factor of 2 to allow visualization within a multi-LOD viewer. This is quite compromising since we can not afford to exploit the high LOD offered by the Gigapixel mosaic. We employed KrPano [KrPano, 2009] panorama viewer for displaying the mosaics using 4-LOD. Nevertheless, if the mosaics were displayable at their full size, a 8-LOD rendering could be possible.

When looking at the third row of Table 5.8, one could argue that the size of the resulted mosaics does not reach the Gigapixel order. Although, a fully spherical mosaic leads to a Gigapixel size, it is important to note that the main purpose of the proposed mosaicing method is to demonstrate the feasibility of generating HR-mosaic imagery by overcoming the nowadays image alignment algorithms. In addition, it worths pointing out that the mosaics presented in Figures 5.24 (a), (b) and 5.25 (b) are already reaching the limits of a general-use computer in terms of computational and memory resources.

Residual errors. The BA scheme includes a self-calibration step and minimizes an error measured in the 2D image space, causing the rejection of correct AF matches and leading to relatively high mis-registration errors, as shown by the fourth row of Table 5.8. In practice we observed that this shortcoming can be overcome by injecting a high number of AF matches. However, this may be costly and when a low number of matches are used, there is a high probability that all of them to be rejected, producing the BA's failure. Since we can not afford this risk, our first concern is to improve the multi-view fine alignment process by simultaneously computing the optimal quaternions using a criterion computed in the 3D space in order to reduce the residual error when using a minimum number of AF correspondences. To this end, the next section proposes an analytical solution for the multi-view fine alignment step.

Runtime. For this experiment we employed the original Rodeon® platform, i.e. without the improvements. Therefore, the searching range for the rotation refinement was considerably high, i.e. $\pm 5^{\circ}$, leading to a computationally expensive rotation estimation stage. The upgraded-Rodeon® reduces the computational time by a factor of 5.83 for an

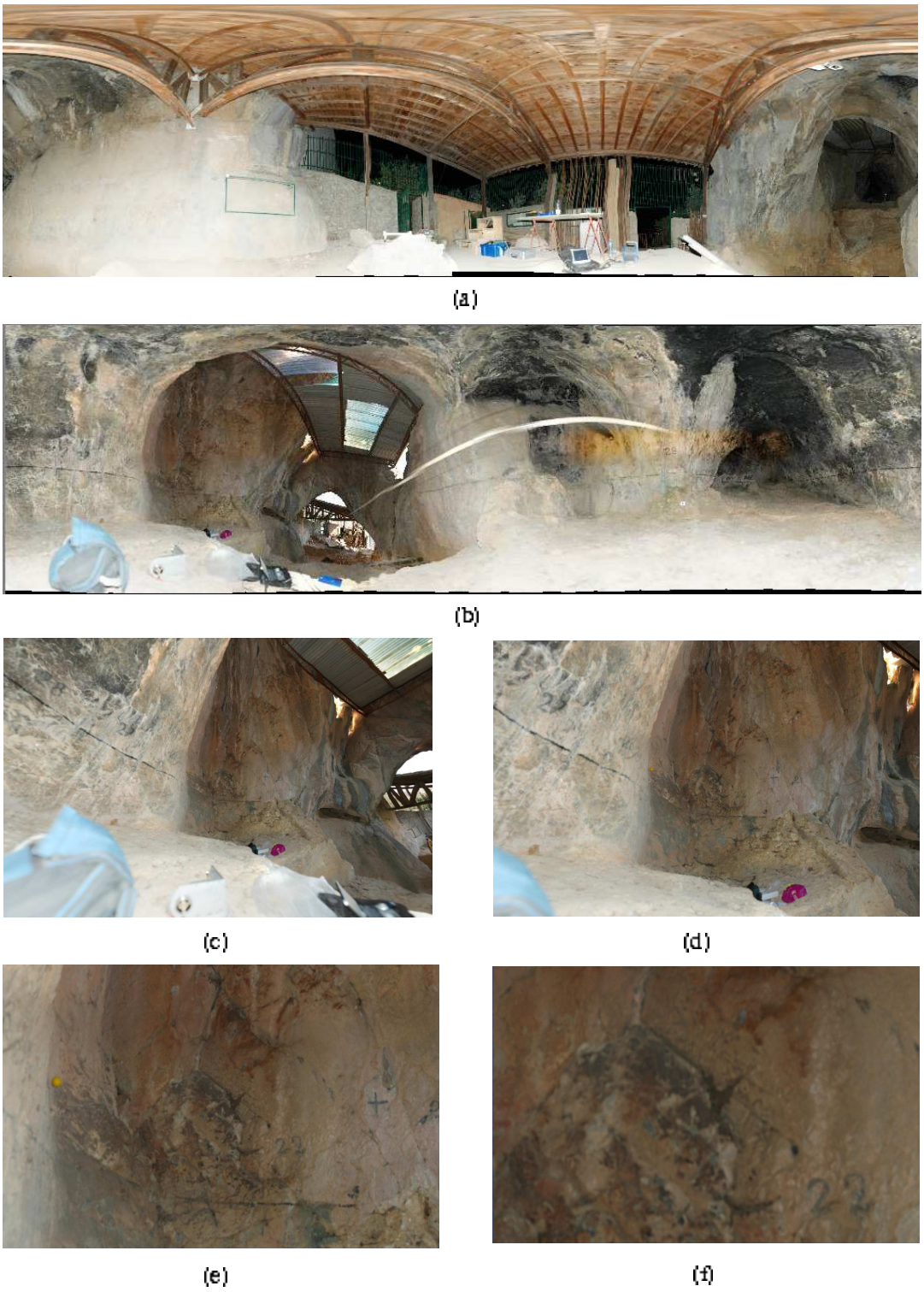


Figure 5.24: Mosaicing tests on data sets acquired in Tautavel prehistoric cave using the Rodeon® platform. The mosaics were generated by injecting the AF matches into the BA process integrated within Autopano Pro v1.4.2. (a) - cave's entrance, (b) - cave's center, (c)-(f) 4-LODs corresponding to the right part of mosaic (b).

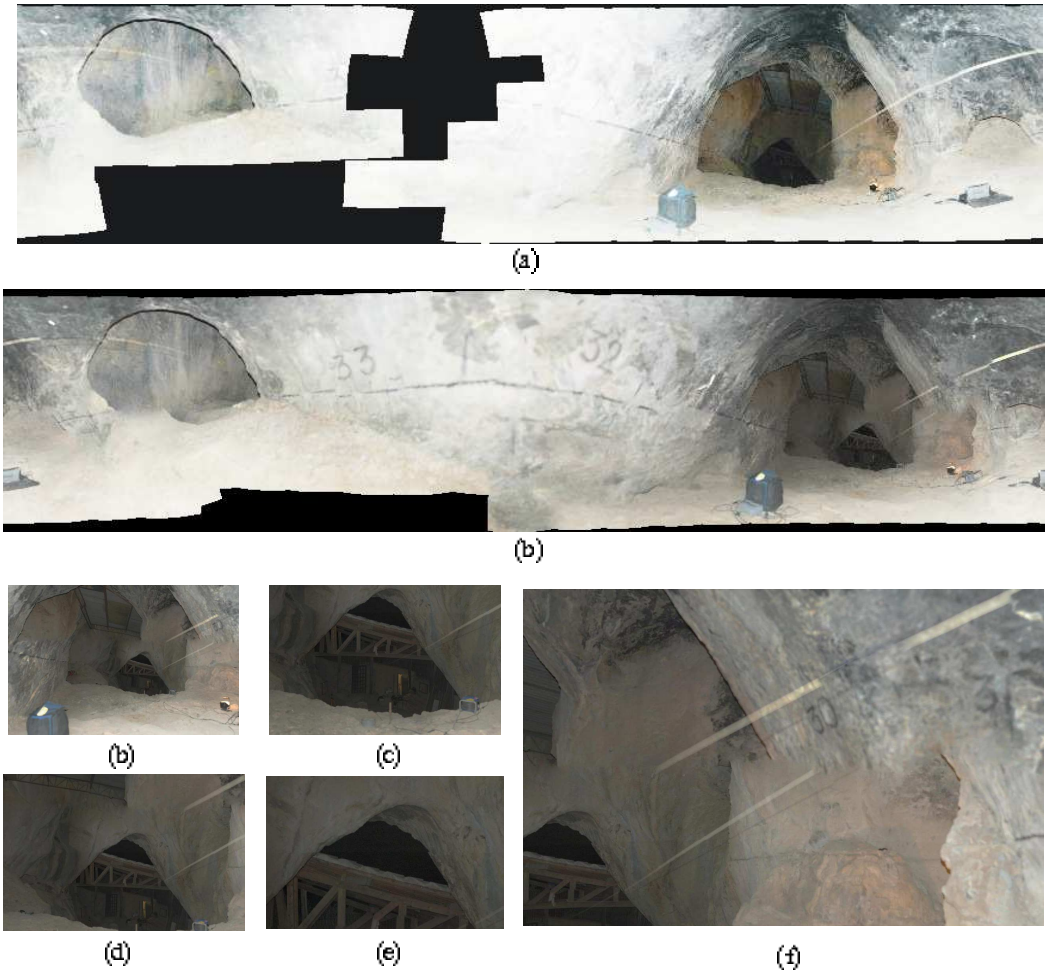


Figure 5.25: Mosaicing test on data set acquired in Tautavel prehistoric cave using the Rodeon® platform. A mosaic generated from images acquired in the cave's background. (a) - mosaic obtained using Autopano Pro : 99 images, mosaic FOV $360° \times 85.22°$,42973×10061 pixels. (b) mosaic obtained using the BA integrated within Autopano Pro powered by the global to local pair-wise alignment procedure presented in section 5.7. (c)-(e) - four LODs corresponding to the left part of the mosaic in (a). (f) - zoom-in on ghosting effects located in the right part of the mosaic.

Mosaic	Figure 5.24 (a)	Figure 5.24 (b)	Figure 5.25 (b)
$\#N_{station}$	272	168	121
FOV(°)	360×108.4	360×105.37	360×84.50
Size(pixels)	43365×13057(567 Mp)	43206×12646 (546 Mp)	42305×9915(418 Mp)
e(pixels)	1.93	1.76	3.10
# AF matches	21840	13440	9680
CPU (time)	8h 12min	5h 33min	3h 50 min

Table 5.8: Qualitative results corresponding to mosaics generated using Autopano Pro and AF matches when running on a 1.66 GHz Linux machine equipped with 2Gb of RAM memory. The mosaics illustrated in Figures 5.24 (a), 5.24 (b) and 5.25 (b) correspond to the cave's entrance, center and background, respectively.

experimental version of the implementation, i.e. without any optimization. Moreover, the number of images to be acquired is reduced to $N_{im} = 32$ which decreases by a factor of 4 the acquisition time.

5.8.2 3D-Cross Bundle Adjustment: Analytical Solution

This section proposes an original analytical solution for refining the relative orientations resulted from the global motion estimation described in Section 5.7.1 which exploits the AF pairings obtained from the local patch matching procedure described in Section 5.7.2.

Let I_i and I_j be two partially overlapped images for which the two aforementioned outputs are available from the pair-wise processing step, i.e.:

- the global rotation $\hat{\mathbf{R}}_{ij}$.
- N homologous AF belonging to the overlapping region defined on $\Omega_{ij} = I_i \cap I_j = \{(\mathbf{u}_i^k, \mathbf{u}_j^k) | \mathbf{u}_i^k \in I_i, \mathbf{u}_j^k \in I_j, k = 0, .., N-1\}$.

Introducing spatial constraints for drift elimination. The simple concatenation of the pair-wise poses' estimates does no introduce spatial constraints, leading error accumulation. It is therefore required to employ a BA technique in order make use of all the AF matches belonging to images which are adjacent to the currently processed image and which are visible from the current camera in order to estimate simultaneously the relative optimal quaternions. This eliminates the drift usually introduced when a simple concatenation of the relative orientation is used by introducing spatial constraints between images belonging to the same station or mosaic node.

Exploit the adjacency information. The proposed scheme exploits the images' adjacency information established through the acquisition setup and minimizes the global registration error over all the images composing the mosaic by using a set of AF features associated to each group of adjacent images (i.e. the set of AF features belonging to the current image which are visible in all its adjacent images). Since images are ordered, the algorithm does not search for overlapping images (or adjacent) allowing considerable computational savings.

Alignment's evaluation using 3D criteria. As shown in Equation (5.20), the BA scheme evaluated in the previous section minimizes the error between the 2D image projections of the transformed 3D vectors which might be biased by the self-calibration

process. In addition, the main goal is to compute the 3D camera motion to which a criterion measured in the 2D image space is less sensible and for this reason our proposal is to employ a criterion measured in the 3D space.

3D-cross criterion formulation. Since in our case the depth coordinate is not known, our solution for the BA process stands in the minimization of the area comprised between the 3D normalized vectors which corresponds to the minimization of the cross product between the 3D vectors $\mathbf{v}_1^k \longleftrightarrow \mathbf{v}_2^k$ corresponding to the AF pairings $\mathbf{u}_i^k \longleftrightarrow \mathbf{u}_j^k$. This can be expressed in terms of both cross and dot product:

$$0 \leq \gamma = \arccos \mathbf{v}_i^k \cdot \mathbf{v}_j^k = \arcsin \|\mathbf{v}_i^k \times \mathbf{v}_j^k\| \leq \pi \tag{5.21}$$

The geometrical meaning of the minimization of the cross product between the corresponding 3D rays is illustrated in Figure 5.26.

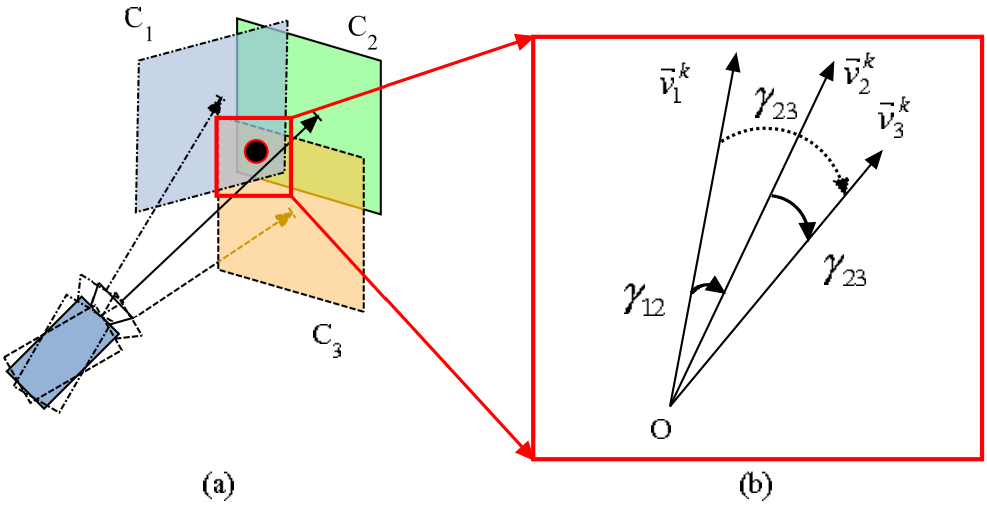


Figure 5.26: (a) The 3D criterion illustrated for $n = 3$ adjacent images. (b)The estimation process minimizes of sum of the angles between the corresponding 3D vectors in each image.

Assuming that we have a set of N corresponding vectors $\vec{\mathbf{v}}_1^k \longleftrightarrow \vec{\mathbf{v}}_2^k$ coming from the corresponding AF matches $\mathbf{u}_1^k \longleftrightarrow \mathbf{u}_2^k$, the cost function for an image pair sums the angles enclosed by each vector couple, being expressed as following:

$$Q_{ij-\times} = \min_R \sum_{k=0}^{N-1} \|\mathbf{v}_i^k \times \mathbf{R}\mathbf{v}_j^k\|^2 \tag{5.22}$$

Multi-view formulation of the 3D-cross criterion. For n images composing a complete spherical mosaic node, several factors need to be taken into account:

- *image adjacency:* in order to model the *adjacency* relationship between I_i and I_j, for all $i, j = 0, .., n-1$ with $i \neq j$, we define the the matrix $\mathbf{W}_{(n\times n)}$ whose terms are defined by a weighting function

$$w_{ij} = \begin{cases} 1, & \text{if } card(\Omega_{ij}) \neq 0 \\ 0, & \text{if } card(\Omega_{ij}) = 0 \end{cases} \tag{5.23}$$

- *feature visibility:* since several features are visible in some images and in others no, we define the visibility matrix $\Theta_{m\times p}$, where m is the number of images composing a group of adjacent images, and p denotes the number of features matched overall m images. We can count whether the feature k detected in image I_i $\mathbf{u}_i^k$, is visible in image I_j by using the following weighting function:

$$\delta_{ij}^k = \begin{cases} 1, & \text{if } \mathbf{u}_i^k \in \Omega_j \\ 0, & \text{otherwise} \end{cases} \tag{5.24}$$

By integrating the two aforementioned factors into Equation 5.22 and by summing over all images composing the mosaic node, we obtain the total cost which need to be minimized when stitching multiple overlapped images acquired from a single view-point, which is expressed hereafter:

$$Q_{mosaic-\times} = \sum_{i}^{n} \sum_{j,i<j}^{n} w_{ij} \sum_{k} \delta_{ij}^k \|\mathbf{v}_i^k \times \mathbf{R}\mathbf{v}_j^k\|^2 \tag{5.25}$$

The technical solution of Equation (5.25) is related to the choice made for the rotation parametrization. We employ quaternions since they allow for elegant and numerically stable solution. Appendix C.6 describes the solution for estimating the optimal quaternion relating two adjacent images, followed by its generalization to the multi-view case in Appendix C.7.

Sequential algorithm. The minimization of the cost function $Q_{mosaic-\times}$ wrt N_q quaternions is done in a sequential fashion using a similar approach as the one introduced by Benjemaa and Schmitt in [Benjemaa and Schmitt, 1997]. Authors introduced an iterative scheme for registration of multiple sets of 3D point clouds by minimizing the sum of squared residual errors between corresponding 3D points in each view. A quaternion-based solution to this problem was proposed in [Faugeras and Herbert, 1986] and [Horn, 1987]. The translation is given by the difference between centroids, while the unit quaternion maximizes the sum of the dot products of corresponding coordinates in the first system with the rotated coordinates in the second system. In this case, the unit quaternion is given by the unit eigen-vector corresponding to the maximum eigen-value of a symmetric 4×4 observations matrix.

In opposite to their case study, in our research work we search for the optimal parameter vector $\mathbf{q}_{node}$ which minimize the sum of square angular errors expressed in $Q_{mosaic-\times}$.

The minimization process starts with a vector parameter $\mathbf{q}_{node}^0$ which can be set to an arbitrary value or initialized by a previous step. At each iteration, all the $N_q - 1$ quaternions are fixed except one of them whose estimation is performed by minimizing the cost function $Q_{mosaic-\times}$. For an arbitrary iteration m, the transition from $\mathbf{q}_{node}^m$ to $\mathbf{q}_{node}^{m+1}$ is done in $N_q - 1$ steps. When all the quaternions are fixed except one, minimizing $Q_{mosaic-\times}$ becomes a simple problem which can be directly solved using the closed-form solution presented in Appendix C.6.

Taking into account the properties of the 4×4 observations matrices $\mathbf{V}_{ij}$, it can be verified that the cost function $Q_{mosaic-\times}$ is lower and upper bounded. The overall criterion $Q_{mosaic-\times}$ is composed by simple concatenation of several relative cost functions $Q_{ij-\times}$, each of which having its minimum given by the smallest eigen-value $\lambda_{ij} = \lambda_{ij}^{min}$ and maximum given by their maximum eigen-value λ_{ij}^{max}. Consequently, it can be shown that the cost function is upper and lower bounded:

$$\sum_{I_i \cap I_j \neq \emptyset} \lambda_{ij}^{min} \leq Q_{mosaic-\times} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} w_{ij} Q_{ij-\times} \leq \sum_{I_i \cap I_j \neq \emptyset} \lambda_{ij}^{max} \tag{5.26}$$

5.9 Conclusions

At the global scope, this chapter proposes a system-oriented solution for generating in-situ Gigapixel mosaics in unstructured and difficult to access environments, without requiring human operator intervention. Since the environment is unknown, the proposed method does not rely on corner-like primitives extraction, providing therefore an environment-independent method.

We conclude by first listing the main issues to solve for when dealing with the in-situ Giga-pixel mosaicing problem in unstructured and difficult to access environments. Once reviewing the main aspects to be addressed, we revisit the algorithm's components and justify our choices. The closer of this section draws the main aspects in which is see the contributions of the chapter.

5.9.1 Addressing key-issues for the in-situ Giga-mosaicing problem

When tackling the Giga-pixel mosaicing problem within an environment-independent method, special attention was given to several factors:

(A) Image matching in feature-less areas. When solving for the automation of the Giga-mosaicing process in whatever environments, we were faced to a well-known issue raised by the image alignment task in presence of feature-less areas. Such an algorithm requires *reliable image matching* in homogeneous, highly textured areas but also in regions presenting repetitive patterns, and natural scenes (such as trees or grass). More specific to the Giga-pixel mosaicing process, is the *accuracy*, since any mis-registration will lead to visible artifacts in the final compositing. Furthermore, the *in-situ* requirement plays an important role in the design of the proposed algorithm, introducing constraints for *automation* and *rapidity*.

(B) Choosing the suitable ingredients. At a first glance, grouping the aforementioned requirements into a single mosaicing framework appeared to be challenging task for the nowadays image mosaicing frameworks. We started by including the suitable algorithms solving efficiently for each requirement:

- *Fast direct method for accurate alignment of HR and low-overlapped images acquired in feature-less areas.* Being given the initial estimates and focusing on providing an accurate image matching algorithm for feature-less areas, our attention was first directed toward direct methods. This guarantees the matchability of low-overlapped images by exploiting all the information belonging to the overlapping region, eliminating the risk of not-founding keypoint features in feature-less areas. The "fast" term is related to fact that for computational savings only pixels extracted from the overlapping areas (i.e. patches belonging to the image border) are correlated within a coarse-to-fine framework.

- *Deviations from parallax-free camera rotation and pinhole camera model.* In our case images are acquired by a motorized pan-tilt head and only small amounts of parallax can be encountered when the optical center is not superposed with the center of

rotation of the platform. In this case, a simple a local patch matching approach can be applied in order to compensate for deviations from purely-rotating camera or pinhole camera model.

- *Fast multi-view fine alignment.* Although accurate, direct methods are computationally too expensive when applying them on HR images for the multi-view refinement stage, whereas BA methods usually employed in conjunction with feature extraction and matching algorithms, are fairly more interesting due to their rapidity.

5.9.2 Revisiting Algorithm's Components

We will now review the design of the proposed mosaicing algorithm combining the aforementioned capabilities.

(A) Pyramidal global-to-local pair-wise image alignment. The pair-wise alignment refines the initial estimates and establishes a global 3D rotation model which, at its turn initializes the non-rigid motion estimation via a local patch matching procedure.

This helps in solving for four issues: (i) it compensates for deviations from free-parallax motion camera rotation, (ii) it allows to establish a set of corresponding "anonymous features" which are reliably matchable in any environment (including feature-less areas) and (iii) it allows to estimate a local displacement for each patch which can be also be used to compute a global translational motion. The last aspect is the most important, as it allows to build the bridge between the pair-wise direct alignment procedure and the BA stage, making possible their jointly use to enable accurate, fast and reliable in-situ Giga-mosaicing.

The proposed scheme has several advantages wrt the existing image alignment methods:

- the direct approach enables fast and efficient computation, provides robustness to noisy initial guesses and initializes the local patch matching procedure sufficiently close to their true homologues, decreasing therefore the outliers' rate.
- the pyramidal approach enables high-resolution image alignment with a reasonable computation time, which is a key aspect for the Gigapixel mosaicing task.
- the local patch matching accuracy ensures that the bundle adjustment step will not get trapped in a local minimum.

Evaluating the environment-independent character of the image matching method. The main ingredient of the mosaicing pipeline which must deal with the feature-less aspect is the pair-wise image alignment process. In order to evaluate the environment-independent character of the proposed method, we tested the pair-wise global-to-local alignment on different acquisition scenarios.

Unstructured and underground environments. We present a first group of trials which test and evaluate the reliability of the proposed system in unstructured underground environments using a camera mounted on a pan-tilt motorized platform capturing HR images with high overlap. We described two trials performed using different experimental setups. The first trial undertaken in Tautavel prehistoric cave tests an experimental version of the algorithm, while the second trial performed in Mayenne Science prehistoric cave validates the algorithm and evaluates the computational savings when an improved acquisition platform is used in order to allow for in-situ processing.

Outdoor structured environments. The second group of trials validates and evaluates the algorithm when used in outdoors structured environments to supply terrestrial mapping

applications. While the first group of trials is susceptible of introducing small amounts of parallax, the second one introduces a higher parallax and aims to stitch poorly-overlapped images with lower resolution acquired on-the-fly by a vehicle equipped with a 10-cameras panoramic moantage.

The presented tests demonstrate the robustness of the algorithm in both structured and unstructured environments, in presence of poorly overlapped regions and relatively high amount of parallax.

(B) 3D Cross-BA for optimal quaternions computation. The patch correspondences between each adjacent image pair are injected in a BA engine for the final optimization. At the first glance we attempt to solve for the multi-view fine alignment using a classical implementation of the BA scheme [Kolor, 2005] and we concluded that the use of a 2D criterion and the self-calibration stage lead to mis-registration errors. Moreover, when AF correct matches are injected into it, the process rejects a majority of the point correspondences, leading to erroneous poses.

The aforementioned shortcomings influenced us to design a closed-from solution for computing the optimal unit quaternion by minimizing the angle between the 3D-rays corresponding to AFs pairings which is given by their cross product. By expanding the criterion to the multi-view case we obtain a sequential algorithm which iteratively computes the relative orientations until convergence. An ongoing work is to implement, test and evaluate the proposed 3D Cross-BA analytical solution on real data acquired in Tautavel and Mayenne Science prehistoric caves.

Possible rendering schemes for dynamic scenes. Our research work focuses mainly on the image alignment process. When it comes to the in-situ mosaic rendering step, rapidity is a priority and therefore a basic rendering pipeline can be used, such as the one proposed in [Garcias and Santos-Victor, 2000]. Nevertheless, off-line an artistic rendering pipeline can be performed using the existing techniques [Shum and Szeliski, 2000], [Lowe, 2004]. A possible future research direction is to attack the mosaicing problem in dynamic scenes. The existing mosaicing frameworks applied on dynamic scenes have as main goal to hide ghosting effects introduced by moving objects in order to produce artistic rendering. In exchange, in our work artistic rendering is not a priority. Moreover, mosaicing make-up is unaffordable for in-situ processing.

GPU implementation. In this chapter we presented a CPU-experimental implementation of the algorithm in order to test and evaluate the behavior of the overall framework on different acquisition scenarios. The algorithm seems to work well on different types of scenery and although the upgraded Rodeon® helps considerably to reduce the computation time, the main bottleneck for Giga-pixel in-situ processing, rendering and visualization is raised by the available computational resources for in-situ processing. Consequently, further improvements take the proposed mosaicing scheme toward a GPU design.

5.9.3 Contributions

While introducing an automatic Giga-mosaicing system, we believe that the main contribution of this chapter stands in the automation of the image alignment task in feature-less areas. While focusing on the design of a robust image matching scheme, novel strategies came up wrt which we see the following contributions:

- **Technical solution.** A global-to-local pair-wise alignment scheme estimates a global motion and leads to a list of homologous image points extractable in any kind of scenery. Since they do not correspond to any salient area, this chapter introduces

them as anonymous features (AF). They can be employed for image matching, tracking and localization purposes.

- **Analytical solution: Cross-BA for multi-view fine alignment.** In order to improve the BA process, we propose an analytical solution for the multi-view alignment stage which minimizes an error in the 3D space given by the angle between the vectors corresponding to AFs. The multi-view formulation takes advantage of sparse matrices, leading to an efficient and fast method for computing the optimal unit quaternions laying between each partially overlapped image couple.

- **Algorithmic contributions.** In order to cope with the in-situ Giga-pixel mosaicing requirements, we exploit the state-of-the-art complementarity by combining pair-wise direct methods with the feature-based BA scheme through the use of locally matched AFs. The two aforementioned ingredients are combined to propose an automatic Gigapixel mosaicing system for generating *in situ* photorealistic mosaics of previously unknown, complex and unstructured underground environments, without requiring human operator intervention. The proposed technique can deal with several key issues of the Gigapixel mosaicing problem in unstructured and large-scale environments, such as: handling the absence of reliably detectable and trackable features, robustness to noisy initial guess and to deviations from pure parallax-free motion.

Wide-range applicability. We report a new technique mosaicing technique employable with different testbeds and acquisition scenarios: either fixed motorized platforms or mobile camera networks mounted onboard ground vehicles. Beside providing Giga-mosaic imagery, the proposed pair-wise method provides the platform's heading, giving the possibility to be employed as a visual compass. Moreover, the mosaicing algorithm can be used to supply site surveys and active vision tasks such as visual-SLAM and recognition. One of our research work applications aims at providing digital recording and virtual visits of prehistorical caves through the world wide web [iCa].

Chapter 6

Generating 4D Dual Mosaics from Image and Laser Data

The previous two chapters were concerned with the automatic alignment of 3D point clouds and HR-images for producing 3D and 2D mosaic views to be exploited further in this dissertation to supply in-situ 3D modeling tasks.

Let us now come back to the main goal of this dissertation which deals with the automation of the 3D modeling pipeline for in-situ photorealist digitization in difficult-to-access and unstructured environments. As shown in Section 2.3 of this dissertation, there are several stages composing the 3D modeling process and each of them must be performed automatically. This chapter deals explicitly with the automation of the image-laser alignment stage in feature-less and GPS-denied areas by proposing an automatic procedure for generating in-situ photorealist 3D models encoded as 4D mosaic views.

We start this chapter by motivating the jointly use of laser and image data and by listing the main key issues which need to be addressed when aiming to supply automatic photorealist 3D modeling tasks while coping with time and in-situ access constraints. The next section comes up with hardware and software solutions forming a 4D mosaicing prototype able the fulfill the aforementioned requirements. Section 6.3 introduces the input data and states the simplified pose estimation problem under calibration constraints. Section 6.4 describes the image-laser alignment algorithm along with preliminary experimental results, while Section 6.5 summarizes our research proposal and draws several future research directions.

6.1 Digital Photorealist 3D Models from Sensor Fusion

When dealing with the automation of the 3D modeling pipeline in difficult-to-access and unstructured underground environments, one has to take into account several constraints in order to solve reliably for the data alignment task in feature-less and GPS-denied areas. This section states several choices and key issues which must be addressed in order to supply in-situ 3D modeling tasks in such environments.

Recovering 3D information beside appearance has now become indispensable due to its tremendous potential for solving reliably for a variety of artificial vision tasks, such as position estimation, navigation, obstacle avoidance, object detection and recognition. The available solutions for all the aforementioned problems can be undoubtedly be improved through the integration of 3D information.

Passive vs. active 3D geometry recovering. Both schools provide 3D geometry but through different environment sensing means, influencing their use wrt both accuracy and computational aspects. A stereo-based framework is subject to features' existence and would be computationally too expensive to be performed in situ. Moreover, the capacity of a stereo-montage to provide depth depends on the baseline. For instance, a stereo-bank mounted on a robot may have a 30 cm baseline, leading to a capacity to estimate depth up to several meters, while a LRF is capable to sense range up to several tens of meters.

Therefore, in order to recover reliably the 3D geometry while overcoming the ambiguity of the feature-matching step, recent techniques are more directed toward the active 3D geometry recovery via 3D LRFs [Dias et al., 2003], [Deveau et al., 2004], [Zhao et al., 2005], [Stamos et al., 2008], [Banno et al., 2008], [Pilania and Chakravarty, 2008].

Key issues of the image-laser data alignment. When tackling the data fusion problem, the main problem for which we have to solve for is the representation of the measurements provided by each sensing device in a common coordinate system. This calls for the computation of the 3D rigid pose, being also refereed to as the multi-sensor calibration problem. The data fusion quality is subject to the 3D rigid transformation relating the sensors, being usually formulated as a minimization problem between common measurements provided by each sensing device. Several laser-image fusion methods were reported over the time in Photogrammetry and Remote Sensing [Reulke et al., 2004], [Deveau et al., 2004], Computer Vision [Dias et al., 2003], [Zhao et al., 2005], [Banno et al., 2008], [Stamos et al., 2008] and Robotics [Cole et al., 2005], [Cole and Newman, 2006], [Newman et al., 2006], [Pilania and Chakravarty, 2008] research communities.

The first community is more interested in off-line laser-image alignment frameworks performed using artificial landmarks and or surveyed ground points. The acquired data is processed by human operators using softwares provided by the acquisition devices. Multi-sensor calibration methods employ a calibrated pattern [Dupont et al., 2005], [Brun and Goulette, 2007] or exploit manually selected common observations [Scaramuzza et al., 2007]. In [Li et al., 2009] authors employ a calibration to get control points. Other approaches employ a detectable calibration pattern [Rodrigues et al., 2008]. However, in risky environments it is difficult to have access and place such a calibration pattern. Consequently, an open issue is how to recover automatically the rigid transformation.

In opposite, the last two research communities have directed their studies toward automatic methods using interest point extraction and matching together with information provided by navigation sensors [Zhao et al., 2005] or by exploiting orthogonality constraints defined over the scene's content [Stamos et al., 2008], the environment saliency [Cole et al., 2005], [Cole and Newman, 2006], or interest features [Banno et al., 2008]. Beside the operational limits raised by the features' existence and the reliability of navigation sensors in GPS-denied areas, an important issue is raised by the image-laser occlusions which are inherent when the two sensing devices have different optical centers, i.e. FMCL system.

Since our research work is concerned with the automation of the 3D modeling pipeline in difficult-to-access and unstructured underground environments, we list hereafter the main key issues which need to be addressed in order to solve efficiently for the 2D-3D data alignment problem in feature-less and GPS-denied areas:

- *Image-laser occlusions.* When both sensors are freely moving, it is difficult and even impossible to perform both image-laser alignment and texture mapping due to occlusions between image and laser data.
- *Data alignment failure in feature-less and GPS-denied areas.* As mentioned before,

automatic solutions for aligning texture maps onto 3D point clouds exploit prior constraints on the scene's content, limiting their applicability to structured or man-made environments.

In-situ 3D modeling requirements. Beside the image-laser alignment shortcomings listed above, additional constraints are imposed by the in-situ deployment of the 3D modeling system.

- *Deal with in-situ constraints.* In difficult-to-access and unstructured underground environments special attention must be given to time and in-situ access. In particular, dense scans acquisition and processing are computationally too expensive to be performed in situ.
- *Provide means to validate in-situ the completeness of the 3D scene model in order to avoid returning on site to complete data collection.* This aspect is concerned with the fully automation of the 3D modeling pipeline, implying in-situ acquisition, processing and visualization. In addition, risky environments preclude access of human surveyors and therefore, supplying 3D modeling tasks through fully automatic procedures is a must.

Figure 6.1 summarizes the main key issues which have to be addressed when solving for the in-situ 3D modeling problem from image and laser data, while the following section describes our research proposal solving for the automation of the data alignment task and coping with the aforementioned in-situ constraints.

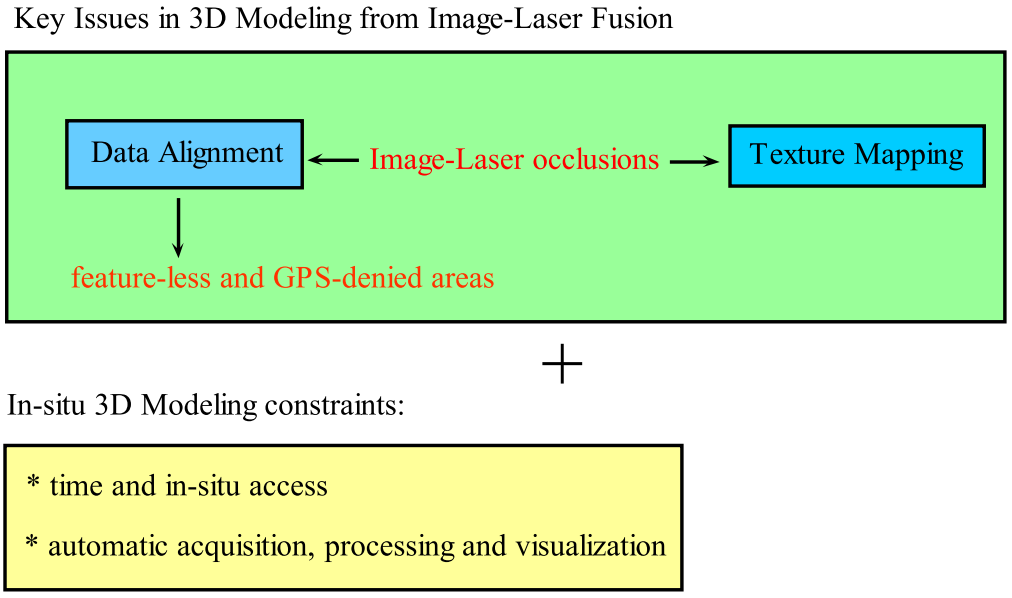


Figure 6.1: Summary of the main key issues which need to be addressed in order to solve for the automation of the image-laser alignment for supplying in-situ 3D modeling in feature-less and GPS-denied areas. The scheme illustrates that both steps, data alignment and texture mapping and subject to image-laser occlusions, while the data alignment its-self requires for an efficient framework reliable in feature-less and GPS-denied areas.

6.2 RACL System for in-situ 3D Modeling via 4D Mosaicing

In order to overcome the image-laser alignment shortcomings raised by image-laser occlusions while fulfilling the aforementioned time and in-situ constraints, we propose the following solutions:

Outlier-free and accurate 3D geometry recovery in feature-less areas through 3D LRFs. In order to address the main issues raised by passive 3D vision techniques related to the features' existence, we capture the 3D geometry through the use of LRFs, allowing to sense reliably the 3D geometry in feature-less areas.

Fast in-situ 3D modeling through complementary color-geometry acquisition (HR-images and low resolution (LR) 3D geometry) and cooperative data fusion. Data fusion is concerned with the combination of different information sources. In [Mitchell, 2007], the author highlights that when attempting to fuse different sensing devices, one must ensure a good compromise between the accuracy (wrt noise) and the integration (i.e. how to reject false measurements). As mentioned in Durrant-Whyte's paper [Durrant-Whyte, 1988a], several data fusion configurations exist: complementary, competitive and cooperating fusion.

In our research work we employ a *complementary* and *cooperative* data fusion approach. We fuse HR-image and LR 3D point clouds (complementary aspect) to deliver photorealist 3D models encoded as 4D mosaic views, which can not be provided by using each sensor separately (cooperative aspect). The proposed combination leads to a 4D mosaicing sensing device which to our knowledge has not been reported by now.

We exploit the laser-image complementarity emphasized also in [Dias et al., 2003] to ensure complete site digitization while fulfilling time and in-situ access constraints. Since for large-scale scenes dense scans acquisition and processing can not be afforded in-situ, our acquisition scenario acquires LR 3D geometry. In opposite to the 3D scanning devices, HR color image's acquisition is instantaneously and the processing time can be reduced via pyramidal schemes. Consequently, we combine HR-images with LR 3D point clouds, leading to a complementary combination of 3D and 2D color data acquisition wrt the acquisition time required by each sensing device in order to fulfill the in-situ requirements.

Occlusions-free image-laser fusion via RACL. Our vision system is based on the acquisition of dual laser-image panoramic views acquired successively from the SVP. We capture the entire 3D spherical FOV from a single 3D pose of the system, using both sensors one after another mounted on a tripod. Therefore, in opposite to the research work reported in [Zhao et al., 2005] and [Stamos et al., 2008], our approach overcomes completely the image-laser fusion problems which may rise due to occlusions when laser and image data are acquired from different view-points. This improves two levels of the entire 3D modeling pipeline: the image-laser alignment and the texture mapping processes.

Automatic feature-less image-laser alignment via 4D mosaicing. The proposed RACL system simplifies the image-laser alignment problem in feature-less and GPS-denied areas by imposing calibration constraints. Since both sensors are considered rigidly attached, the dominant motion between the two mosaics consists in a 3D rotation and a small inter-sensors parallax amount, both being unknown. The proposed RACL system allows to solve automatically for the image-laser alignment problem via a mosaic-based framework, generating in-situ a 4D mosaic for each spatial position of the platform. A 4D mosaic is a 4-channel panoramic view encoding depth and RGB-color information coming from laser and HR-color camera, respectively.

4D mosaic-driven in-situ 3D modeling. When a traditional 3D modeling pipeline

is employed, several 3D models are generated from different 3D poses of the system in order to cover the site in its entirety. During the global 3D scene model computation, several problems are encountered during the 3D model matching stage due to occlusions which lead to an ambiguous feature-matching process.

In order to ensure in-situ the 3D scene model completeness, we propose the use of a 4D mosaic-driven acquisition scenario. In this context, several 4D mosaic views must be automatically acquired, processed, aligned and merged. The proposed 4D mosaic views provide a fully spherical FOV, allowing for long-term feature tracking and facilitating therefore the 3D model matching task via a 4D-panoramic matching process.

Figure 6.2 summarizes the ingredients included in our image-laser fusion scheme in order to overcome the main key issues of the in-situ 3D modeling problem emphasized in Figure 6.1.

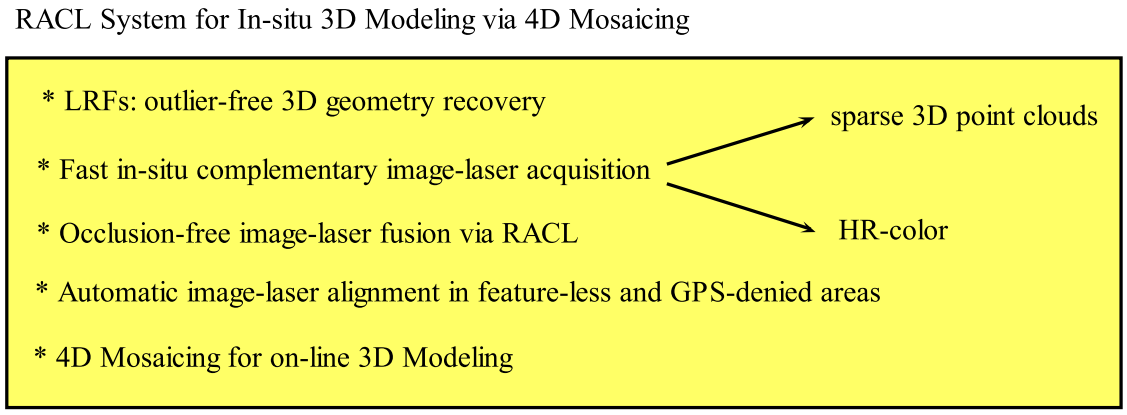


Figure 6.2: The proposed complementary and cooperative image-laser fusion resulting in a 4D mosaicing sensor for in-situ 3D modeling.

6.3 Panoramic-based Image-Laser Alignment

After presenting the main ingredients included in our image-laser alignment strategy, the following two sections provide a brief description of the input data and state the simplified pose estimation problem under calibration constraints.

6.3.1 Panoramic Sensing Devices

The proposed image-laser alignment method is powered by two panoramic views, supplying geometry and color information. We state hereafter several omnidirectional imaging devices capable to capture geometry and color information from real world and justify our choices.

3D mosaic. The Digital Era lead to ubiquitous electronic devices embedding powerful digital processing units which enable massive data acquisition and computing. Although expensive, the recently released 3D scanners allow to acquire accurate and high-density geometry. The close-range ScanStation^{TM}2 manufactured by Leica shown in Figure 6.3 (a) allows to acquire a FOV of $[360° \times 270°]$ with a speed acquisition of 50000 points per second. MESA Imaging designed the SR4000 3D time of flight camera showed in Figure 6.3 (b) which offers high-resolution 3D image data in real-time. It can be used in conjunction with the Giga-pixel mosaicing algorithm described in Chapter 5 to generate

a depth panoramic view. It worth noting that this camera does not provide the intensity information.

On the downside, the aforementioned sensors are limited to low-depth sceneries and therefore they can not supply 3D mosaics in large-scale and complex sceneries.

Low-cost, long-range LRF. In our research work we focus to design a low-cost 3D mosaicing sensor, capable to sense long-range sceneries and to acquire the 3D geometry and the associated intensity backscattering. To this end, we employ the Trimble GS200 LRF illustrated in Figure 6.3 (c) with an addressability of 700 m. The sensor is employed along with a 3D mosaicing acquisition scenario used in conjunction with the automatic 3D mosaicing algorithm described in Chapter 4. Since we aim at using the proposed sensor in complex environments, we enable the capturing device to sense small-occluded areas by acquiring partial mosaic views, avoiding the generation of a fully mosaic view for each 3D pose of the system and decreasing therefore the processing time.

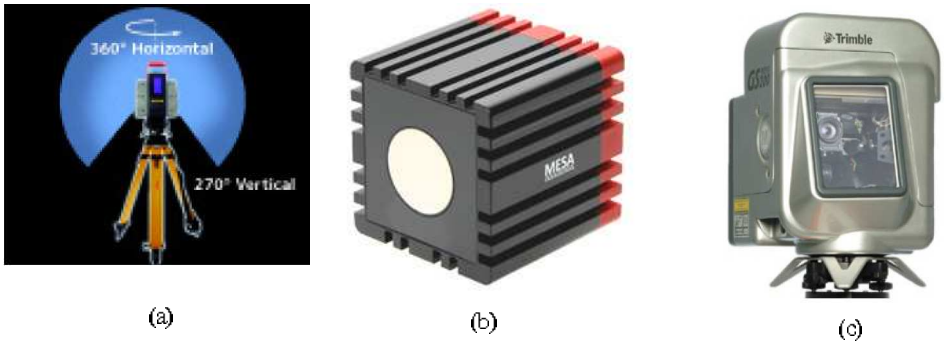


Figure 6.3: Possible 3D mosaic sensing devices. (a) The ScanStation 2 manufactured by Leica. (b) The SR4000 3D time of flight camera manufactured by MESA Imaging. (c) Trimble GS200 employed in our research work.

HR-RGB mosaic. The second main input is represented by the color mosaic which enables the photorealist component of the 3D model. In our research work we employ a motorized pan-tilt head in conjunction with the Giga-pixel mosaicing algorithm described in Chapter 5 to produce a Giga-pixel mosaic.

Other special-purpose imaging devices were designed to supply wide-angle imagery. Fisheye cameras allow to capture directly a hemispherical FOV but they introduce high distortions. Omnidirectional imaging can be achieved through the use of catadioptric cameras [Nayar, 1997].

Although various panoramic imaging devices are now capable to capture directly a fully spherical FOV, on the downside, they do not allow to acquire partial spherical FOV in order to sense occluded areas in complex environments. More precisely, for each 3D pose of the system a fully spherical mosaic is acquired which leads to very high computational costs and causes data redundancy. This is the main reason due to which we prefer to compose mosaic views by stitching several images, allowing to sense occluded areas when needed.

6.3.2 Data Input and Problem Statement

Figure 6.4 illustrates the two inputs of the image-laser alignment procedure. In order to facilitate the visualization of the FOV imaged by each sensor, Figure 6.4 depicts both the 3D spherical and the 2D image projections associated to each input, i.e. the 3D

mosaic generated by the laser and the 2D mosaic obtained from the Gigapixel camera which was down-sampled to meet the 3D mosaic resolution. It can be observed that both sensors are capturing the same FOV, having their optical centers separated by a 3D rotation and a small inter-sensor parallax. In order to build photorealistically textured panoramic 3D models, one must register the 3D spherical mosaic M_{BR-3D} and the color Giga-mosaic M_{HR-RGB} in a common reference coordinate system in order to perform the texture mapping stage.

Pose estimation under calibration constraints. Since the two capturing devices (laser scanner and camera) are supposing to acquire the same FOV, they can be either rigidly attached or used successively, one after another. However, in both cases, it is difficult to calibrate the system such that the parallax is completely eliminated. Consequently, it is possible to model the transformation between the two sensors through a 3D euclidian transformation with 6-DOF (i.e. three for rotation and three for translation) as illustrated in Figure 6.4. The following section is dedicated to the description of the image-alignment algorithm allowing to compute transformation relating their corresponding optical centers.

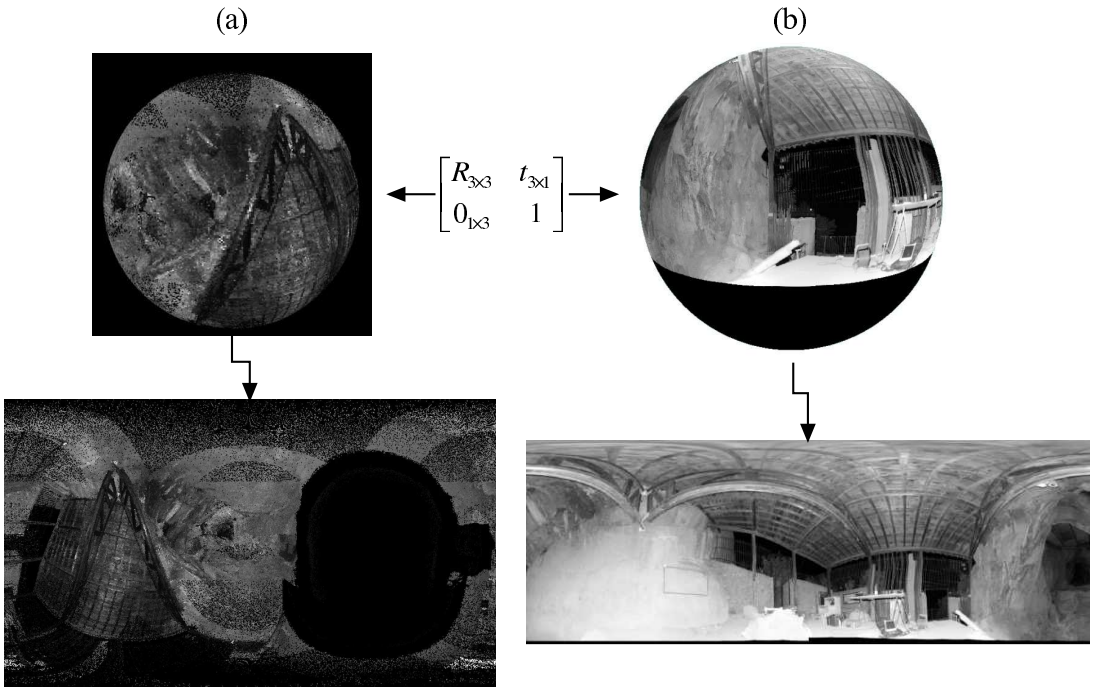


Figure 6.4: The two inputs of the panoramic-based image-laser alignment procedure exemplified on a data set acquired in Tautavel prehistoric cave. We illustrate the spherical and image plane projections associated to each input. (a) M_{BR-3D} - the scan matcher output by the 3D mosaicing process described in Chapter 4. FOV $360° \times 180°$, size: 2161×1276, angular steps $[\delta\theta, \delta\varphi]_{BR-3D} = [0.002906°, 0.00246°]$, (b) the optical mosaic obtained using the algorithm described in Chapter 5. FOV: $360° \times 108.4°$

6.4 Automatic Pyramidal Global-to-local Image-Laser Alignment

We employ a direct correlation-based technique within a feature-less framework. In order to cope with time and in-situ access constraints, we cut down the pose estimation combinatory using a pyramidal framework.

Figure 6.5 illustrates the image-laser fusion pipeline which can be split in two main processes, each of which being detailed through the following description. Since the entire pose estimation method is very similar to the pair-wise global-to-local alignment described in Chapter 5, the following subsections resume several specifications related to its appliance on fully spherical mosaic views.

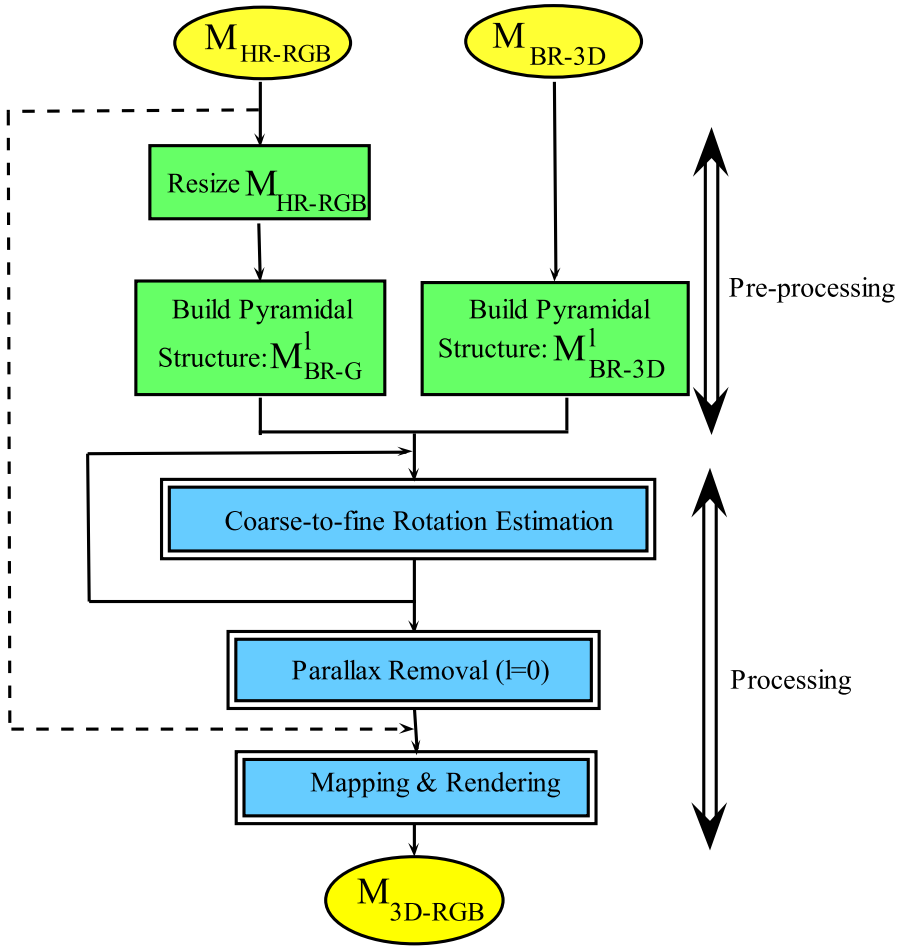


Figure 6.5: Image-laser fusion pipeline. Inputs: 3D mosaic M_{HR-RGB} and 2D Gigapixel color mosaic M_{BR-3D} illustrated in Figures 6.4 (a) and (b), respectively. The pre-processing and processing steps are highlighted in green and blue, respectively.

6.4.1 Pre-processing

The proposed image-laser alignment method correlates the reflectance acquired by the LRF with the green channel of the optical mosaic M_{HR-G}. To do so, we first recover automat-

ically the parameters of the spherical acquisition through a 2D triangulation procedure in order to compute the 2D projection of the 3D mosaic. This stage of the algorithm is very important as it provides the topology between the 3D points and allows fast interpolation.

Down-sampling the green channel of the M_{HR-RGB} mosaic. Our algorithm starts by down-sampling the green channel of the optical mosaic in order to meet the 2D laser mosaic's size. The result obtained is illustrated in Figure 6.6 (a).

Important observation. Once the pose is computed, the high potential of the Gigapixel mosaic can be exploited in order to improve the 3D model resolution through shape-from-X techniques. A method for densifying a sparse 3D point cloud from image data can be found in [Harrison and Newman, 2009].

Generating pyramidal structures for each input: M_{BR-G} and M_{BR-3D}. We generate pyramidal structures of $L_{max} = 3$ levels for both inputs $M_{BR-3D} = \{M^l_{BR-3D} | l = 0, .., L_{max-1}\}$ and $M_{BR-G} = \{M^l_{BR-G} | l = 0, .., L_{max-1}\}$, where the mosaic size ranges from $[2162 \times 1278]$ up to $[270 \times 159]$ corresponding to levels $l = 0, .., L_{max}$.

6.4.2 Pose Estimation

The pose estimation procedure employs a hybrid scheme, the 3D rotation is computed by minimizing a radiometric criterion in the 2D mosaic space, while the translation is computed by back-projecting the rotationally aligned mosaics in the 2D space via a local patch matching procedure. The proposed approach lead to a two-steps rigid transformation computation process: first, the 3D global rotation $\mathbf{R}_{(\theta,\varphi,\psi)}$ is computed in a pyramidal fashion, while the second step is dedicated to the inter-sensor parallax compensation being performed only at the highest resolution level.

Correction of 3D mosaic distortions. As mentioned in Chapter 4, the 3D mosaic acquisition combines several bands acquired through laser's rotations which may introduce wavy effects within the 3D mosaic geometry. These effects are captured within the inter-sensor parallax computation step which is performed through a non-rigid motion estimation procedure. Consequently, in order to correct the 3D mosaic's geometry, the alignment procedure is performed by aligning the 3D mosaic onto the 2D optical one, M_{BR-G}.

Global rotation estimation. The algorithm employs a patch-based correlation strategy using quaternions in order to solve for the optimal 3D rotation. In order to avoid the use of navigation sensors and features' extraction, the rotation estimation is performed in a full search fashion. Since the same type of operation is performed at each pyramidal level $l = 0, .., L_{max} - 1$, let us now drop the superscript l through the following description.

Assume that an arbitrary 3D rotation $\mathbf{R}_{(\theta,\varphi,\psi)}$ belonging to the rotation solution space $\mathcal{P}_{RS}$ is applied to the 3D point cloud. Each transformed 3D point $\hat{\mathbf{p}} = \mathbf{R}(\mathbf{q})\mathbf{p}$ is back-projected in the 2D mosaic space using the intrinsic parameters of the spherical geometry $\mathcal{S} : \{\delta\theta, \delta\varphi, \theta_{min}, \theta_{max}, \varphi_{min}, \varphi_{max}\}$ previously recovered through a 2D triangulation procedure. The transformed points are converted in spherical coordinates (θ, φ) in order to retrieve the corresponding pixel locations $\mathbf{u} = [u, v]^T$ in the 2D mosaic space. This can be expressed using the spherical projection of the 3D mosaic $\mathcal{S}_{BR-3D}$ under the following form:

$$\hat{\mathbf{u}} = \mathcal{S}^{-1}_{BR-3D}\mathbf{p} \tag{6.1}$$

Consequently, a pixel belonging to the 3D mosaic M_{BR-3D} can be mapped into the optical mosaic space M_{BR-G} by using the following composed projection:

$$\hat{\mathbf{u}}_{BR-G}(\mathbf{R}) = \mathcal{S}^{-1}_{BR-G}\mathbf{R}(\mathbf{q})\mathcal{S}_{BR-3D}\mathbf{u}_{BR-3D} \tag{6.2}$$

The grey level associated to the image point location $\hat{\mathbf{u}}_{BR-3D}$ is obtained via bilinear interpolation in the 2D projection of 3D mosaic M_{BR-3D}. Thus, for an arbitrary rotation value $\mathbf{R} \in \mathcal{P}_{RS}$, Equation (6.2) provides us with the transformed mosaic $\hat{M}_{BR-3D}(\mathbf{R})$ projected in the optical mosaic geometry M_{BR-G}. The optimal rotation is obtained by minimizing the dissimilarity in brightness between M_{BR-G} and $\hat{M}_{BR-3D}(\mathbf{R})$.

When comparing the 3D mosaic inputs illustrated in Figure 6.6 (a) to the green-channel of the resized optical mosaic M_{BR-G} shown in Figure 6.4 (b), one can observe that each sensing device has different responses and consequently, the ZNCC score is used to achieve robustness wrt illumination changes. The optimal rotation $\hat{\mathbf{R}}$ minimizes the ZNCC score measured over the entire image space expressed by the following equation:

$$\hat{\mathbf{R}} = \arg \max_{\mathbf{R} \in \mathcal{P}_{RS}} \sum ZNCC(M_{BR-G}, \hat{M}_{BR-3D}(\mathbf{R})) \tag{6.3}$$

Since the 2D projection may result in zero and lost pixels $\hat{\mathbf{u}}$, the procedure employs a down-weighting function to correlate only valid pixels, as presented in Chapter 5.

Figure 6.6 illustrates the results obtained by the optimal rotation estimate. The alignment correctness can be visually inspected when looking at Figures 6.6 (a) and (b). Figure 6.6 (c) shows that the superposition of the two images does not result in grey-level due to the different responses given by the sensing devices. Figure 6.7 (b) illustrates a close-up view of the superposed mosaics showing that the global rotation does not model completely the motion separating the camera and the laser, and consequently the inter-sensor parallax must be introduced within the estimated motion model.

Parallax removal. We recover the parallax between the laser's and the optical mosaicing platform by performing a local patch matching procedure at the highest resolution of the pyramidal structure. As for the local patch matching procedure described in Chapter 5, this stage of the algorithm uses the rotationally aligned mosaics.

The patch matching procedure outputs a 2D translational motion for each patch, estimating a non-rigid motion over the entire mosaic space. This vector field is used for the parallax removal stage. In addition, the non-rigid motion allows to compute a mean translation motion model defined over the entire mosaic space $\overline{\mathbf{t}}_{2D}$. The parallax is removed in the 2D image space by compensating each $\overline{\mathbf{t}}_{2D}$, obtaining therefore the warped 3D mosaic $\hat{M}_{BR-3D}$ aligned onto the 2D mosaic. Figure 6.7 (c) depicts the result of the laser-camera alignment procedure.

Accuracy. Although the Giga-pixel mosaic produced using the Autopano Pro software (details are presented in Chapter 5) has a RMS of 3.74 pixels, these residual errors become negligible in the down-sampled mosaic M_{BR-G} used for the registration process. A sub-pixel accuracy can be achieved by using a bicubic fitting, as described in Chapter 5.

6.4.3 Texture mapping and rendering

This is the final stage of the 3D modeling pipeline which actually finalizes the 4D-mosaicing process. Since the main goal of the research work presented in this dissertation is concerned with the in-situ 3D modeling problem, we are mainly interested in producing a fast rendering technique for visualization purposes in order to validate in-situ the data acquisition correctness. To this end, a simple point-based rendering procedure may suffice. Nevertheless, off-line a more artistic rendering can be performed by sending data to a host wirelessly connected to the target.

In-situ point-based visualization. The employed method simply associates the RGB-color to its corresponding 3D coordinate. In order to emphasize the photorealist

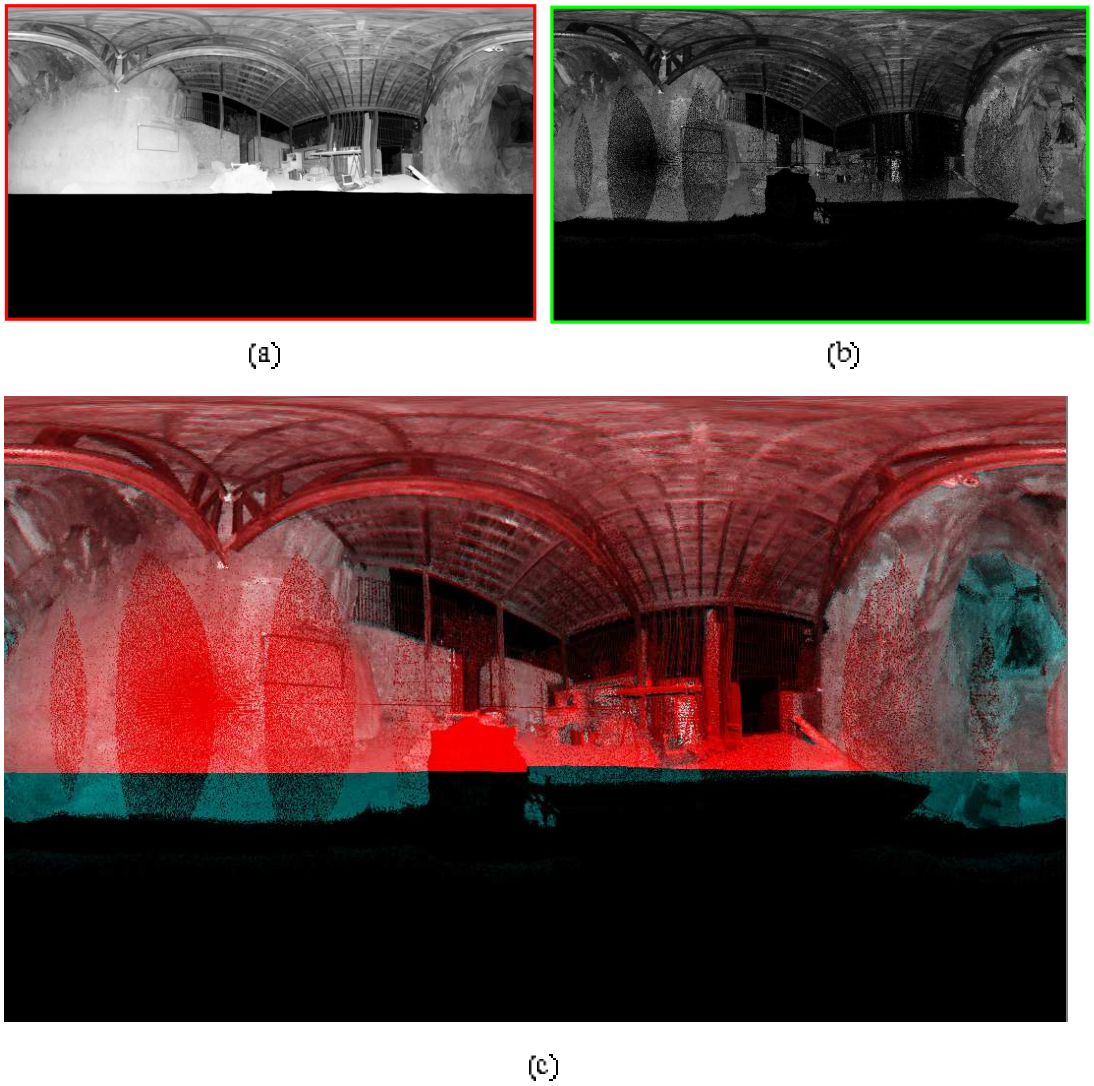


Figure 6.6: Experimental result on a data set acquired in Tautavel prehistoric cave. (a) The the green channel of the down-sampled optical mosaic M_{BR-G} meeting the 3D mosaic size: FOV $360° \times 180°$, size: 2161×1276, angular steps $[\delta\theta, \delta\varphi]_{BR-G} = [0.00296°, 0.002458°]$. For visualization purposes, we add null-pixels to form a complete spherical FOV. (b) The 3D mosaic $\hat{M}_{BR-3D}$ warped in the optical mosaic's geometry using the optimal rotation estimate $\hat{\mathbf{R}}(\theta, \varphi, \psi) = [92.82°, -86.667°, -66.667°]$. (c) Superposed aligned mosaics: M_{BR-G} - red channel, $\hat{M}_{BR-3D}$ - greed channel.

rendering results obtained when using high-resolution texture maps, Figure 6.8 compares the rendering results obtained by first using the intensity acquired by the 3D scanning device illustrated in Figure 6.8 (a), while the rendering using the texture maps obtained from the color mosaic is shown in Figure 6.8 (b).

Off-line mesh-based rendering. We apply a 2D meshing algorithm developed in our laboratory by Mathieu Brèdif and assign to each polygon the RGB-color corresponding to its 3D coordinates. Figures 6.9 illustrates the rendering results showing that the complex surface geometry of the environment lead to 3D point cloud discontinuities which are difficult to handle by the 2D meshing algorithm. In such environments, a more elaborated

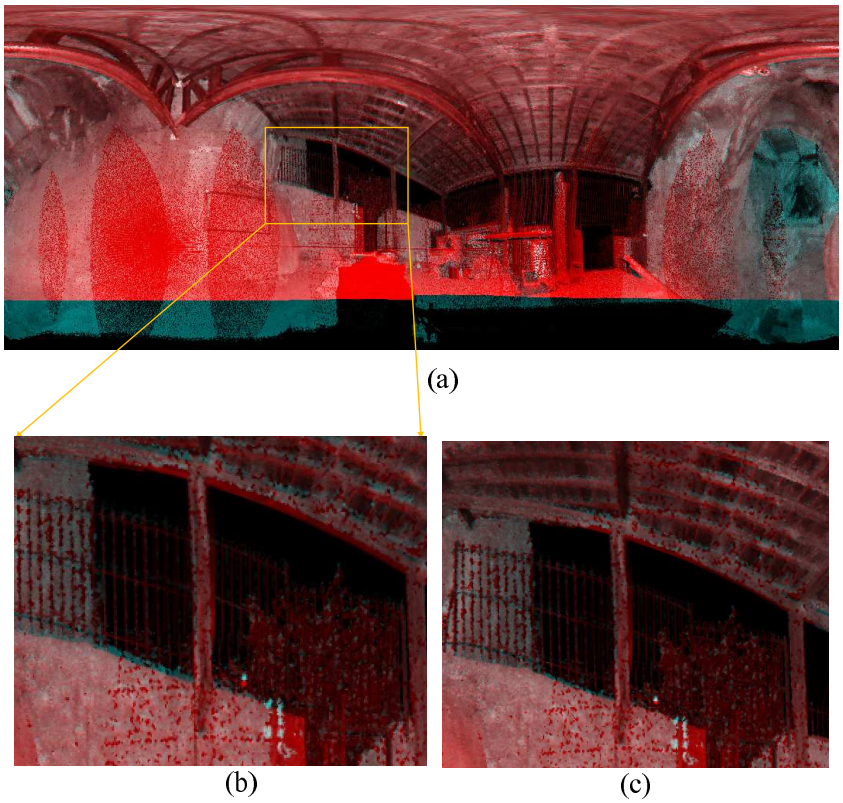


Figure 6.7: Experimental results of the parallax removal procedure obtained on data sets acquired in Tautavel prehistoric cave: (a) Superposed aligned mosaics: M_{BR-G} - red channel, $\hat{M}_{BR-3D}$ - greed channel. (b) zoom in - before parallax removal, (c) zoom in - after parallax removal. The compensated parallax amount: $\bar{\mathbf{t}}_{2D} = [-1.775, -0.8275]^T$ pixels.

meshing algorithm must be designed in order to provide robustness to missing data.

6.5 Conclusion

This chapter investigates the image-laser solutions to address the automation of the 3D modeling pipeline in feature-less and GPS-denies areas. Although the proposed solution exploit the 3D geometry captured through active 3D vision means, it can be applied to 3D point clouds acquired through passive 3D vision techniques. The research studies presented in this chapter are concretized in a 4D mosaicing sensor prototype and contribute to the automation of the image-laser alignment process. We summarizes hereafter the main features of the aforementioned studies.

4D mosaicing sensor for omnidirectional and photorealist in-situ 3D modeling. In order to solve for the image-laser alignment problem in unstructured and underground environments, we propose hardware and software solutions giving rise to a 4D mosaicing sensor prototype providing omnidirectional and photorealist 3D models.

- **Hardware solution.** We came up with a hardware solution consisting in a double-head panoramic imaging device embedding a panoramic HR-color camera and a 3D LRF. The two sensors are rigidly attached and their relative position is unknown.

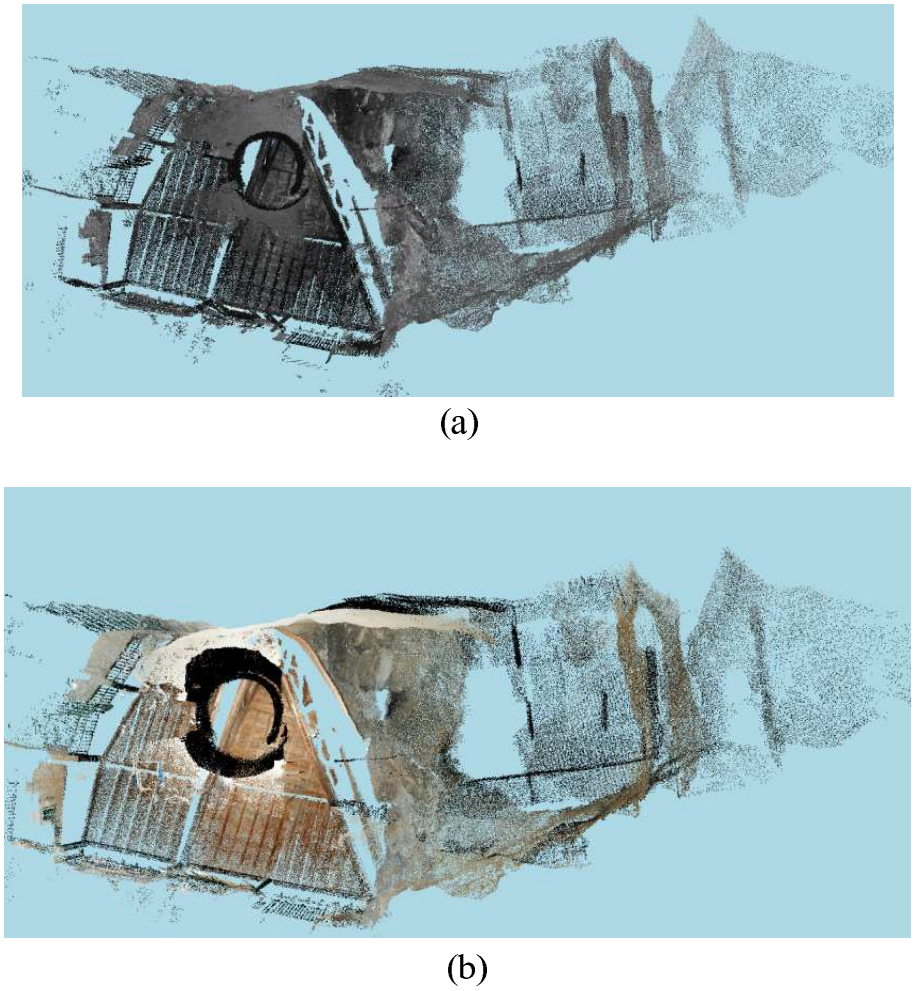


Figure 6.8: Texture mapping results. (a) The 3D point cloud displayed using the intensity acquired by the LRF. (b) The colored 3D point cloud using the down-sampled optical mosaic M_{BR-RGB}.

Figure 6.9: Mesh-based rendering of the Tautavel prehistoric cave. (a) Outdoor view. (b) Indoor view of the 3D model.

- **Acquisition scenario.** The 4D mosaicing sensor employs a specific acquisition scenario, each of its heads being set to acquire the necessary data in order to cover a fully spherical FOV to generate color and 3D mosaic views. In addition, the acquisition protocol takes care of time and in-situ constraints by acquiring low resolution 3D point clouds and HR images.
- **Software.** Although both inputs are provided by the 3D and 2D mosaicing procedures described in Chapters 4 and 5, respectively, let us assume that both inputs are available from previously processing steps. The 4D mosaic sensor embeds automatic procedures for the image-laser fusion task. We propose a mosaic-based framework in order to estimate the 3D pose separating the optical centers which is further exploited within the texture mapping stage. We employ a point-based rendering step to be performed in-situ to allow fast visualization.

Automatic image-laser pose estimation. The image-laser alignment step solves one of the main shortcomings standing behind the automation of the 3D modeling pipeline. With help from both hardware and acquisition scenario, addressing the image-laser alignment task becomes a more simple problem, allowing us to resume the pose estimation process to a rotation computation and a small-parallax removal step. The pose estimation process exploits a mosaic-based scheme and switches between the 3D and 2D spaces along the estimation process. The pose estimation is designed in two steps: the global rotation computation is performed within a coarse-to-fine approach, being followed by a local patch matching procedure which provides means for parallax compensation. The two advantages of the proposed image alignment method are its simplicity and capability to solve simultaneously for the following issues:

- automatic rotational alignment between the camera and the laser with toleration to small inter-sensor parallax amounts;
- robustness to feature-less and GPS-denied areas;
- on-line acquisition, processing and visualization.

Validation and further improvements. This chapter presented preliminary results of the image-alignment fusion method letting us concluding the feasibility of the proposed scheme. Future work is concerned with the validation of the proposed method on a recent data acquisition campaign performed in the Mayenne Science prehistoric cave in order to allow the in-situ demonstration of 3D modeling applications using the 4D mosaicing prototype.

The main future research work aims at exploiting the high-potential of the 2D Gigapixel mosaic to densify the 3D point cloud. This allows us to create rich textured 3D maps which can provide a valuable information for addressing several problems related to photorealist 3D modeling and autonomous navigation schemes.

Following the application type, one may choose to employ either a basic- or an artistic-rendering scheme. Since our research work concerns the in-situ world modeling problem, for computational savings, we employ a fast point-based rendering technique. For off-line uses, such as creating virtual models for digital heritage and virtual traveling applications, it is required to develop a rendering scheme capable to deal with with missing data caused by the accidental surfaces inherent to complex and large-scale environments.

In-situ uses of the 4D mosaicing sensor. Our research studies are mainly concerned with the in-situ use of the 4D mosaicing sensor. We are mainly interested in

exploiting the great potential 4D mosaic views for in-situ world modeling, localization and exploration purposes. We list hereafter several in-situ uses of the 4D mosaicing sensor.

- **4D-mosaic-driven in-situ 3D modeling.** In complex environments, several views need to be acquired in order to ensure the complete 3D modeling of the site. To do so, we propose the use of a mosaic-driven 3D modeling scenario, in which the system acquires, aligns and merges them dynamically in order to ensure in-situ the 3D scene model completeness. The 4D mosaics are powerful tools, encoding both geometry and color information, allowing to disambiguate the data matching task which is inherent to outliers in complex environments.
- **4D-mosaic hybrid SLAM.** The proposed 4D mosaic can be used as a basic entity for panoramic-SLAM purposes, helping to disambiguate the data association problem and providing long-term feature tracking.
- **Visual servoing.** 4D mosaic views encode rich representations of the real world which can be used along with visual servoing procedures to provide active vision to unmanned systems.

Figure 6.10 illustrates an inside view of the 4D mosaicing sensor prototype, highlighting the image-laser solutions emerging from it for supplying off-line and in-situ applications. Since the research work presented in this dissertation aims at designing a vision-based system for generating in-situ complete 3D models of complex environments, we are mainly concerned with the in-situ 4D mosaicing uses which are integrated within the ARTVISYS system to supply autonomous site digitization and exploration applications.

Autonomous site digitization and exploration. We aim at exploiting the potential of the 4D mosaic views to guide a 3D modeling system to act intelligently on-the-fly and sense the occluded areas in order to ensure in-situ the 3D scene model completeness. In this context, the 4D mosaics represent basic entities designed to supply in-situ 3D modeling and exploration missions taking place in difficult to access environments. For this reason, the next chapter integrates the 4D mosaicing sensor within a modular software architecture and describes the solutions provided by the 4D mosaic views for each processing block composing it.

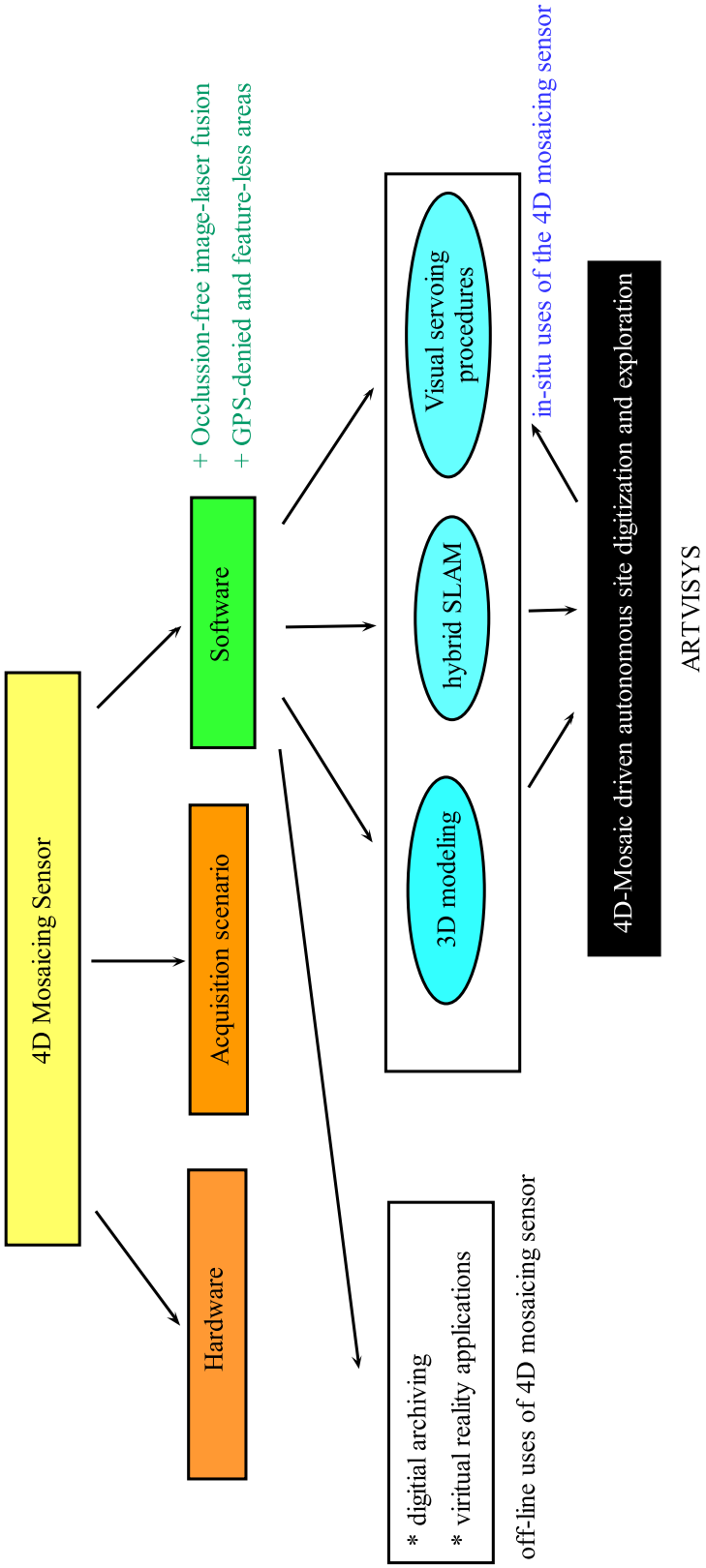


Figure 6.10: Summary of the 4D mosaicing sensor prototype and the image-laser solutions emerging from it for supplying off-line and in-situ applications.

Chapter 7

Toward 4D Panoramic-driven Site Exploration

This chapter proposes the integration of the 4D mosaicing sensor within a vision-based system designed to supply autonomously the digitization of complex environments in their entirety.

Since this aspect is intrinsically related to the system's autonomy through the world modeling capability, in Section 7.1 we investigate the potential of the 4D mosaicing sensor to provide a vision-based autonomy model which is instantiated to the site digitization and exploration case. In the next section we employ it to design the software architecture of an autonomous site digitization and exploration system, having as nucleus the 4D mosaicing process which gives rise to the ARTVISYS system. It's visual control loop includes several processing blocks which require the 3D pose estimation between partially overlapped 4D mosaic views. After reviewing the available solutions solving the 3D model matching problem in Section 7.3, we exemplify a more difficult case on a data set acquired in the Tautavel prehistoric cave. Section 7.5 proposes a 4D-panoramic-based solution of automatic 3D model matching by designing a *viewpoint invariant hybrid descriptors* - VIHD, to be used in conjunction with an *unambiguous matching* strategy. Several existing solutions are proposed to supply the estimation of the 3D rigid transformation relating two partially overlapped 3D models, taking into account several system's configurations. The closure of this chapter summarizes the main contributions and provides several future research directions concerning the remaining processing blocks of the visual control loop.

7.1 Proposed Visual Autonomy Model

Our research work focuses on the world modeling functionalities in feature-less and GPS-denied areas. Supposing that proprioceptive sensors are not embedded on a mobile platform, we propose to study and evaluate how far can we go with the system's autonomy by exploiting only exteroceptive sensors (camera and laser). Appendices D.1 and D.2 provide a brief review of the existing methods dealing with site digitization and exploration problem, emphasizing the relationship between the visual autonomy and the world modeling procedures which we illustrate in Figure 7.1.

For this reason, in this chapter we propose a vision-based autonomy model powered by the 3D world modeling process, emphasizing the high potential of passive and active 3D vision alone to supply unmanned systems' autonomy. Nevertheless, the proposed model

can be further improved with proprioceptive devices when reliable.

To this end, this section integrates the 4D mosaicing sensor within a vision-based exploration framework, providing mapping, localization and navigation in the already explored and built world model. We integrate the 4D mosaicing sensor within the vision-based autonomy architecture illustrated in Figure 7.2 whose composing blocks are detailed hereafter.

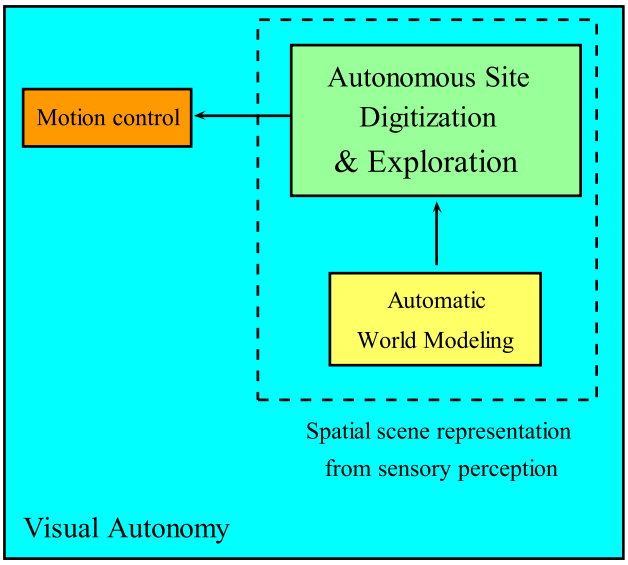


Figure 7.1: The vision-based autonomy of unmanned mobile systems is intrinsically related to the capacity of the system to autonomously digitize and explore the environment, which at its turn requires automatic world modeling procedures.

1. Sensory perception. Unmanned systems must be equipped with instinctual perception such as seeing, hearing and touching. Visual information is our main concern, which is usually gathered through exteroceptive sensors (camera, lasers, sonars) which are embodied using different configurations. In our case study, we employ a camera-laser hybrid device giving rise to the 4D mosaicing sensor presented in Chapter 6.

2. Artificial vision through dynamic world modeling. As for humans, vision is an elementary functionality which provides knowledge about the system's surroundings, allowing situational awareness during the execution of the required tasks. Therefore, embedding unmanned platforms with automatic procedures for environment sensing and modeling is vital as they provide them with rich and understandable information about the real world. This gives rise to an artificial vision engine powered by a dynamic world modeling process.

3. Scene understanding. In order to exploit the high potential of the artificial vision engine provided by the world-modeling process, scene understanding functionalities must be developed in order to extract high-level semantics characterizing the scene's content. This is of primer importance since it impacts on the future decisions and actions of the platform. This calls for two subsequent blocks: characteristics' extraction from images and definition of cognitive rules governing the environment surrounding the system.

(a) Scene characteristics extraction. The scene understanding process starts by extracting different characteristics from the scene model via signal and image processing techniques (region segmentation, optical flow, edge extraction, entropy computation, etc.).

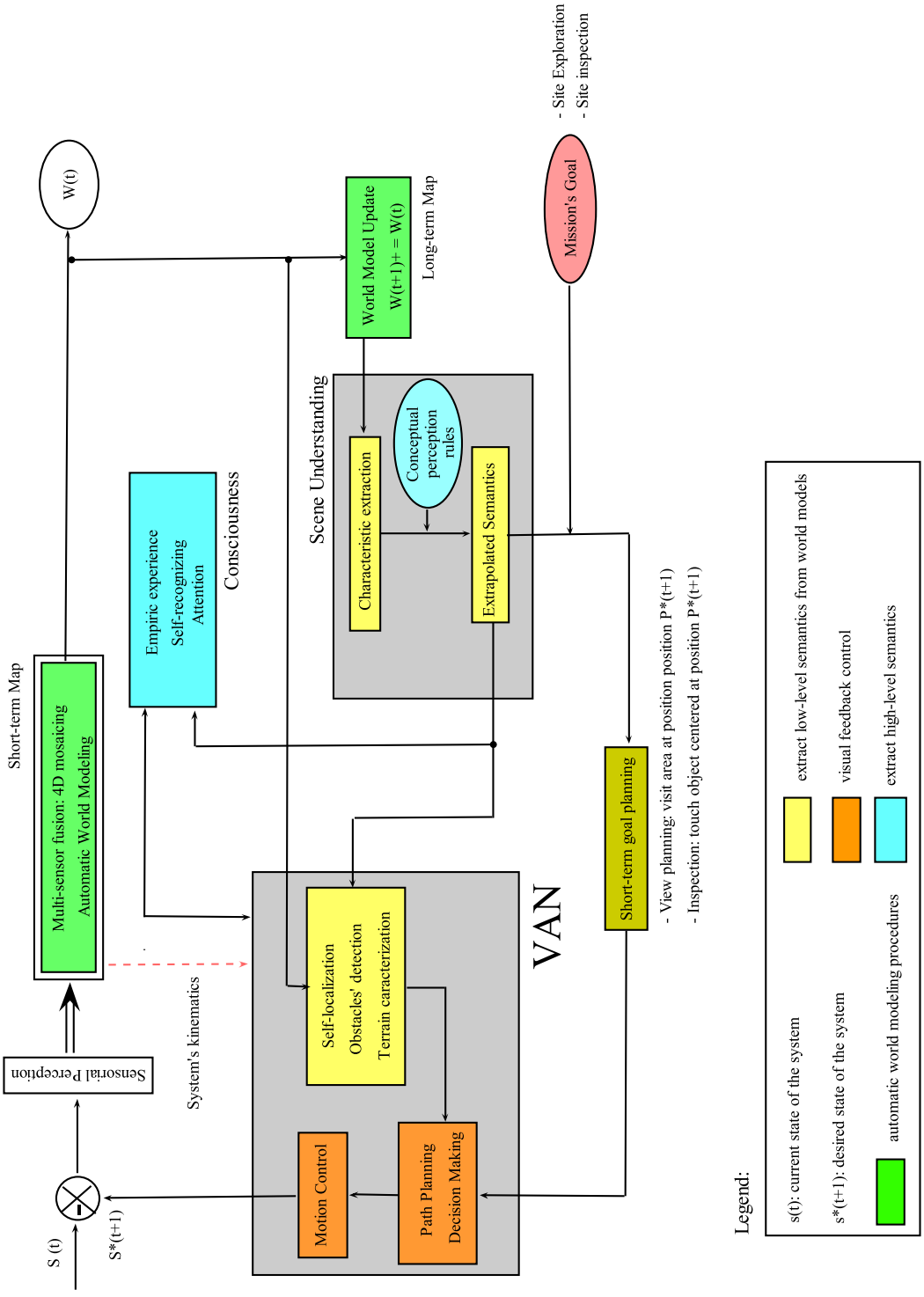


Figure 7.2: Vision-based autonomy model: the VAN process represents the vision-based autonomous navigation module described in this section.

(b) Conceptual perception. Characteristics alone are meaningless and reality perception rules need to be defined in order to provide basic semantic interpretation based on which high-level semantics can be further extrapolated. For instance, optical flow compu-

tation provides knowledge on scene's dynamic, region segmentation and entropy computed in several areas allows us to characterize them (i.e. planar or highly unstructured), discontinuity depths allows us detecting unvisited places and so on. To this end, an underlaying module is required in order to relate the platform to the exterior world through senses and ideas as a result of collective and individual reasoning. For this reason, basic notions providing scene interpretations are required to provide situational awareness, in order to answer questions such as: what a depth discontinuity means, or to what can be associated light or other colors and how to classify materials (wood, rock, water) surrounding the system. All these operations are related to the exploration task, which provides knowledge and contributes directly to the system's autonomy.

4. Short-term goal planning. The environment is previously unknown, being subject to unexpectedly changes. Following the scene understanding results, several goals may be prioritized, influencing the action flow. This module provides short-term planning based on the 3D world model build so far in order to drive the platform toward the final goal. For instance, for environment digitization and exploration the short-term goal is to visit the occluded areas detected by the scene understanding module which requires the next best view computation. Another example can be given for geological inspection of objects classified as "interesting" during the previous stage. This block is powered by fast decision making resources built upon the reasoning principles defining the scene understanding module.

5. Visual-based autonomous navigation (VAN). The platform must be capable to autonomously navigate within the already generated scene model in order to reach a particular 3D position. For this reason, the navigation process generates possible motion paths to the target using the internal world model, the size of the platform and its maneuverability. Visual odometry and obstacle detection procedures interfere with the motion control process via visual servoing procedures.

(a) Self-localization. During navigation and exploration stages, the system must always be aware of its positioning wrt its inside world model. The self-localization procedure answers the system's question "Where am I?", taking in charge kidnapping cases. Since the system must be capable to navigate in GPS-denied areas and unstructured terrains (i.e.odometry is not reliable), the implementation of vision-based localization procedures are therefore a must.

(b)Short-term piloting. The environment can be dynamic and therefore the path planned may be deviated by recently appeared objects in the trajectory. In order to deal with unpredicted situations, a pilot module is charged with the path updating process.

(c) Interacting with surroundings. As stated in the exploration strategy introduced in [Bolduc et al., 1996], the navigation module must approach the obstacles in order to allow their classification by the scene understanding module in "interesting "and "not interesting". During an exploration mission, the system is required to "touch" the interesting ones and to conclude on their properties. This environment exploration module can also help for the navigation task by providing information about the material properties of the system surroundings (traversable surfaces, water, wood, rock, humans, etc.).

6. Consciousness. Studies on a machine consciousness approach [Moreno and de Miguel, 2006], [Moreno and de Miguel, 2008] shown that these methods can provide a controlled behavior when dealing with unexpected situations. Learning capabilities are expected to be improved as they are driven by attention and subjective experience acquired on-the-fly should improve considerably the system autonomy when dealing with uncertainty.

Other contextual constraints, such as real time processing and robustness to large variability of the real world, interfere with the aforementioned functionalities.

The proposed visual-autonomy paradigm provides a general architecture applicable for a wide range of missions taking place in unknown environments. The next section illustrates how we apply it in our research work to propose a vision-based autonomous site exploration strategy scheme to supply site surveys in complex and underground environments.

7.2 Visual-actuated 4D Mosaicing Sensor for Site Digitization and Exploration

Our research work is willing to produce complete and photorealistically textured 3D models of complex and difficult to access environments. Therefore, once accessing the site, the system must be capable to ensure in-situ the 3D scene model completeness in order to avoid to return on site to collect new data. This calls for an intelligent system capable to acquire and process data on the fly, and to exploit it in order to guide and control the system to ensure the 3D scene model completeness.

In this dissertation we propose to exploit the high potential of textured 3D maps encoded as 4-channel panoramic views to supply visual servoing procedures for vision-based autonomous site digitization and exploration operations. The high-detailed 3D maps provide valuable help for exploration, allowing to search for reachable targets, to find traversable terrains and to localize the system. Figure D.4 synthesizes the available solutions for the autonomous site digitization and exploration problem and the proposed strategy described in this section.

To address this issue, we employ the visual-autonomy paradigm from Figure 7.2 and integrate the 4D mosaicing sensor within a visual-based site digitation and exploration architecture, giving rise to a vision-based 3D modeling system, ARTVISYS for which a global overview was provided in Chapter 3.

Due to occlusions, several 4D mosaic views need to be acquired from different 3D spatial positions of the system in order to ensure the digitization of the entire site. As mentioned in Chapter 3, we employ a 4D-mosaic driven acquisition scenario in a stop and go fashion. Traditional systems employ a camera with limited FOV, leading to a low situational awareness and limited 3D perception. The main advantage of the 4D mosaicing sensor is that it captures directly a a fully spherical 3D model and photorealistically textured from a single 3D pose of the system, increasing spatial perception and allowing for long term tracking features between successive sensor readings. Figure 7.3 illustrates a zoom-in view on the visual feedback loop actuating the 4D mosaicing sensor, showing its composing processing blocks which are briefly detailed hereafter.

1. Long-term world model update. This procedure is called at each new spatial position of the system, when a 4D mosaic is acquired. The following two subsequent operations are required in order to integrate the currently acquired 4D mosaic, noted $M_{4D}(t)$ into the global 3D scene model generated so far, noted $W(t)$.

(a) 4D Mosaic matching. The system acquires and generates a 4D mosaic view at each spatial position. As they are generated, the 4D mosaics are sequentially aligned and integrated into a global 3D scene model. After each acquisition, the generated 4D mosaic must be aligned wrt a global coordinate system in order to be merged with the previously generated ones, forming thus the global 3D scene model. To this end, this chapter provides an automatic solution to ensure reliable 4D-mosaic matching in unstructured environments.

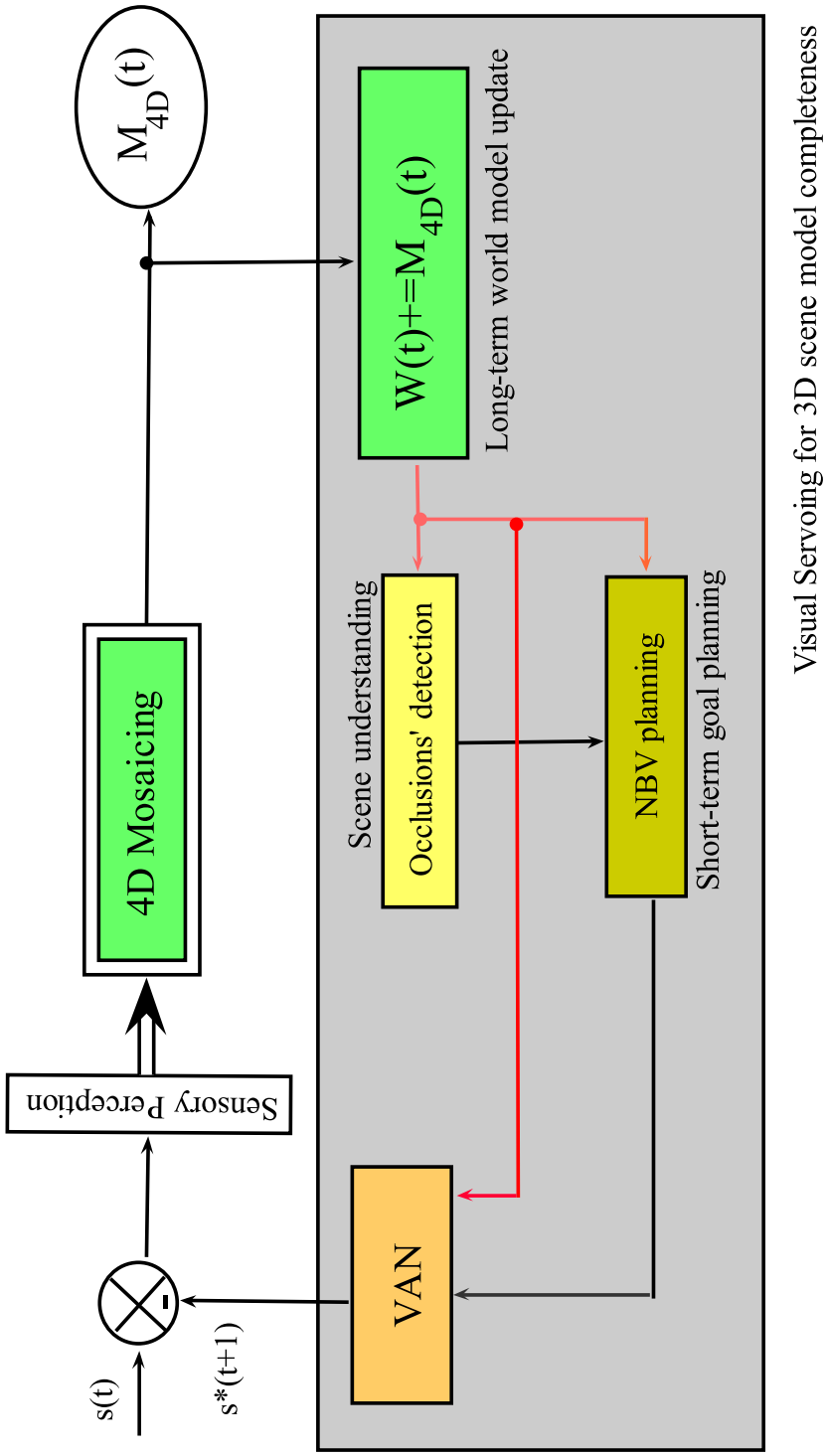


Figure 7.3: Visual-actuated 4D mosaicing sensor for site digitization and exploration.

This is the main ingredient of the visual feedback loop as it allows for several functionalities to be computed based on the currently built 3D model, such as: occlusions detection, view and path planning, obstacle detections and system's localization.

(b) 4D Mosaic merging. Once the 3D pose has been computed, the 4D mosaic

views must be merged in a common reference system in order to produce a global scene model. During this stage, special attention must be given to noisy points which must be discarded using down-weighting rules and to data redundancies. The acquired mosaics are complementary, i.e. hardly- or non-visible areas in one view were sensed from other viewpoints. Therefore, space carving methods [Kutulakos and Seitz, 2000] must be used in order to integrate data sensed from all viewpoints covering the occluded areas.

2. Next Best View (NBV) planning. In order to ensure in-situ the 3D scene model completeness, the system exploits the global 3D scene model $W(t)$ generated so far to find the best next 3D position of the system from where the new 4D mosaic in order to maximize the visible volume, while minimizing cost's displacement and data redundancy. The NBV's computation process requires knowledge about depth discontinuities in the 3D model acquired so far $W(t)$. This might be seen as an **occlusion-based exploration** strategy.

3. Autonomous navigation to rich the NBV. Supposing that a NBV position $s^*(t+1)$ has been computed, the system must be capable to navigate autonomously and reach it. This calls for obstacles' detection, self-localization, path planning and motion control procedures.

All the aforementioned functionalities are composing the control loop of the ARTVISYS system, actuating the 4D mosaicing sensor to perform site digitization and exploration.

The main process powering the entire visual feedback loop is represented by the 4D mosaicing matching procedure (step **1.a**) which allows to integrate all 4D-mosaic views within a global 3D scene model. For this reason, the following sections are dedicated to the 3D model matching process, illustrating the potential of the 4D mosaicing sensor in solving reliably for the data matching task in GPS-denied and unstructured environments. Since the remaining processing blocks composing the visual feedback loop are strongly related to the 4D mosaicing sensor's features, we provide several research directions in the closing section of this chapter.

7.3 The 3D Model Matching Problem

The recent advances in 3D modeling lead to a growing interest for retrieving 3D models from real world [Snavely et al., 2006]. With help from 3D graphics hardware and CPUs, 3D data has become cheap enough to be processed and displayed on a general computer. The world wide web hosts 3D models gathered by peoples from all over the world, allowing to spread high-quality 3D models. (avalon.viewpoint.com). The issue moves from "how we generate 3D models?" to "how we recognize them?". This is an active research topic in shape-based recognition, retrieval, clustering and classification.

Up-to-date research work reported real-time performances on large-scale 3D reconstruction from video data [Pollefeys et al., 2008]. Although image-based matching techniques provide loop-closing constraints to the bundle adjustment, they are powerless when dealing with 3D surface irregularities. When using only a texture-based criterion it is difficult or even impossible to match low-overlapped 3D models acquired under high viewpoint changes in large-scale environments. In addition, as stated in [Wu et al., 2008], the accuracy required by ICP is not achievable with 3D point clouds recovered from stereo. Several limitations were also encountered when attempting to match 2D features in complex environments where due to the accidental terrain the problem gets even worse in absence of odometry.

This issue is addressed in [Wu et al., 2008], were Viewpoint Invariant Patches - VIP are introduced for matching textured 3D models produced from video data. The VIPs are designed as orthonormal patch projections on which SIFT descriptors are computed. The VIP descriptor is subject to the orthographic projection which is valid only on planar surfaces. Since our research work is concerned with the 3D model matching in unstructured environments, planar surface hypothesis required to build orthonormal patch projection is not validated. Although the matching results illustrate that VIPs outperform SIFTs, the accuracy of the pose estimation is limited by the error of the passive 3D geometry recovery process.

It is therefore necessary to acquire accurate 3D geometry and to provide more reliable matching while avoiding the use of salient features in order to deal with unstructured environments. For this reason, several research works including this one have directed their studies toward image-laser solutions.

A very similar work to ours is presented in [Miro and Dissanayake, 2008], in which authors propose the use of the 3D visual sensor illustrated in Figure 7.4 for SLAM to supply SAR missions in unstructured environments. Authors employ a 3D camera manufactured by MESA® and a high-resolution color camera mounted on it, being fixed wrt each other. The two sensors are used to extract 3D features and to design a SLAM algorithm based on an Extended Information Filter (EIF) [Thrun et al., 2005]. SIFT features extracted in the textured images that can be identified in the range images, giving access to the full 3D information which is exploited to retrieve the sensor motion between two consecutive camera poses via a least-square fitting procedure [Arun et al., 1987] and RANSAC [Fischler and Bolles, 1981]. The new transformation can be used to combine several textured range images to form a global model of the environment [Ellekilde et al., 2005]. On the downside, the MESA range-sensor is limited to 10 m range, being therefore unsuitable for large-scale sceneries.

Figure 7.4: Image-range solution presented in [Miro and Dissanayake, 2008]. (a) MESA and camera montage. (b) The rescue arena University of Technology Sydney illustrating the difficulty to navigate in unstructured environments.

7.4 4D Mosaic-driven Acquisition Scenario in the Tautavel Prehistoric Cave

A first data acquisition scenario was performed in the Tautavel prehistoric cave in October 2007. In order to illustrate the difficulty to acquire data, Figure 7.5 depicts an indoor view of the cave, illustrating the area where the system could not be placed. In this case, an aerial digitization system could be more suitable to be used.

In this context, we acquired three stations, one situated at the cave's entrance, a second station after the forbidden whole and a third station in the cave's background area. Figure 7.6 illustrates the corresponding 2D mosaic views generated using the algorithm

Figure 7.5: A global view of the interior of the Tautavel prehistoric cave. The red rectangle highlights the forbidden area for sensor's positioning.

proposed in Chapter 6 which are down-sampled in order to allow fast computing. The goal is to recover the 3D rigid transformation relating the 4D mosaic views. When looking at Figures 7.6 (a) and (b), the difference in viewpoint is considerably high and when visually inspected, common areas can barely be detected.

Figures D.5 and D.6 from Appendix D.3 illustrate that Harris and SIFT features' were extracted in a sufficient number. In Figure 7.7 (c) several matches can be visually inspected, showing the week capacity for matching reliably Harris features between panoramic views acquired under high viewpoint variations and illumination changes. In addition, SIFT matches were found in a limited number (4 and 8 false matches were found at levels 1 and 0, respectively) allowing us to conclude that traditional 2D features do not allow for outlier-free pose estimates and that the matching process must be driven by 3D information.

In our research work we are interested in coding 3D descriptors allowing to reliably

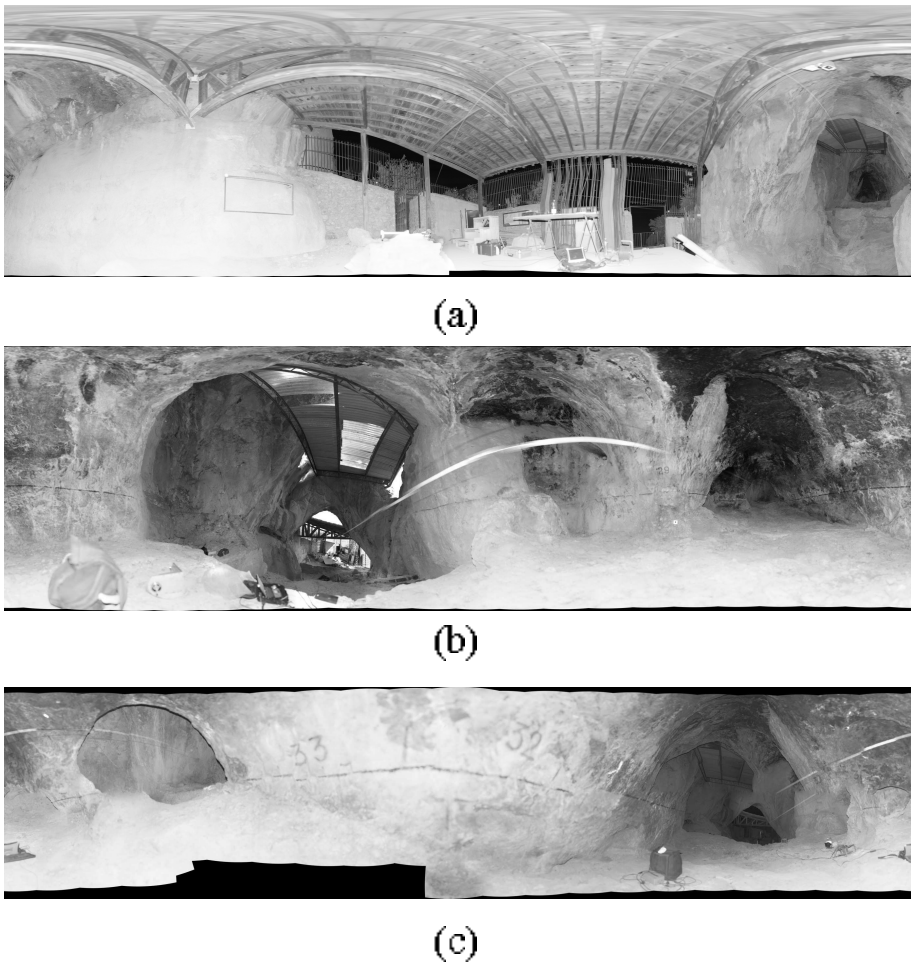


Figure 7.6: 2D Mosaic views acquired in Tautavel prehistoric cave. (a) caves entrance: M^1_{2D} size 2710×816, (b) cave's middle M^2_{2D} size 2771×811, (c) cave's bottom M^3_{2D} size 2775×652.

match 3D models in feature-less environments over large-scale scenes. To this end, the next section proposes a 3D model matching procedure which exploit the 4D mosaicing sensor capabilities to solve reliably for data matching problem over large-scale scenes.

7.5 4D-Panoramic-based Solution for Automatic 3D Model Matching

When designing an automatic solution, one has to take into account the three main stages of the data matching pipeline: (a) descriptor extraction, (b) feature matching, and (c) pose estimation.

The proposed framework takes advantage of the panoramic views which provide long-term feature tracking and exploits the accuracy of the 3D geometry recovered through LRFs to design hybrid descriptors for eliminating the ambiguity of the traditionally 2D features matching methods. To this end, we design a class of hybrid descriptors, encoding both appearance and geometry information which will drive the matching process toward

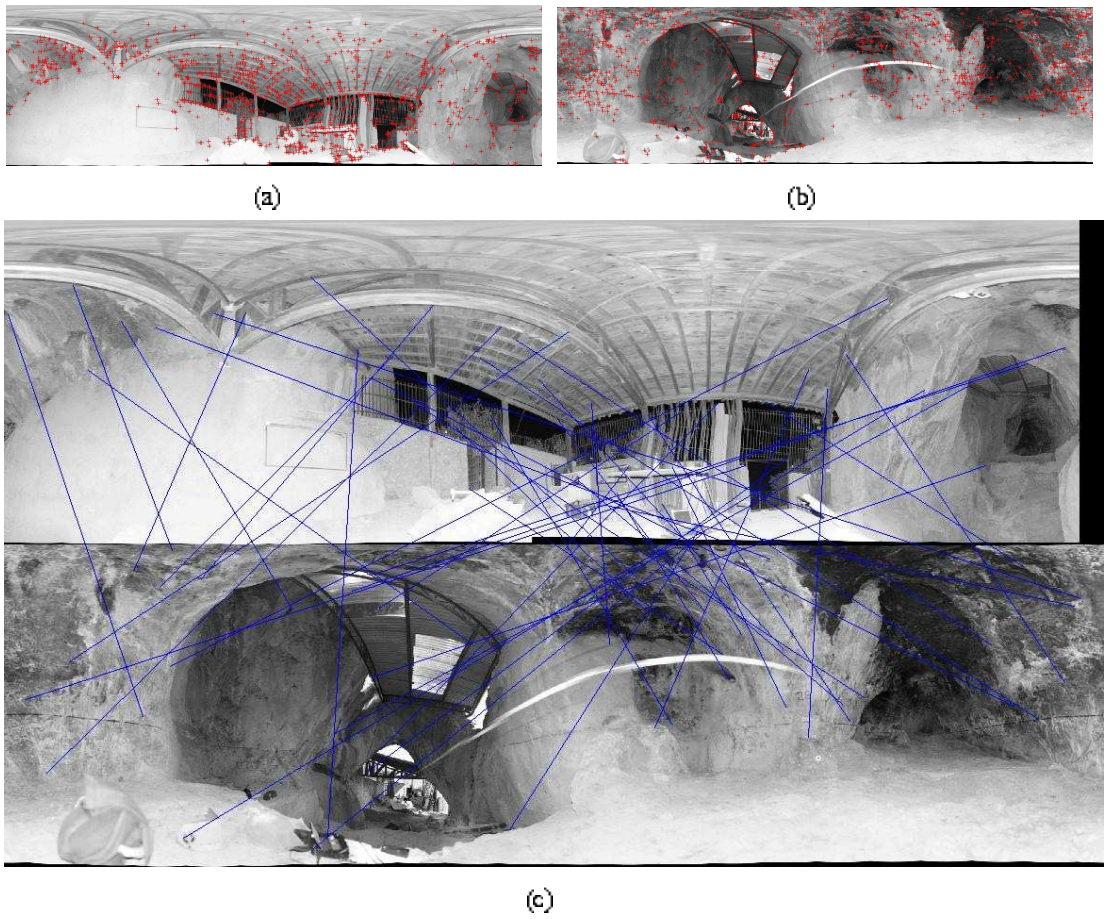


Figure 7.7: Harris matches extrction on M^1_{2D} and M^2_{2D} mosaic views. ♯1136 matches between (a) M^1_{2D} and (b) M^2_{2D}. (c) one match over 75 is displayed.

reliable data matching in feature-less environments by filtering out the false matches based on intrinsic geometrical properties and topological consistency between triads belonging to the same 4D mosaic view, noted M_{4D}. Figure 7.8 illustrates the proposed approach which is detailed throughout the following subsections.

7.5.1 Viewpoint Invariant Hybrid Descriptors - VIHD

We proposed the design of hybrid features, $\mathbf{d} = [(x, y, z), (u, v)^T, \mathbf{t}, \vec{\mathbf{n}}]^T$ encoding the 3D coordinates $(x, y, z)^T$, the pixel location in the 2D projection of the 4D mosaic $(u, v)^T$, the color components $\mathbf{c} = [R, G, B]^T$, and the normal $\vec{\mathbf{n}}$ to the tangent plane at the surface at the 3D coordinates of the descriptor $\mathbf{d}$, given by $(x, y, z)^T$.

Various characteristics may be added, following the features which might be relevant to the environment description. In our case, it is very useful to encode the entropy of the patch (η), characterizing the variation of the surface geometry. Since salient descriptors are subject to illumination changes, causing lost features, we recommend to use the unprocessed color provided by texture maps.

The anonymous features resulted from the 2D Giga-mosaicing process described in Chapter 5 can used to by-pass the extraction phase. Additional processing stages are required for uniform sampling and to achieve viewpoint-invariance. In complex environ-

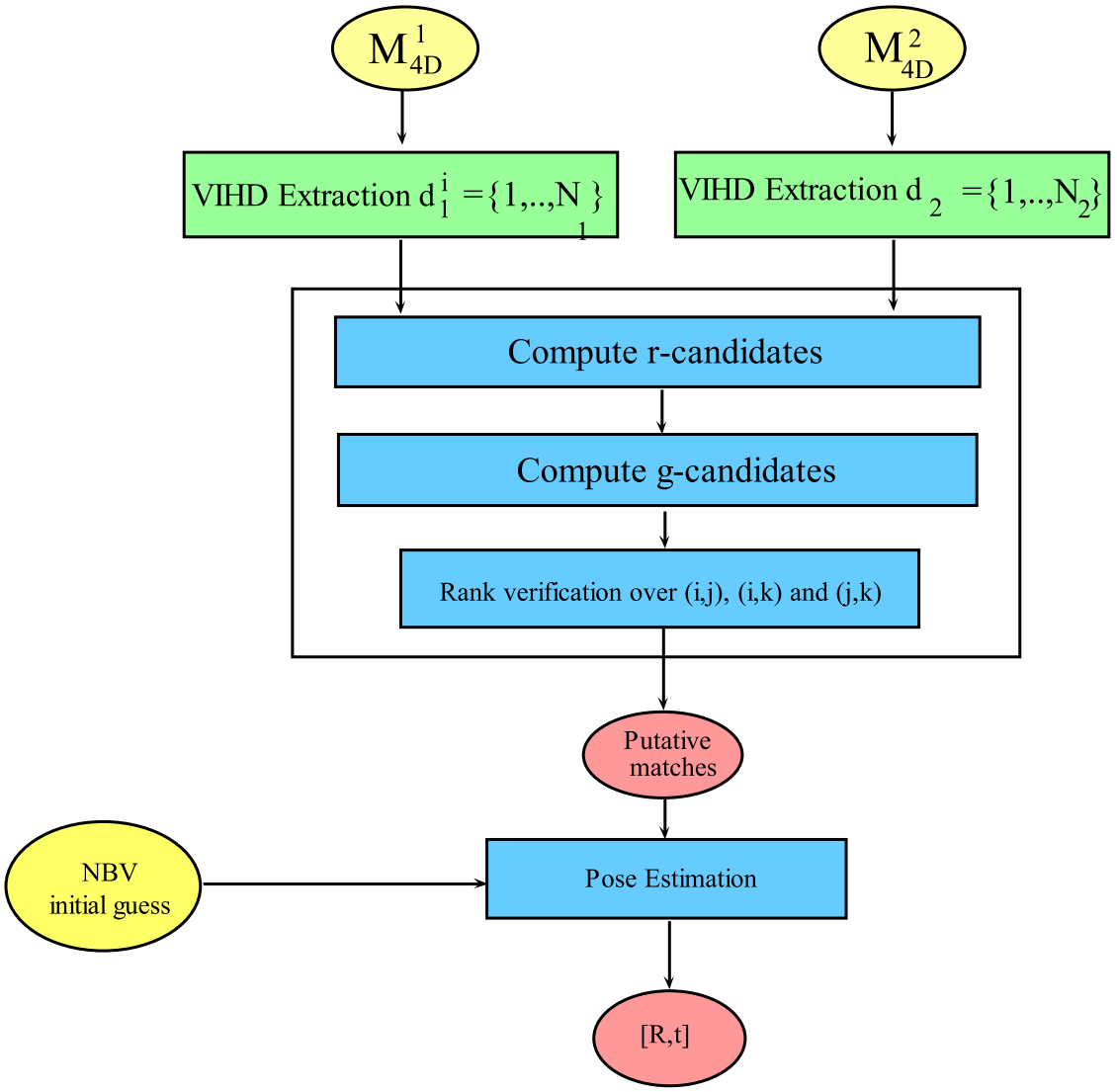


Figure 7.8: The proposed image-laser solution provided by the 4D mosaicing sensor for data matching and pose computation.

ments, the orthonormal projection is not validated and for this reason we suggest the use of circular patches projected on the sphere surface using curve fitting.

7.5.2 Unambiguous Matching

Suppose that a list of descriptors were extracted from each mosaic $D[M^1_{4D}] = \{\mathbf{d}^1_i, i = 0,..,n_{D_1}-1\}$ and $D[M^2_{4D}] = \{\mathbf{d}^2_j, j = 0,..,n_{D_2}-1\}$, where $n_{D_i}, i = 1,2$ denote the number of the VIHD descriptors extracted in M^i_{4D} with $n_{D_1} < n_{D_2}$. Our matching strategy follows two stages. First a radiometric criterion establishes *r-candidates* which are verified in a second stage using a geometric criterion, producing *g-candidates*.

Finding radiometrically (R) eligible candidates matches. Each feature $\mathbf{d}^1_i \in D[M_{4D}]^1$ is matched against each feature $\mathbf{d}^2_j \in D[M_{4D}]^2$ to form a cluster with the best R "eligible" candidates which might correspond to $\mathbf{d}^1_i$. The eligibility criterion is based on

the ZNCC score computed over a squared windowed area of size w.

In order to deal with occlusions, we must be able to decide whether a feature point is lost or not. This aspect is taken into account by introducing the following rule:

$$\begin{cases} ZNCC(\mathbf{d}_i^1, \mathbf{d}_j^2) \in [-1, 0.5] \rightarrow w(i,j) = 0 \\ ZNCC(\mathbf{d}_i^1, \mathbf{d}_j^2) \in (0.5, 1] \rightarrow w(i,j) = 1 \end{cases} \tag{7.1}$$

Eligible candidates for $\mathbf{d}_i^1$ are expressed under the following form:

$$\mathbf{Q}_{(i)} = \{\mathbf{d}_{i,q,r}^2, q = 0, .., R-1 | w(i,j) = 1\} \tag{7.2}$$

where q is the counter and r denotes the rank of features $\mathbf{d}_j^2 \in M_{4D}^2$, with $card(\mathbf{Q}_i) = R$. As shown in Figure 7.9, this phase of the algorithm outputs a bag of *R-candidates* for each feature $\mathbf{d}_i^1$ extracted from M_{4D}^1.

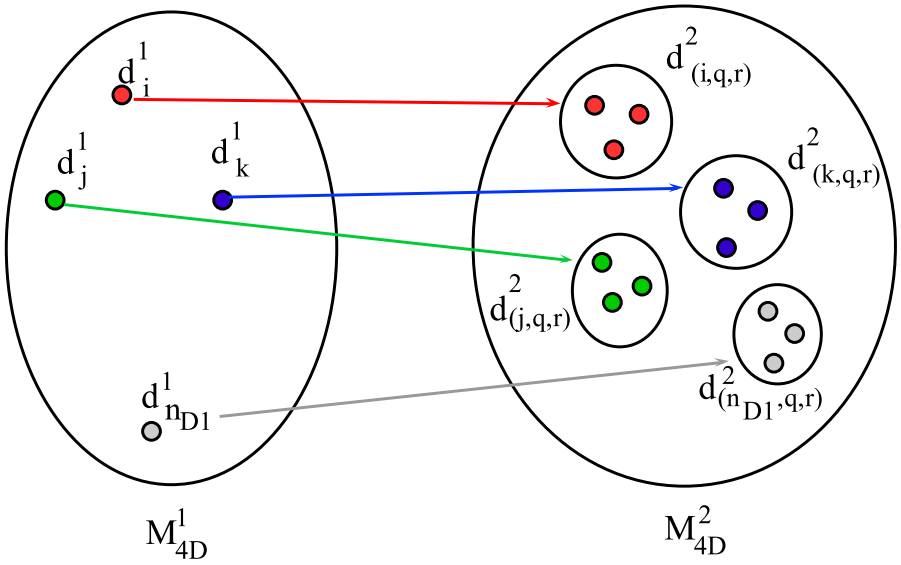


Figure 7.9: R-candidates for each features $\mathbf{d}_i^1 \in M_{4D}^1, i = 0, .., n_{D1} - 1$.

Selecting geometrically (G) consistent triads. For each candidate pair $\mathbf{d}_i^1 \leftrightarrow \mathbf{d}_{i,q,r}^2$ the geometrical coherence is verified wrt the inter-mosaic topological relationship between descriptor triads $(\mathbf{d}_i, \mathbf{d}_j, \mathbf{d}_k)^1 \in D[M_1]$ and their corresponding bag of *R-candidates* $(\mathbf{d}_{i,q,r}, \mathbf{d}_{j,q,r}, \mathbf{d}_{k,q,r})^2 \in D[M_2]$ established at the previous step.

The inter-mosaic topological consistency is defined in terms of Euclidian norm and angular distance between feature couples taken over feature triads in each mosaic. The geometrical consistency is verified on pairs (i,j), (i,k) and (j,k) composing each triad and provides potential matches by comparing the spatial relationship between features in M_{4D}^1 and features in M_{4D}^2.

Valid homologous descriptors composing the triad $(\mathbf{d}_i, \mathbf{d}_j, \mathbf{d}_k)^1$ are found by searching over triads formed by *R-candidates* $(\mathbf{d}_{i,q,r}, \mathbf{d}_{j,q,r}, \mathbf{d}_{k,q,r})^2 \in D[M_2]$, the one which verifies the inter-mosaic topological relationship, noted $(\mathbf{d}_{(i,r)}, \mathbf{d}_{(j,r)}, \mathbf{d}_{(k,r)})^2$. The neighboring relationship is defined in terms of norm and angular distances between the corresponding 3D points $\mathbf{p} = (x, y, z)^T$:

$$\mathcal{V}(i,j) : \begin{cases} \mathcal{V}_1 : \|\mathbf{p}_i^1 - \mathbf{p}_j^1\| = \|\mathbf{p}_{i,q,r}^2 - \mathbf{p}_{j,q,r}^2\| \pm \xi_{\|\|} \\ \mathcal{V}_2 : \mathbf{p}_i^1 \times \mathbf{p}_j^1 = \mathbf{p}_{i,q,r}^2 \times \mathbf{p}_{j,q,r}^2 \pm \xi_{\times} \end{cases} \tag{7.3}$$

The quantities $\xi_{|||}$ and $\xi_{\times}$ define the norm and the angular tolerance measures, respectively which can be established through automatic thresholding methods [Sezgin and Sankur, 2004]. Pairs of features verifying simultaneously $\mathcal{V}_1$ and $\mathcal{V}_2$ provide a list of *G-candidate* pairs $(\mathbf{d}^2_{i,q,g} \leftrightarrow \mathbf{d}^2_{j,q,g})$, with $card\{\mathbf{Q}_{i,g}\} = G$, where g denotes the ranking of the best G candidates in a decreasing order. The g-candidates $(\mathbf{d}^2_{i,q,g} \leftrightarrow \mathbf{d}^2_{j,q,g})$ are topologically consistent with $(\mathbf{d}^1_i, \mathbf{d}^1_j)$ within the bounding domains $\square\xi_{|||}$ and $\square\xi_{\times}$. A similar procedure is performed for pairs (i,k) and (j,k) leading to *G-candidates* pairs $(\mathbf{d}^2_{i,q,g} \leftrightarrow \mathbf{d}^2_{k,q,g})(i,k)$ and $(\mathbf{d}^2_{j,q,g} \leftrightarrow \mathbf{d}^2_{k,q,g})(j,k)$ which are topologically consistent with $(\mathbf{d}^1_i \leftrightarrow \mathbf{d}^1_k)$ and $(\mathbf{d}^1_j \leftrightarrow \mathbf{d}^1_k)$, respectively. The proposed topological consistency criterion can be formalized as following:

$$\begin{cases} \mathcal{V}(\mathbf{d}_i, \mathbf{d}_j)^1 \Leftrightarrow \mathcal{V}(\mathbf{d}_{i,q,g}, \mathbf{d}_{j,q,g})^2 \\ \mathcal{V}(\mathbf{d}_i, \mathbf{d}_k)^1 \Leftrightarrow \mathcal{V}(\mathbf{d}_{i,q,g}, \mathbf{d}_{k,q,g})^2 \\ \mathcal{V}(\mathbf{d}_j, \mathbf{d}_k)^1 \Leftrightarrow \mathcal{V}(\mathbf{d}_{j,q,g}, \mathbf{d}_{k,q,g})^2 \end{cases} \tag{7.4}$$

The corresponding valid triad $(\mathbf{d}_{i,r}, \mathbf{d}_{j,r}, \mathbf{d}_{k,r})^2$ is found by confronting the ranks of the homologous features for each couple composing the triad, i.e. the following three conditions must be simultaneously verified:

$$\begin{cases} \mathbf{d}^2_{\mathbf{i},r} : \mathcal{V}(\mathbf{i},j) \wedge \mathcal{V}(\mathbf{i},k) \Leftrightarrow \text{rank}(\mathbf{d}^2_{i,q,g})(\mathbf{i},j) = \text{rank}(\mathbf{d}^1_{i,q,g})(\mathbf{i},k)) \\ \mathbf{d}^2_{\mathbf{j},r} : \mathcal{V}(i,\mathbf{j}) \wedge \mathcal{V}(\mathbf{j},k) \Leftrightarrow \text{rank}(\mathbf{d}^2_{j,q,g})(i,\mathbf{j}) = \text{rank}(\mathbf{d}^1_{j,q,g})(\mathbf{j},k)) \\ \mathbf{d}^2_{\mathbf{k},r} : \mathcal{V}(i,\mathbf{k}) \wedge \mathcal{V}(j,\mathbf{k}) \Leftrightarrow \text{rank}(\mathbf{d}^2_{k,q,g})(i,\mathbf{k}) = \text{rank}(\mathbf{d}^1_{k,q,g})(j,\mathbf{k})) \end{cases} \tag{7.5}$$

As shown in Figure 7.10, the above tests lead to a geometrical-consistent triad:

$$(\mathbf{d}_i, \mathbf{d}_j, \mathbf{d}_k)^1 \leftrightarrow (\mathbf{d}_{i,r}, \mathbf{d}_{j,r}, \mathbf{d}_{k,r})^2 \tag{7.6}$$

and three homologous pairs:

$$(\mathbf{d}^1_i \leftrightarrow \mathbf{d}^2_{i,r}, \mathbf{d}^1_j \leftrightarrow \mathbf{d}^2_{j,r}, \mathbf{d}^1_k \leftrightarrow \mathbf{d}^2_{k,r}) \tag{7.7}$$

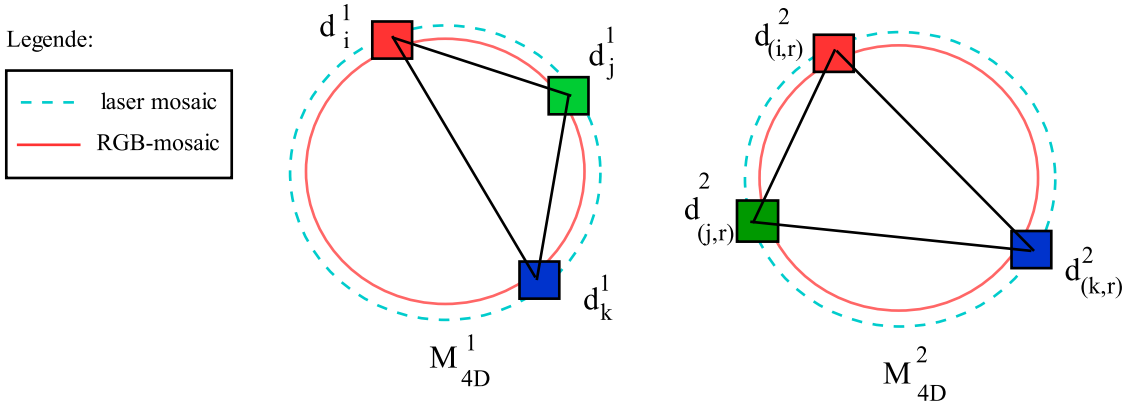


Figure 7.10: VIHD Matching.

The matching process continues by exploiting the circular positioning of features and searches for a geometrically consistent triad with $(\mathbf{d}_{i+1}, \mathbf{d}_{j+1}, \mathbf{d}_{k+1})^1 \in M^1_{4D}$. To this end, the algorithm continues by testing the next triad $(\mathbf{d}_{i,r+1}, \mathbf{d}_{j,r+1}, \mathbf{d}_{k,r+1})^2 \in M^2_{4D}$, as shown in Figure 7.11. The process leads to a list of homologous hybrid descriptors $(\mathbf{d}^1_m \leftrightarrow \mathbf{d}^2_m), m = 0, .., n_{D12} - 1$ which are used for the 6-DOF pose estimation process for which several solutions are provided in the next section.

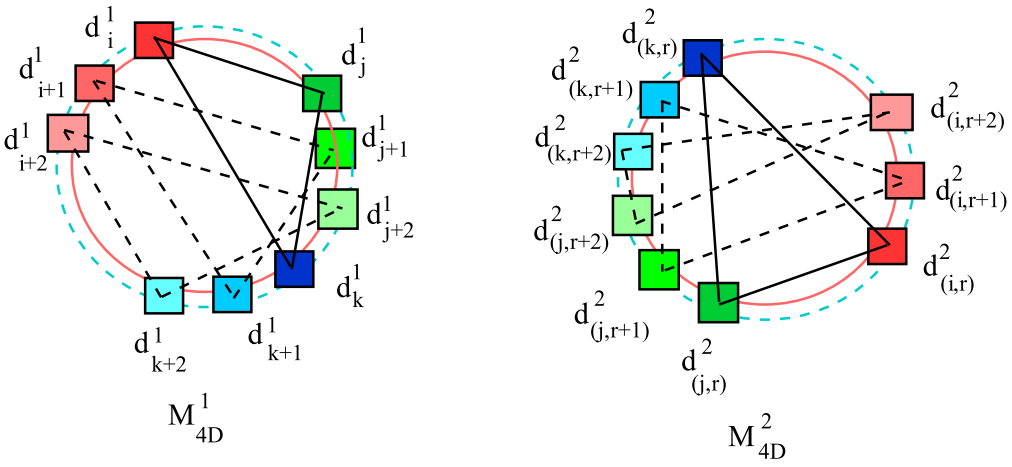


Figure 7.11: VIHD-triad matching process exploiting the circular positioning of features provided by the mosaic view.

7.5.3 6-DOF Pose Estimation using Next Best View

Following the number of homologous features available, several pose estimation schemes are available for the estimation process.

- **Closed-from solutions.** The main advantage of these methods is that they provide the solution in a single step. Reported techniques address the problem by splitting the estimation process. A closed form solution for absolute quaternion computation can be found in [Horn, 1987]. In the same paper, a solution for computing the 3D transformation using matches of three non-collinear points is presented. Once the rotation is computed, the translation can be recovered using the difference of the centroids between the reference points and the rotated points in the second reference system.

- **LS-solutions.** In opposite to closed-form solutions, the parameters are estimated simultaneously but iteratively. The process minimizes the sum of squares of residual errors iteratively until a negligible error is obtained. Nevertheless, for practical issues, when an initial estimation is available it can be used to reduce the number of iterations.

In this dissertation, our main concern is to design a pose estimation process which can take advantage of the 4D mosaicing sensor and its integration within the autonomously site digitization and exploration system introduced in Chapter 3 - ARTVISYS. In this context, the pose estimation process can benefit of the NBV process integrated within the ARTVISYS's software architecture to initialize the pose estimation procedure. This is an advantage provided by the 4D mosaicing sensor it-self, since the NBV process exploits the currently 3D built model. The 3D pose can be refined using non-linear optimization schemes, such as Levenberg-Marquardt [Moré, 2006]. A similar estimation scheme is introduced by Low in [Low, 2006] which employs the NBV position to provide a coarse alignment for 3D model registration via ICP algorithm.

7.6 Conclusions and Future Research Directions

This chapter proposes the integration of the 4D mosaicing sensor within a vision-based system architecture for supplying autonomous site surveys in complex and difficult to access environments. The use of the 4D mosaicing sensor provides image-laser solutions which allow to address the automation of the 3D modeling process. Moreover, when integrated within a visual feedback control loop, the 4D mosaicing sensor allows to ensure the 3D scene model completeness. The research perspectives of this chapter are focusing on providing several research directions for the processing blocks composing the visual control loop of the ARTVISYS system.

7.6.1 Conclusions

This chapter investigates the potential of the 4D mosaicing sensor in solving for several problems intrinsically related to the automatic world modeling process, ranging from unmanned system's autonomy needed to explore autonomously an unknown environment, passing through visual servoing procedures required to ensure the 3D scene model completeness and finishing with the pose estimation between partially overlapped 4D mosaic views to provide a global 3D scene model - being the main input of the visual servoing procedures. We summarize hereafter the main investigations presented in this chapter.

Vision-based autonomy model. Since the autonomous site digitization and exploration problem is intrinsically related to the unmanned system's autonomy, we exploit the potential of the 4D mosaicing sensor to establish a vision-based autonomy model, highlighting two aspects: (i) evaluate how far we can exploit 3D vision techniques to provide autonomy to unmanned mobile platforms, and (2) apply the proposed vision-based autonomy model to solve for the site digitization and exploration problem, giving rise to an artificial vision-based system - ARTVISYS.

ARTVISYS: integration of the 4D mosaicing sensor within an autonomous 3D site digitization and exploration system. When designing ARTVISYS, we have taken into account different solutions ranging from exploration to autonomous navigation, passing through SLAM and photorealist 3D world modeling techniques. The proposed system is willing to fill the gap between the existing methods by integrating the photorealist 3D modeling capability provided by the 4D mosaicing sensor within a visual control loop in order to ensure the 3D scene model completeness.

The main nucleus of ARTVISYS is represented by the 4D mosaicing process which generates automatically omnidirectional 3D scene models, while providing visual data for scene understanding, view-planning and autonomous navigation procedures. The aforementioned procedures form the visual feedback loop of the ARTVISYS system, which provides onboard reasoning and decisional resources in order to ensure in-situ the 3D scene model completeness.

ARTVISYS-solutions for automatic 3D modeling. We shown the high potential of the 4D mosaic views in addressing efficiently several open issues in automatic 3D world modeling: (i) the elimination of data matching ambiguities in feature-less areas using image-laser fusion, (ii) long-term features tracking offered by panoramic views and (iii) the use of the NBV computed from the currently built 3D model which accelerates considerably the pose estimation process between poorly-overlapped views.

- **6-DOF pose estimation.** We introduced a visual feedback system architecture to ensure the 3D scene model completeness and we focus on its main ingredient: *com-*

pute the pose between several sensors readings in order to (1) align and integrate the currently acquired 4D mosaic view in a global 3D scene model, (2) to provide localization of the platform and (3) to allow for occlusion detection for view planning. To this end, we proposed the use of hybrid descriptors to disambiguate between feature matches, constraining the matching process with a 4D-panoramic driven matching procedure.

- **Unambiguous data matching using image-laser hybrid descriptors.** The 4D mosaicing sensor enables the extraction of hybrid descriptors from image and laser data. Their use is recommended together with a two steps matching algorithm: first, radiometric candidates and established which are further filtered using an intra-mosaic topological consistency criterion verified on triads belonging to the same mosaic views.

7.6.2 Future research directions

The research directions of this chapter are directed toward the exploitation of the 3D scene model to supply the visual control loop in order to ensure the 3D scene model completeness, being actually one of the main near feature research perspectives of this dissertation. The proposed 4D mosaic matching procedure takes in charge the generation of the global 3D scene model which feeds the processing blocks composing the visual control loop. The development of the remaining functionalities (4D mosaic merging, NBV computation, visual-based navigation procedures and sensor's control) must cope with several time and in-situ access constraints.

View Planning. The system must be able to use the already generated 3D scene model in order to compute the next best 3D pose from which a new 4D mosaic must be acquired. The next best view process is a function of the detected occlusions, the currently generated 3D scene model, subject to the system's kinematics and to the stationing possibilities provided by the environment. For detecting the stationing possibilities, the system must characterize its surroundings from visual data in order to detect whether the terrain allows it's stationing in that area. In particular, the system must detect flat or non-flat terrain, humidity or other extern factors which may impact on the system stationing during 4D mosaics' acquisition. The output is further exploited by the next 3D pose estimation procedure which provides the main feedback for system's control.

Path Planning. The system must be able to perform path planning in order to navigate from the current position to the estimated next best 3D pose computed at the previous step. Nowadays, path planning procedures rely on machine learning processes leading to simple and binary reasoning, limiting the system's capacity to react rapidly in unpredictable situations and to deal with uncertainty. In order to avoid the learning step which is time consuming and limits system's capacities, we propose to employ a learning-through-history and reinforcement learning schemes, giving the capability to extrapolate semantics and actions empirically. Such an approach gives the possibility to act intelligently on the fly, while accomplishing the required mission, i.e. the 3D scene model completeness in our case study. The path planning procedure exploits the currently generated 3D scene model which provides environment perception and awareness about the system's surroundings. The 3D scene model can be exploited to infer semantics about the environment which must be taken into account within the path planning process.

Autonomous navigation. After generating the path, the system must be able to execute it by navigating autonomously through the environment, without human operator

intervention. This is a hard task to achieve which nowadays is accomplished by sending visual data to a host where is processed by computer vision experts to send commands to the target [Griffiths et al., 2006], [Mathies et al., 2007], [Li et al., 2007]. For these reasons, currently existent systems do not have enough capacity for taking onboard decisions and their real applications rely on heavily human operator intervention. This highly-dependency on human interaction is subject to memory bandwidth and communication latency, causing unmanned systems failure to react rapidly to unpredictable situations. Autonomous navigation calls for:

- *Simultaneous Localization and Mapping procedures (SLAM).*

  As mentioned in Section 2.5.3, the proposed image-laser site digitization scheme allows to solve for the SLAM problem efficiently in feature-less and GPS-denied areas. The framework is powered by the 4D mosaic views, providing both appearance information and geometry. Their joint use allows for reliable data matching for localization purposes, place recognition in feature-less areas and loop closing procedures. Such a dual SLAM scheme provides the capacity to generate photorealist and dense 3D maps of the environments without relying on navigation sensors (GPS, INS, magnetic compasses, dead-reackoning techniques).

- *Obstacle detection and fast decision making capabilities.* The remaining procedures which have to be solved are the obstacle detection and the decision making modules. The latter provides feedback for system's command and control in order to achieve 3D scene model completeness. A good solution would be to supply these processes within a general framework based on biologically inspired computer vision algorithms exlploiting the currently built 3D model.

Perform in-situ the global 3D scene model rendering. Optionally, the 3D scene model rendering can be generated on-the-fly, in a dynamic fashion. One open issue is how to deal with the high amount of data when performing 3D modeling missions in large scale environments. We propose the use of a multi-level scene rendering procedure, capable to generate 3D models with different levels of detail in order to cope with real time and power consumption constraints.

The aforementioned research work represents the design of a vision-based system embedded with automatic 3D modeling capabilities. The proposed system is capable to perform autonomously site digitization and exploration in previously unknown and difficult-to-access environments. The system captures 3D and appearance information to generate dynamically in-situ complete 3D scene model of the environment while localizing it-self within the generated model. This is a basic function which must be embedded onboard mobile systems aimed at performing different missions in complex environments. In a second step, the system can be upgraded with additional functionalities for accomplishing specific tasks in hostile environments, such as scene understanding, exploration, disaster response, searching and rescuing, etc.

Chapter 8

Conclusion and Research Perspectives

This dissertation evaluates the potential of a hybrid image-laser system for generating autonomously complete and photorealist 3D models in challenging environments, without requiring human operator intervention. The presented research focuses on two main aspects: (i) the automation of the 3D modeling pipeline, targeting the automatic data matching in feature-less and GPS-denied areas for in-situ world modeling and (ii) the exploitation of the generated 3D models along with visual servoing procedures to provide unmanned systems with autonomous site digitization and exploration capabilities.

Our investigations are projected into a vision-based system prototype introduced as ARTVISYS which addresses the automation of the 3D modeling pipeline and based on the 3D model generated so far, it explores and digitize the environment to ensure the 3D scene model completeness. The system is aimed at generating textured 3D models encoded as 4D mosaic views which are integrated within a global 3D scene model sequentially. Finally, we propose to exploit the currently built 3D model to provide feedback to the system in order to ensure in-situ the 3D scene model completeness, giving rise to an artificial vision engine powered by a site digitization and exploration process.

8.1 General Conclusions

One might wonder what can be gained from the image-laser fusion and in which measure such a hybrid system can generate automatically complete and photorealist 3D models in difficult to access and unstructured underground environments.

In such environments, special attention must be given to the main issue standing behind the automation of the 3D modeling pipeline which is represented by the capacity to match reliably image and laser data in GPS-denied and feature-less areas. In addition, time and in-situ access constraints require fast and automatic procedures for in-situ data acquisition, processing and interpretation in order to allow for in-situ verification of the 3D scene model completeness. Finally, the currently generated 3D model represents the only available information providing situational awareness based on which autonomous behavior must be build in order to enable the system to act intelligently on the fly and explore the environment to ensure the 3D scene model completeness.

The proposed automated 3D modeling pipeline gave rise to several solutions for automatic data matching from which other stand-alone sub-systems emerged. Our frameworks

are projected into a vision-based system prototype - ARTVISYS embedding software solutions designed to be used along with a mosaic-driven site digitization and exploration acquisition scenario.

Since the autonomous site exploration problem is intrinsically related to the unmanned system's autonomy through the world modeling capability, we first investigate in which measure such a hybrid system can solve for the unmanned system's autonomy. To this end, we propose a vision-based autonomy model and the software architecture of a vision-based system designed to supply autonomously site digitization and exploration missions in difficult to access and unstructured environments.

8.1.1 Contribution to the automation of the 3D modeling pipeline

When studying the image-laser potential in solving for the automation of the 3D modeling pipeline in GPS-denied and feature-less areas, we establish a strategy which solves reliably for the data matching problem through the use of an image-laser solution designed within a panoramic-based framework.

We design a *complementary* and *cooperative* image-laser fusion. The *complementary* aspect is related to the data acquisition. In order to deal with time and in-situ access, the proposed acquisition protocol consists in acquiring low resolution 3D point clouds and high-resolution color images in order to generate photorealist 3D models. Their *cooperative* use lead to in-situ generation of omnidirectional and photorealist 3D models encoded as 4D mosaic views, which are not achievable when using each sensor separately. Since in practice the acquisition of low resolution 3D point clouds in complex environments leads to a high amount of depth discontinuities, one may consider the input as a sparse 3D point cloud.

4D Mosaicing sensor. This dissertation introduces a dual sensor designed to generate in-situ omnidirectional 3D models encoded as 4D mosaic views from image and laser data. The 4D mosaicing algorithm has as main inputs 3D and 2D mosaic views. Although both sensors are rigidly attached, the image-laser 3D motion is unknown and consists of a global 3D rotation and a small inter-sensor parallax. This acquisition strategy allows us to perform occlusion-free image-laser alignment and texture mapping processes, giving rise to omnidirectional 3D models encoded as 4D mosaic views. Moreover, the matching between several 4D mosaic views is driven by color and 3D information, allowing to disambiguate the data matching tasks in feature-less areas.

8.1.2 Contribution to data matching in GPS-denied and features-less areas

Since our research work aims at generating photorealist and complete 3D models of difficult to access and unstructured underground environments, this dissertation attacks the main ingredient standing behind the automation of the 3D modeling pipeline: *the data matching problem in GPS-denied and feature-less areas.* To this end, we propose a camera-laser dual sensor and investigate its potential for addressing the ill-poseness of the data matching problem in complex and unstructured environments.

We first focus to provide solutions for the automatic data alignment problem for simple case, i.e. SVP and small-parallax case, in order to provide image-laser solutions, i.e. the 4D mosaic views, to solve for a more difficult problem: the data alignment under wide viewpoint changes in large scale environments. During our research study, the following data alignment algorithms emerged, each of which being usable as independent processes.

Free- and small-parallax data alignment in feature-less areas. We solve for the pose estimation problem for free- and small-parallax acquisition scenarios within accurate and environment-independent frameworks.

We developed pair-wise pose estimation procedures for both image and laser data, based on which multi-view alignment algorithms were designed to generate in-situ 2D and 3D mosaic views, respectively.

- **Laser data matching for in-situ 3D mosaicing.** Chapter 4 presents an automatic scans matching procedure capable to generate in-situ 3D mosaic views by stitching several partially overlapped scans acquired from the SVP via an environment-independent framework. The proposed method is powered by a pair-wise scans alignment procedure, which estimates the rigid rotation using an intensity-based pyramidal framework to compute quaternions via dense correlation. The multi-view alignment procedure employs a graph-based approach used along with the topological inference criterion [Sawhney et al., 1998] to verify the consistency of the pair-wise pose estimates.

- **Image data matching for 2D Gigapixel mosaicing.** We propose a relative rotation estimation scheme for high-resolution image alignment using a patch-based correlation procedure via quaternions. The global rotation estimate is refined via a non-rigid estimation process which outputs a list of homologous image points. Since they do not correspond to any corner-like features, they are introduced as *anonymous features*.

  In order to generate a 2D-Gigapixel mosaic views, we investigate the use of the proposed pair-wise pose estimation scheme within an existent bundle adjustment framework [Kolor, 2005], powered by AF matches. The results obtained let us conclude that the self-calibration step and the use of a 2D residual error lead to a high number of rejected AFs.

  Theoretical solutions. In order to solve for the aforementioned problems, we proposed a closed-form solution for estimating the absolute quaternion by minimizing a residual error measured in the 3D space, i.e. the angle between homologous vectors corresponding to AF pairings given by their cross product. The multi-view fine alignment scheme - *cross-BA* - leads to a sequential process which estimates the absolute quaternions by minimizing the cross product between corresponding 3D vectors in multiple views.

Wide-baseline and unambiguous data matching driven by 4D mosaic views and hybrid descriptors. What is more important about the 4D mosaic views is that they are high level data structures providing means (i.e. geometry and color information) to solve for a more difficult problem, i.e. wide baseline data matching in feature-less areas. When used along with a mosaic-driven acquisition scenario, the 4D mosaic primitives eliminate the feature matching ambiguity by imposing spatial constraints between feature triads belonging to the same mosaic view. This allows us to approach the estimation scheme of outlier-free estimates, which is of prior concern in our research work as it allows the processing of several processing blocks composing the visual feedback loop such as: view planning, trajectory planning, and motion control.

Theoretical solutions. We exploit the 4D mosaic to extract viewpoint invariant hybrid descriptors - VIHD to be used with a two-step data matching process. First, radiometric

candidates and established which are further filtered using an intra-mosaic topological consistency criterion verified on triads belonging to the same mosaic view.

8.1.3 4D Mosaic-driven autonomous site digitization and exploration

Chapter 7 proposes the integration of the 4D mosaicing sensor within a vision-based system to supply site digitization and exploration of difficult to access and unstructured environments.

Since the autonomous site digitization and exploration problem is intrinsically related to the unmanned system's autonomy via the world modeling capability, we first investigate in which measure the 4D mosaicing sensor can solve for the system's autonomy problem and established a purely-visual autonomy model to be embedded onboard mobile unmanned mobile systems designed to supply complex missions in challenging environments, site surveys being one of them.

The proposed visual autonomy model was further instantiated to the autonomous site digitization and exploration case, giving rise to the **ARTVISYS** system which comes together with a 4D mosaic-driven acquisition scenario and embeds automatic softwares for supplying the entire 3D modeling pipeline. The 4D mosaicing sensor represents the nucleus of the 3D world modeling process which powers visual servoing procedures in-charged with the 3D scene model completeness.

Since the processing blocks composing the visual feedback loop exploit the global 3D scene model, we evaluate the 4D mosaic's potential to address the pose estimation problem. To this end, we propose image-laser solutions for disambiguating the data matching process which is inherent to outliers in feature-less areas when using either image or laser data alone.

8.1.4 Software Quality Validation

Figure 8.1 provides an overview of the results emerged from the research studies presented in this dissertation. The proposed algorithms are implemented in C++ and run on a Linux laptop equipped with an Intel 1.66 GHz and 2Gb of RAM memory. At the first side, we focus to solve for the data matching task while runtime issues were addressed using pyramidal frameworks. In addition, initial guess provided by physical instrumentation and calibration constraints are used to limit the searching space.

Experimental results illustrate the performance of the algorithm in both areas: underground unstructured environments (in three prehistoric caves situated in France) and in outdoor structured environments (Paris, France). Moreover, different acquisition scenarios were employed using both fixed and mobile acquisition devices.

8.2 Short-term Research Perspectives

The research perspectives of the near future are concerned with the in-situ demonstration of an improved version of the proposed system, starting with the implementation of several theoretical solutions, passing through their validation and finishing with the in-situ demonstration on the proposed system.

Future improvements and validations of the 4D mosaicing sensor. The first part concerns the improvement of the 4D mosaicing sensor by including the theoretical solutions highlighted in blue in Figure 8.1 and their validation on a recent data acquisition

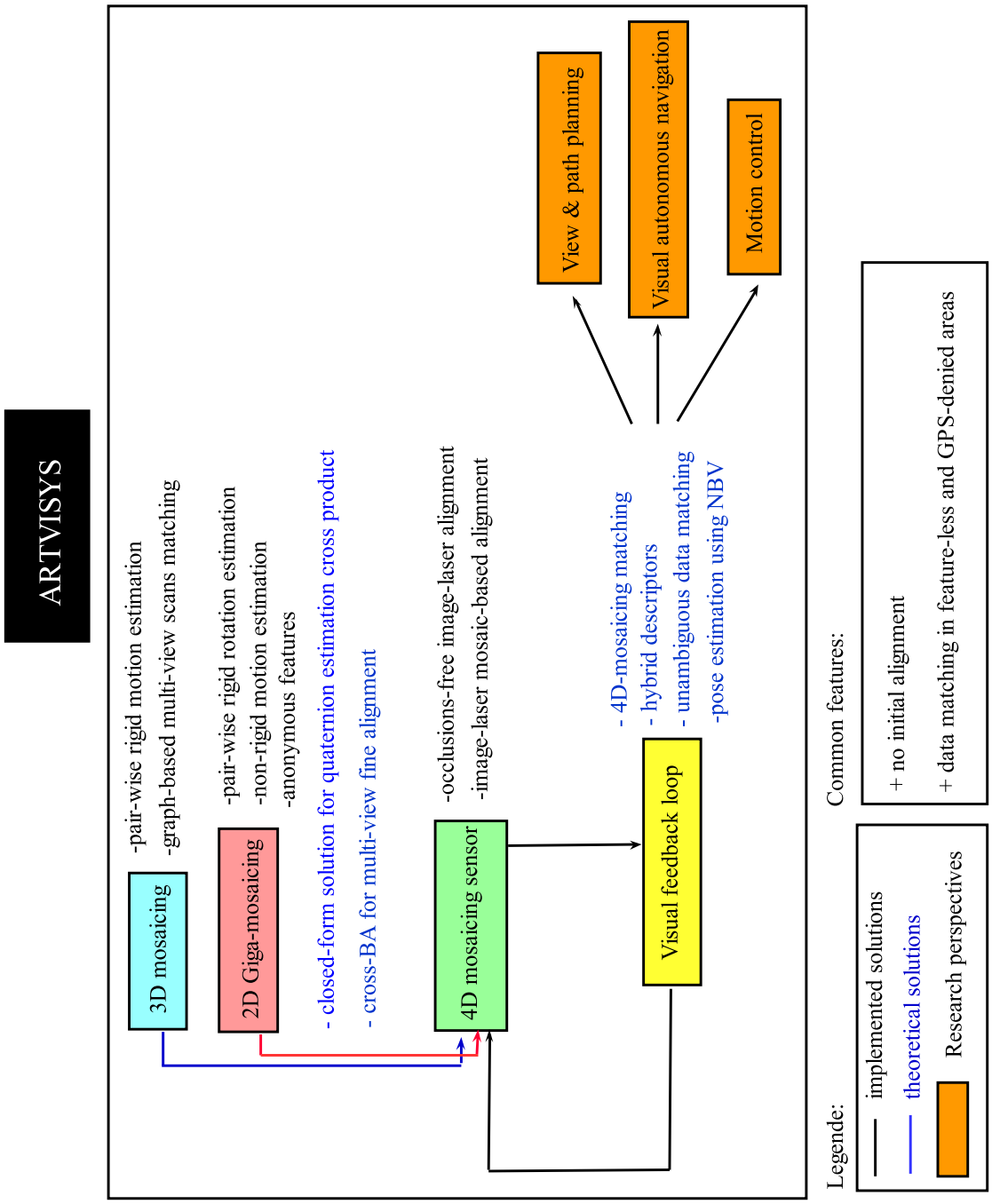


Figure 8.1: Contributions and short-term research perspectives of the research work presented in this dissertation.

campaign recently undertaken by the French Mapping Agency in the Mayenne Science prehistoric cave (France). Since the proposed solutions do not rely on feature-based frameworks, this should lead to an operational system capable to generate automatically 4D mosaic views in both unstructured and structured environments.

Visual feedback loop of the ARTVISYS system. In order to ensure a complete site digitization, a visual feedback loop actuates the 4D mosaicing sensor to generate

dynamically 3D scene models and to act intelligently on the fly in order to explore the environment. The research perspectives of the near future are concerned with the development of the processing blocks composing the visual control loop of the ARTVISYS system for which a brief description is presented in Chapter 7. The research directions are mainly oriented toward the use of biologically inspired computer vision algorithms to develop reinforcement learning ad planning schemes [Bishop, 1998].

Being given the challenging environment presented by the Mayenne Science prehistoric cave, the development of the ARTVISYS system on these data sets should lead to a system capable of demonstrating in-situ the feasibility of autonomous site digitization and exploration missions.

Site digitization and exploration using multiple autonomous cooperative systems.

As stated in [Estlin et al., 1999], utilizing multiple cooperating platforms to achieve the overall goal of the mission has definitely several advantages. Systems embedding complementary sensors can be deployed simultaneously to greatly increase the collected information, giving the possibility to perform tasks whose feasibility is beyond the limits of a single system.

When using such scenarios, open issues are concerned with communication, control [Murray, 2007] and cooperative navigation [Sanderson, 1998], mapping [Rocha et al., 2005] or situational awareness [Touzet, 2000]. Systems share information and exploit it via onboard reasoning resources in order to decide and communicate these decisions. In this context, there is a tradeoff between the communication capabilities and the amount of shared information between systems.

When reaction time is critical, due to communication latency, system's onboard autonomy must be addressed. This includes autonomous navigation and re-planning of its own tasks wrt its surroundings. Indeed, the multi-platform cooperation increases the gain of an exploration mission, but before using it, one must address the system's autonomy problem.

Consequently, a priority is to solve for the system's autonomy in order to maximize the onboard individual capabilities, while limiting inter-systems communication and cooperation to team-working and emergency response situations which are not feasible by a single system, reducing therefore the communication issues. To this end, the next section focuses on the main research perspective of this dissertation which targets to embed human-like intelligence onboard unmanned mobile systems through through vision, giving rise to a fully autonomous system capable to supply both single- and team-working scenarios.

8.3 The use of ARTVISYS as a general-purpose system

When studying the potential of the 4D mosaicing sensor for supplying autonomously site digitization and explorations missions, we were faced to a more general problem: the unmanned systems autonomy. For this reason, in Chapter 7 we investigate in which measure the 4D mosaicing sensor can solve for the autonomy problem and proposed a vision-based autonomy model.

We saw that 3D vision techniques provide powerful resources for mapping and localization together with the possibility to infer semantics about the environment. This allows system's awareness when evolving in an unknown and changing environment for performing complex missions in high-risk environments, relieving human operator's life. Moreover, it

is possible to enrich unmanned platforms with artificial intelligence functionalities powered by a computer vision engine.

Consequently, ARTVISYS represents a promising research direction to solve for the autonomy of unmanned systems. More precisely, its mapping through exploration capability can be exploited as an artificial vision sense to supply reasoning and decisional resources as well as actions via visual servoing procedures. This represents the main research perspective of this dissertation which aims at providing a purely-vision solution for addressing the unmanned systems autonomy problem.

The global scheme of the proposed system illustrated in Figure 8.2 emphasizes the modular design which allows to easily enhance the system's performances. ARTVISYS embeds vision-based environment perception capabilities at which visual servoing algorithms can be added, giving rise to unmanned platforms able to accomplish complex missions in previously unknown and difficult to access environments.

Figure 8.2 illustrates the ARTVISYS's behavior for both cases: when the system runs in open loop and when the visual servoing procedures are included within the main process. During the *open loop* the system generates in-situ a photorealist 3D mosaic, providing a fully 3D spherical view of the system's surroundings for a single 3D pose of the system. The proposed system embeds purely vision algorithms for performing automatically the entire the 3D modeling process which consists in on-line data acquisition and processing, enabling in-situ 3D models generation.

When including visual servoing procedures within the main process, it is possible to enhance the system's performances by integrating different capabilities within the main process.

In our research work, several procedures are required in order to enable the system to explore autonomously an unknown environment for generating complete 3D scene models. They are included in the visual control loop to provide path and view planning, obstacle detection and avoidance and other visual-based autonomous navigation on-board functionalities which are related to the system's kinematics, to the 3D locations where the system is authorized to move within the map and to the detected occluded areas.

The modular design of ARTVISYS allows for ease upgrade by integrating other functionalities for accomplishing a wide variety of complex missions taking place in hostile environment, such as: site inspection and monitoring, disaster response or searching and rescuing missions.

Quantum computing and autonomy model. The high potential of the world modeling capability in providing autonomy to unmanned platforms has been studied through separated processing blocks, leading to a limitation in terms of computational resources for in-situ processing. During today's lunar exploration missions, an image is send on Earth in 25 s and unmanned platforms employs radar units and celestial navigation systems. The presence of computer vision experts in the loop precludes the possibility of analyzing the site and while taking decisions on-the-fly. Driven by the computational issues posed by traditional computing, the visual perception's high potential has not been exploited in its entirety for conceiving a complete visual autonomy model.

Although computers have become considerably faster, they are cursed to manipulate and interpret binary bits to provide computational results and following the Moore's low, the limits of the integrated-circuit-based computing power will be reached by the next decade.

Nevertheless, a new hope arise from quantic computing, which can a quantum bit (qubit), being capable to exist in the classical 0 and 1 states, but also in a weighted su-

perposition of both. An operation on such a qubit acts on both values simultaneously, a two-qubit would act on 4 values and so on, increasing the *quantum parallelism* exponentially. By designing the suitable algorithm it is possible to exploit this parallelism to address computational-burden issues raised by classical computers. Today such algorithms exist in a very few number (e.g. Shor's and Grover's algorithm). Researchers estimate that a traditional computer requires 10 million billion billion years to factorize a 1000-digits number, while Shor algorithm provides the result in 20 minutes.

In this context, we can conclude that there are no functional limits for embedding human-like intelligence onboard unmanned systems in order to solve for the unmanned systems autonomy problem. The issue here is related to the human aptitude to conceive the appropriate program which imitates brain's functionalities. This is a challenging problem, since nowadays neuroscientists are still searching to model brain mechanisms.

Biologically-inspired computer vision algorithms. Exploiting sensory perception to develop biologically inspired computer vision algorithms can bring a useful contribution to design adaptive algorithms to answer to a high-variability of environment types. Moreover, biologically inspired techniques provide valuable insight to new design principles for robotics. In this context, we are confident that computer science and robotics can contribute to a better understanding of biological systems.

This dissertation idea-flow starts with the automation of the 3D modeling pipeline to enable in-situ 3D modeling in high-risk and complex environments. Next, we propose to exploit the currently built 3D model with visual servoing procedures for supplying site digitization and exploration purposes.

Our studies let us concluding that the world modeling capability of ARTVISYS can be exploited as an artificial vision engine to supply several functionalities to solve for the unmanned systems autonomy problem. We design a vision-based autonomy model to be used along with biologically inspired computer vision algorithms implemented using quantum computing.

The proposed research perspective puts together several research fields in computer science, biology, neuroscience, physics and mechanics and robotics, which may benefit together of the impact of such a experimental system to clear several open questions in each area, such as in biology and neuroscience, while solving for the unmanned systems autonomy problem.

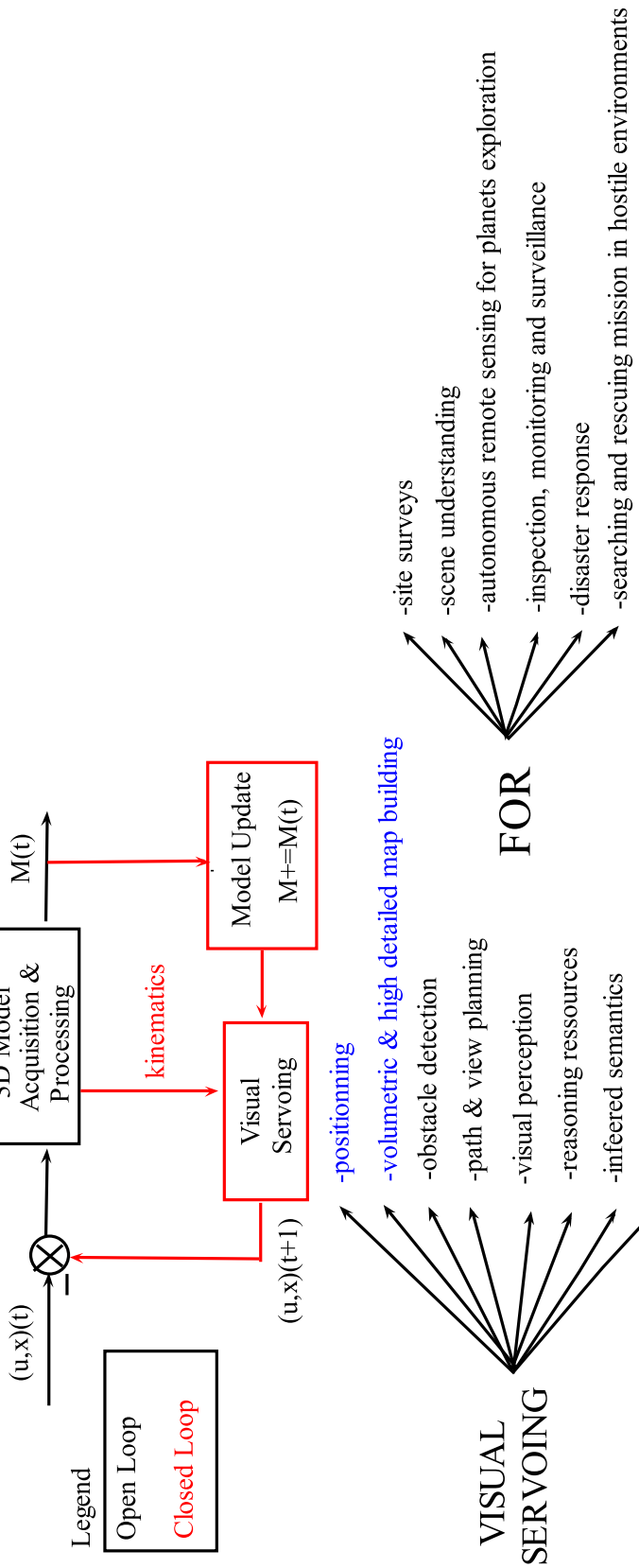


Figure 8.2: Open and closed loop uses of the ARTVISYS system.

Appendix A

Complements to Chapter 2

A.1 Laser-range sensing techniques

This section presents a complement to Section 2.2.2.2 and provides a brief description on active 3D geometry recovery techniques. Based on the way they exploit the reflected laser beam, there are two different laser-range sensing techniques.

Triangulation Sensors. Capturing devices belonging to this class represent the most common range scanners. They recover the disparity measure between the emitted and the reflected laser beam into a laser-sensitive CCD array. Triangulation sensors are equipped with a lighting system which projects a light pattern onto the surface to be scanned. A CCD camera senses the reflected light from the surface and a software provided by the scanner computes an array of depth values, which can be further converted into 3D point coordinates expressed in the scanner coordinate system, by exploiting the calibrated position and orientation of the light source and sensor. The main drawback of the triangulation-based sensors is that they require a suitable clear visible field of view for the source and the sensor in order to observe the surface being scanned. In addition, the quality of the captured data is sensible to the surface's reflectance properties. Triangulation sensors yields inaccurate data in presence of shiny materials, with low surface albedo or having significant subsurface scattering.

Time-of-flight. Such sensing devices measures the time that the laser-beam takes to travel to the target and to return back to the sensor. By exploiting the velocity of the laser beam (speed of light) and the accurate measurement of the time taken, the distance between the object and the sensor can be computed. Laser range finders are the heart of the so-called time-of-flight 3D scanners, incorporating high-precision scanning abilities. These systems have been developed with near real time rates, being employed in large scale environments sensing missions (e. g. 100 m). Time-of-flight systems require accurate time measurements, limiting therefore the accuracy of the measured depths.

A.2 The state-based formulation of SLAM

The probabilistic SLAM was first introduced in 1986 in IEEE Robotics and Automation Conference, when probabilistic techniques were just starting to get spread within the robotics and artificial intelligence frameworks. Since robot mapping techniques are subject to uncertainty and sensor noise, research works were directed toward probabilistic techniques. Solutions to the probabilistic SLAM problem search an appropriate representation

for observation and motion models to allow for a consistent and efficient computation of the prior and posterior distribution. Probabilistic SLAM solutions were developed as mathematical derivations of the recursive Bayes rule.

Estimation-theoretic methods for mapping and localization includes Peter Cheeseman, Jim Crowley, Hugh Durrant-Whyte and Raja Chatila, Olivier Faugeras together with many other useful contributions. Since then, several key contributions were reported: Smith and Cheeseman [Smith and Cheesman, 1987], Durrant-Whyte [Durrant-Whyte, 1988b] introducing statistical basis concerning the relationship between landmarks and the geometric uncertainty.

The standard state-space approach using additive Gaussian noise leads to the use of the Kalman filter for robot's localization. Mapping allows an extension the this framework by estimating the landmarks' locations, beside the robot's pose. Closed loops, i.e. a second encounter of a previously visited area of the environment, play an important role for error bounding, by deforming the already mapped area such that a topologically consistent model is created. We resume hereafter some of the main filters in SLAM.

Several research works attempting to provide filters for SLAM lead to Kalman filter and it's variants: Extended Kalman Filter [Dissanayake et al., 2001] (EKF), Information Filtering [Thrun et al., 2000] and it's related Extended Information Filtering(EIF) Thrun et al. [2002]. EKF includes the non-linearities from outside world by approximating the robot motion using linear functions. When using EKF-SLAM techniques, the observation update step requires that all the landmarks and the joint covariance matrix be updated every time an observation is made, yielding a computational time which grows quadratically with the number of landmarks. This problem is mainly due to the fact that each new landmark is correlated to all other ones, being a fundamental need for the long-term convergence of the algorithm. In Guivant and Nebot [2001] authors introduced the Compressed Extended Kalman Filter (CEKF) which reduces the computational requirements of EKF, without affecting the accuracy of the results.

A second class of filters aiming to solve for the SLAM problem belongs to the Particle Filter (PF) approach, which is actually a recursive Bayesian filter implemented as Monte Carlo simulations (SMC). The use of particles allows to handle high-nonlinearities of sensors and non-Gaussian noise. This capability causes a growth in the computational complexity on the state dimension as new landmarks are being measured, which makes it unsuitable for real-time applications. As a consequence to this drawback, SLAM frameworks employ PF for localization purposes in combination with other SLAM techniques, such as FastSLAM. The FastSLAM was introduced by Montemerlo [Montemerlo et al., 2002] which shifted the designed of the recursive probabilistic SLAM. In contrast to EKF which commits a single data association for the entire filter, FastSLAM performs a local data association for each particle. The PF is used to sample over robots paths, requiring less memory usage and computational time than EKF or KF.

Expectation maximization (EM) methods iterate two steps: an expectation step (E-step) when the posteriors over the robot's poses are calculated, and the maximization step (M-step), in which the map is calculated given the poses' expectations. A main advantage over the KF methods is that EM can handle the data association problem by localizing repeatedly the platform wrt the map in the E-step. The M-step takes into account the detected features which are either reinforced in the next E-step or eliminated. However, EM methods need to process several times the same data, becoming inefficient for real-time application. When estimating the robot's pose, the computational cost grows exponentially with the size of the map and the error is unbounded, leading to unstable maps after long

cycles. A possible solution to this issue is to solve for the data association problem, which corresponds to the E-step elimination. As a consequence, authors combine EM - for mapping (M-step), with a PF-localizer to refine odometry readings.

A.3 Existing 3D Modeling Systems

The research work presented in this dissertation is concerned with the photometric and geometric recovery for generating in-situ photorealist and complete 3D digital models in complex and large scale underground environments. Thus, this section presents several representative 3D modeling systems and points out their limits with respect to the aforementioned research goal. Various 3D modeling systems have been developed promoting a wide range of applications:

- cultural heritage of large-scale objects and monuments: Stanford's University Michelangelo Project [Levoy et al., 2000], [Bernardini and Rushmeier, 2002], Great Buddha Project [Ikeuchi et al., 2007], Bayon Temple [Banno et al., 2008], IBM's Pieta Project [Wasserman, 2003], Columbia University's French Cathedral Project [Allen et al., 2003b], detailed 3D modeling of castles [El-Hakim et al., 2007] and digital recording of aboriginal rock art [El-Hakim et al., 2004].

- 3D modeling of urban scenes [Peter K. Allen and Blaer, 2001], [Stamos et al., 2008], [Stamos and Leordeanu, 2003], [Reed and Allen, 2000].

- modeling from real world scenes [Zhao et al., 2005], [Dias et al., 2003], [Huber, 2002], [VIT, 2000].

- natural terrain mapping [Huber and Herbert].

- underground mine mapping [Surmann et al., 2003], [Thrun et al., 2003], [Magnusson and Duckett, 2005], [Magnusson and Duckett, 2007].

- autonomous planetary exploration using both: vision [Giralt and Boissier, 1992], [Goldberg et al., 2002], [Matthies and Shafer, 1987], [Mathies et al., 2007] and range measurements [Rekleitis et al., 2009], [Hebert et al., 1989].

Through the following description, we review a non-exhaustive list of systems closely related to our research goal, i.e. in-situ 3D modeling in large-scale unstructured environments. Beside generating in-situ complete and photorealist 3D models, the proposed 3D modeling system aims at providing the possibility to perform 3D measurements onto 3D digital models for scene understanding and data annotation purposes. To this end, accurate geometry recovery is one of our main concerns.

First, we analyze passive 3D modeling systems' performances, in particular, how far away we can push the capacities of IBM to gather accurate 3D geometry. Second, our attention moves to laser-based systems to emphasize their limitations for producing photorealist 3D models. Third, we briefly review the existing dual systems and study their capacities with respect to our research goal in order to motivate further the choice of our system's design and 3D modeling strategy.

A.3.1 Image-based Systems

Photosynth®. Szeliski's group at Microsoft Research has recently reported an interesting system for virtual tourism [Snavely et al., 2006] which relies on a combination of IBR and IBM along with image retrieval and annotation techniques for exploring photos collection in 3D. IBM, in particular a robust SFM approach is employed for recovering the viewpoint of each photograph, to estimate a sparse 3D model of the scene and to geo-register the latter onto global 3D maps (Digital Elevation Models-DEMs). IBR methods are used to enable the user to smoothly navigate between photographs. Authors demonstrated the reliability and the broad applicability of the proposed system on numerous sites, ranging from Notre Dame to the Great Wall of China and Yosemite National Park. The proposed system is interested in providing data annotation for virtual tourism purposes, paying the price of heavy manual processing for the geo-registration step. In addition, sparse and noisy 3D geometry obtained via SFM cannot allow 3D accurate measurements of the surface geometry, which is one of our main concerns is our research work.

City scanning. CSAIL group at MIT attacked the problem of photometric rendering in large scale urban scenes [MIT, 2000] using a large amount of pose-annotated images (via GPS measurements) for generating spherical mosaics. The poses (rotation and translation) between adjacent nodes (overlapping mosaics) is computed through the use of vanishing points matches and Hough transformation. While encoding complete scene representation of the system's surroundings through the mosaics' use, with an image-based approach it is difficult to model high detailed architectural environments. Furthermore, the final rendering does not allow accurate 3D measurements. Finally, as a consequence of using geometric features for pose computation, this method cannot be employed in unstructured environments.

Zisserman's group [Fitzgibbon and Zisserman, 1998] aims at developing a system for fully automatic construction of Graphical Models of scenes when the input is an image sequence. The authors employ feature extraction and matching through couples and triple consecutive images. The final result is a 3D point cloud which samples irregularly the real 3D scene geometry. While extending the limits of the image-based methods, this approach has two inherent drawbacks: sparse depth estimates and the use of Computer Aided Design (CAD) model which is a coarse approximation in the areas which does not support 3D measurements.

IBM for navigation. In [Johnson et al., 2007] reports the use of the Descent Image Motion Estimation Systems (DIMES) during the Mars Exploration Rover (MER) landings for estimating of the lander's velocity before touchdown. Currently, the latest mission on Mars [Mathies et al., 2007] used successfully stereovision to build 3D models which for both: automatic terrain assessment and visual odometry. In the case of automatic terrain assessment the triangulated 3D points are used to evaluate the traversability of the terrain immediately in the front of the robot, which is defined as a regular grid of squared patches. In the case of visual odometry, the model is used to identify and track features of the terrain to mitigate the effect of slip [Howard and E. W. Tunstel, 2006].

ESA's ExoMars Rover panoramic camera (PanCam). The ESA Aurora programme will deploy the ExoMars rover to deliver the exobiology payload *Pasteur* to the surface of Mars by 2016. A 0.7 kg panoramic camera (PanCam [Griffiths et al., 2006]) was designed to fulfil the digital terrain mapping requirements of the mission as well as to provide multi-spectral geologic stereo panoramic views. The PanCam can also be used for high resolution imaging of inaccessible locations on crater walls and to observe retrieved

subsurface samples before ingestion into the Pasteur payload. A recent paper [Paar et al., 2009] describes the current status and implementation of the PanCam vision ground processing workflow. Authors are introducing the *PROX* software, a 3D vision processing pipeline capable to support several key functionalities such as panorama mosaicing as well as generation of textured triangular meshes and DEMs from stereo images. Since time is a limiting factor for the success of such endeavours, authors' main concern is to develop an automatic and fast processing pipeline in order to overcome the main drawbacks of several robotics surface missions, mainly on Mars, which are currently operated with tremendous manual processing. However, the proposed *PROX* 3D vision toolbox aims to deliver a fully automatic processing pipeline to be performed by the host on Earth which consequently does not allow for in-situ data processing and interpretation. This keeps the rover-host dependency issue still open, since during the mission the system must be able to deal with memory bandwidth, communication latency or real time decision making to handle unpredictable situations. The only reliable and efficient solution to this problem is to develop a automatic processing pipeline to be perform in-situ.

IBM systems provide passive 3D vision and color information with light and low-cost sensors, rapid acquisition and processing. On the other hand, 3D laser range finders (LRFs) capture range data to build terrain with centimeter accuracy. Although very important in our research context, such accuracy would be very difficult to attain with most stereo vision systems when dealing with unstructured environments, being very difficult to detect stable features. Furthermore, LRFs sensors provide 3D point clouds without needing processing, which is very important for site inspection purposes (3D measurements). Finally, since LRFs do not rely on ambient lightening, there is no need to address the problems arising from adverse lighting conditions.

A.3.2 3D Laser-based Systems

3D Mapping. A popular testbed for 3D mine mapping tasks is the Groundhog robot illustrated in Figure 2.1f). The first framework embedded onboard Groundhog consists of a volumetric mine mapping system reported in [Thrun et al., 2003]. The platform is equipped with four 2D laser range finders, without odometry sensor. Authors employ an ICP-based method for 2D scan registration to recover relative pose estimates, yielding locally consistent maps. In order to build globally consistent maps, the local poses initialize a slightly modified version of ICP to find the correspondences between robot's poses at different points in time. Volumetric maps are obtained by integrating the resulted maps and poses with 3D information acquired by additional scanners pointing toward the ceiling and the floor of the mine. Unfortunately, the proposed reconstruction method is only valid for planar environments. Later, Groundhog was upgraded with odometry measurement devices and in [Nuchter et al., 2004] and [Baker et al., 2004a] authors reported two mine mapping frameworks which integrate odometry readings to provide initial pose estimate to ICP-like methods for scans' registration.

In [Surmann et al., 2003] authors developed an autonomous mobile robot equipped with a 3D laser range finder for 3D exploration and digitalization. In [Nuchter et al., 2005] authors introduced the Kurt3D robot for 3D mapping of rescuing environments using a client-server architecture wirelessly connected. The 3D laser range finder is build by using a SICK 2D range finder and a standard servo motor to capture pitch motion. Due to the erroneous vehicle's poses, authors are considering the geometric structure of the overlapping 3D scans for registration. The registration process chooses extremum of

scans as natural landmarks, which correlate to corners and jump edges. The host preforms the robot's tele-operation for interactive 3D mapping (i.e. manual correction of the initial 6DOF pose and restart of matching algorithm). The operator is charged to minimize the number of the acquired 3D scans to save time for victim detection. Since time is a major concern, authors attempt to speedup the ICP algorithm by reducing the 3D data and using a kd-tree to partition the space in order to speed up the matching process.

Later in [Cole and Newman, 2006] authors solved for 3D SLAM problem in outdoor environments by introducing a modified version of ICP with a wider convergence basin, overcoming the risk to fail in undesired minimas, due to erroneous odometry readings. In [Cole et al., 2005] a similar pose refinement algorithm for matching natural salient features. For rapidity purposes, the method selects points randomly to verifies their saliency. Therefore, in presence of homogeneous surface, this approach may lead to erroneous poses estimates.

When dealing with 3D mapping in large scales environments, an open issue is that the central loop of ICP requires to store the complete point cloud data, which is computationally unaffordable. Biber attacked this problem using the normal distribution transform (NDT) for 2D scan registration [Biber and Strasser, 2003]. The key aspect to this representation stands in the model parametrization which consists in a combination of normal distributions, encoding the probability of finding a surface point at a certain position. Later, an extension to the 3D case (3D-NDT) is introduced [Magnusson and Duckett, 2007] for 3D mine mapping purposes. Authors present a quality assessment of the proposed method wrt the standard ICP algorithm showing that 3D-NDT is less error prone when using very low sample ratios and faster thanks to a more efficient scan surface representation. The proposed method exploits initial poses' estimates from the robot's two dimensional odometry which is subject to high errors when driving on undulated terrain. Since the odometry poses are too far from the true ones causing algorithm's failure, the framework's success is highly subject to the algorithm's parameters. Both algorithms were evaluated wrt the sensitivity to the initial error and it was noticed that 3D-NDT starts to fail for smaller values in the initial error than ICP.

In [Borrmann et al., 2008] authors addressed several open issues related to 3D mapping and 6DOF localization by extending a 2D SLAM method presented in [Lu and Milios, 1997]. Authors solved for several key issues which are inherent when dealing with 3D SLAM problem, such as: additional complexity due to the 6DOF, leading to a solution space which increases exponentially, non-linearities and the massive amount of data. Authors overcome the aforementioned issues using a Taylor expansion and Cholesky decomposition within a globally consistent scan matching algorithm initialized from odometry readings.

The aforementioned systems have several common drawbacks starting with the use of the vehicle's pose (GPS, IMU, odometry) which initializes feature-based ICP-like technique for pose refinement. Such frameworks are not reliable since radiometric and/or geometric features' existence cannot be guarantee in unstructured environments. These methods cause two major shortcomings to current the 3D scans' alignment scheme due to the matching step (i.e. the initialization phase which usually exploits information from vehicle's pose) and to the registration step usually performed by ICP. Usually, gross errors in the initial pose estimates causes ICP to get stuck in local minimas. Chapter 4 of this dissertation proposes an automatic pyramidal 3D scan alignment methods which replaces the two processing steps performed by the aforementioned scheme, overcoming therefore a major drawback of the existent scan alignment algorithms. The proposed technique does not require initial estimation and nor feature extraction and matching.

3D Terrain Modeling. In [Huber and Herbert] authors have wisely combined terrestrial with low altitude laser data acquired with an Unmanned Aerial Vehicle(UAV) for 3D unstructured terrain modeling. Authors have completely overcame the major bottleneck of the ICP method by eliminating the need of the initial alignment through the use of a surface matching algorithm based on shape descriptors [Johnson, 1997], [Huber, 2002]. Authors demonstrated the feasibility of a 3D mine mapping system [Huber and Vandapel, 2003b] using the same 3D scans alignment approach, emphasizing therefore the generality of the proposed system. On the downside, the shape descriptors rely on normal estimates, whose accuracy is subject to the scans' density.

Path planning. In [Rekleitis et al., 2009] authors introduce a path planning technique for planetary exploration using a Lidar sensor which acquires directly high resolution 3D mosaics for a given spatial position of the system. While assuring the completeness through the use of mosaics, high resolution scans causes both, data redundancy and unaffordable processing time for nearby mosaics matching. In this dissertation we propose a 3D mosaicing system by matching automatically several low-resolution scans. This allows to acquire both, complete and partial mosaics to eliminate occlusions which occur very often in complex environments, avoiding the acquisition of the already-acquired data. In addition, mosaics enable completeness, whilst low resolution scans allow for fast in-situ processing.

A.3.3 Dual Systems

Image-laser complementarity for 3D modeling has been extensively emphasized by several notable papers [Dias et al., 2003], [Levoy et al., 2000], [Ikeuchi et al., 2007], [Peter K. Allen and Blaer, 2001], [Zhao et al., 2005] leading to a predominant use of active 3D vision for capturing 3D surface geometry and color images for photorealistic 3D model rendering.

The main key issue now is how the image-laser pose is computed. In addition, it is still unclear how the texture mapping step is performed when the camera and the laser have different optical centers (refereing to 3D modeling pipeline described in section 2.3.3, FMCL systems). Researchers have extensively attempt to address the two aforementioned open issues to develop dual systems designed for automatic and photorealist 3D modeling. We provide a brief description of several existent systems having the above mentioned research goal.

Various 3D modeling systems were developed for cultural heritage purposes: Great Buddha Project [Ikeuchi et al., 2007] and Digital Michelangelo Project [Levoy et al., 2000] proposing an interactive 3D modeling pipeline of large-scale objects. A common drawback of the proposed frameworks is that they require heavy operator's intervention for both steps: data acquisition and processing. In [Ikeuchi et al., 2007] two methods are proposed to refine the image-laser alignment: if the two sensors are separated by a short baseline, calibration would be a suitable solution. Otherwise, a reflectance edge-based method is proposed. However, in presence of homogeneous surfaces and for high-resolution images, the edge detection may fail, leading to erroneous pose estimates.

In [Dias et al., 2003] authors added 3D points from stereo into the range data giving the possibility to increase 3D information and fill holes in the model. The image-laser registration process rely on semi-automatic initial alignment between laser-image and initial camera calibration [Tsai, 1987] based on edges, which is further refined by matching triangulated 3D points against the range data.

The VIT group [VIT, 2000], [Beraldin and Cournoyer, 1997], [El-Hakim et al., 1997] developed a mobile platform equipped with a range capturing device and nine cameras. The relative positions of the range sensor wrt each cameras are known from an off-line calibration procedure. This is a limiting factor, since the off-line calibration is required regularly. An additional high resolution camera provides texture information. The pose between the high resolution camera and the range sensors requires operator's intervention for finding corresponding points.

For large-scale man-made environments, such as urban areas, it is possible to create 3D models by combining a simple set of primitives such as rectangular blocks, pyramids and cones. Façade system reported by Devebec in [Devebec et al., 1996] is a good example of a primitive-based method. The proposed 3D modeling process starts with a manual initialization followed by a primitive fitting procedure via disparity minimization between 2D edges and 3D model edges projected in the 2D image space. Recent research work reported in [Werner and Zisserman, 2002] has eliminated the Façade system's need of for an initial guess. The main drawback of primitive-based approaches is that they impose orthogonality constraints over the scene's type, limiting their applicability to structured environments.

In [Zhao and Shibasaki, 1999] a system embedding a laser scanner and a camera sensor rigidly attached (RACL) is used for providing panoramic textured range images. Authors exploit the assumption that neighbor image are horizontally in order to register long range images sequences.

In [Sequeira et al., 1999] authors report a calibrated RACL system embedding a 3D modeling pipeline which relies on feature extraction and matching for pose estimation, limiting the applicability of the system to structured environments.

In AVENUE [Peter K. Allen and Blaer, 2001], [Stamos, 2001] the authors promote a project integrating range and intensity sensing for photo-realistic 3D modeling in urban environments. Aiming to minimize the amount of the user interaction, authors employ Façade's idea [Devebec et al., 1996] (i.e. line correspondences between the model and the image) for camera pose estimation and Backer's method [Becker and Bove, 1995], [Becker, 1997] for camera self-calibration consisting in exploiting parallel and orthogonality constraints. However, since such hypothesis cannot be applied to general 3D scenes, authors are studying solutions for extending the system to a general case study. In addition, since the image and the laser have different optical centers (FMCL system), the feature matching and texture mapping steps are highly sensible to occlusions in either image or laser data.

In [Zhao et al., 2005] the authors designed a 3D modeling system composed by a video camera and a Lidar for building aerial 3D models in semi-urban large-scale environments. Authors took advantage of the airborne-sensing context which leads to a negligible camera-Lidar baseline wrt the scene's depth, overcoming therefore the image-laser alignment shortcomings due to occlusions. A novel 3D "modeling-through-registration" technique is introduced using stereo from video and active 3D vision. In order to perform texture mapping coming from video, authors register passive 3D vision coming from stereo onto 3D point clouds acquired by the lidar. This method takes care of two exigencies: accurate geometry provided by Lidar data and photorealism through the use of texture from video data. On the downside, the framework relies on Harris corners [Harris and Stephens, 1998] extraction and matching for the stereo process and exploits initial alignment from DGPS and ground points surveys to register the inferred 3D points onto the Lidar data. This makes this method unsuitable for feature-less and GPS-denied areas, which is one of our major concern in this dissertation.

A coarse modeling of complex environments to be used for communication between users and mobile robots is reported in [Miura and Ikeda, 2009]. The system is equipped with two pan-tilt-zoom (PTZ) high resolution cameras and a range sensor. The proposed framework extracts texture from images and maps it onto plane segments generated from range data. The texture mapping step is the most expensive process of the entire 3D modeling pipeline. Authors performed on-line tests using two wireless-connected PCs which yield latency due to the slow wireless communication.

Beside the image-laser data alignment problem, another open issue of dual FMCL systems is the rendering of the registered range and intensity data. This problem was first discussed in [Chen and Williams, 1993] and attacked in [McMillan and Bishop, 1995], [Rademacher and Bishop, 1998] and [Shade et al., 1998]. Coorg in his PhD thesis [Coorg, 1998] introduces a solution based on median extraction technique under the assumption that the "correct" texture is visible from most images. Pulli in [Pulli et al., 1997] and Devebec in [Devebec et al., 1996] provide efficient solutions for view-dependent texturing, without handling the problem of texture occlusion. That means that parts of the scene which do not correspond to the modeled object but appear on the input images are erroneously texture-mapped onto the model.

The above mentioned methods are not suitable for generating in-situ complete and photorealist 3D scene models in previously unknown environments due to several reasons. First, they rely on manual data alignment provided either by an operator or by navigation sensors. In addition, image-laser alignment methods suppose the existence of either radiometric or geometric features which cannot be guaranteed in previously unknown environments.

Generally, the existent dual 3D modeling systems may be classified wrt two criterions:

- following *the application type*, we can find either semi-automatic frameworks designed for cultural heritage of large-scale objects or monuments, and automatic modeling processing for large-scale urban-like environments. The latter category imposes scene orthogonality constraints and relies on the existence of radiometric and geometric features, limiting their application to structured environments.

- following *the sensors disposition*, there are RACL and FMCL dual systems. The first category lead to the development of reliably and efficient 3D modeling systems capable of automatic functioning in structured environments. The latter category lead to manual, semi-automatic and automatic frameworks resulting in inaccurate 3D scene models due to occluded areas in either image or laser data.

Appendix B

Complements to Chapter 4

B.1 Depth Mode

In complement to Section 4.5.3, we describe below the depth mode of the rotation estimation procedure. If the intensity is not provided by the capturing device, the proposed scan matcher automatically switches to the *depth mode* of the rotation estimation procedure.

Analogue to the *intensity mode*, for each quaternion $\dot{\mathbf{q}}(\psi, \mathbf{n}[\theta_{\mathbf{n}}, \varphi_E]), (\psi, \theta_{\mathbf{n}}) \in \mathcal{P}_{SA}$, we map pixels $\mathbf{m}_2^j$ from D_2 in the D_1's space using the spherical projection expressed in Equation (4.8).

The optimal rotation is obtained by minimizing the difference in depth between the two panoramic images D_1 and D_2 in the overlapping region, expressed as follows:

$$\mathbf{E}^D(\psi, \mathbf{n}) = \frac{1}{D_{max} N_{12}} \sum_{j=0}^{j=N_2-1} \Phi_j^D |D_2(\mathbf{m}^j) - D_1(\hat{\mathbf{m}}^j_{\psi,\mathbf{n}})| \tag{B.1}$$

In order to obtain a depth dissimilarity score defined on the interval $[0, 1]$, the mean of absolute differences is divided to D_{max}, which denotes the maximal depth value computed over the entire scan sequence defined as:

$$D_{max} = max\{d_{max}(S_i), i = 0, .., N_{sequence} - 1\} \tag{B.2}$$

Φ_k^D denotes the characteristic function which penalizes "lost" and "zero" pixels. Φ_k^D is computed on depth images by applying Equation (4.10) to depth images D_1 and D_2 instead of I_1 and I_2, respectively.

The optimal rotation is given by the minimal dissimilarity measure, thereby maximizing the overlap $\hat{\mathbf{O}}^D[\hat{\psi}, \hat{\mathbf{n}}]$:

$$\hat{\mathbf{R}}^D[\hat{\psi}, \hat{\mathbf{n}}] = \arg \min_{(\psi,\mathbf{n}) \in \mathcal{P}_{SA}} \mathbf{E}^D(\psi, \mathbf{n}) \tag{B.3}$$

B.2 Complement to Section 4.6

The overall processing flow of the pair-wise scan-matcher is presented in Figure B.1. Figure B.2 illustrates the pair-wise scan matching results on a data set acquired in Moulin de Languenay prehistoric cave.

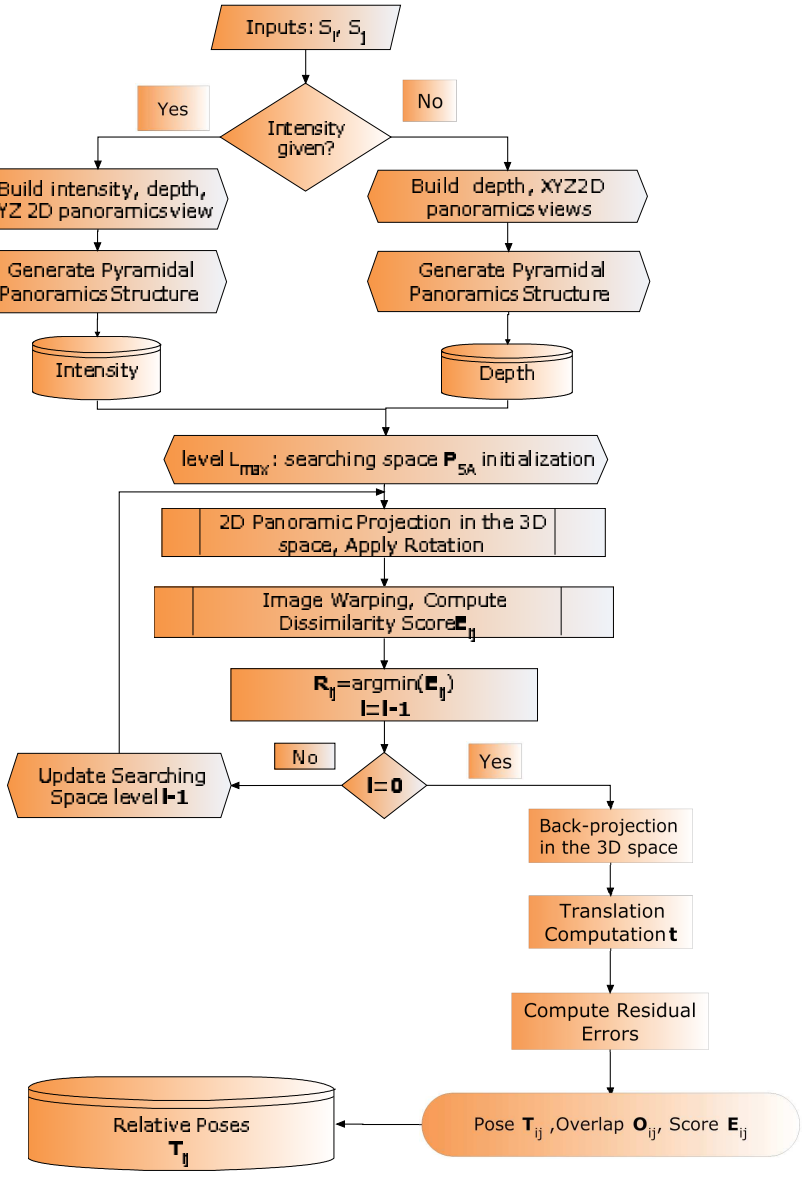


Figure B.1: The global pipeline of the pair-wise scan matcher.

B.3 Complement to Section 4.8

The global pipeline of the multi-view scans alignment process is shown in Figure B.3. The multi-view scan matching result obtained on the data set acquired in Moulin de Languenay prehistoric cave is illustrated in Figure B.4.

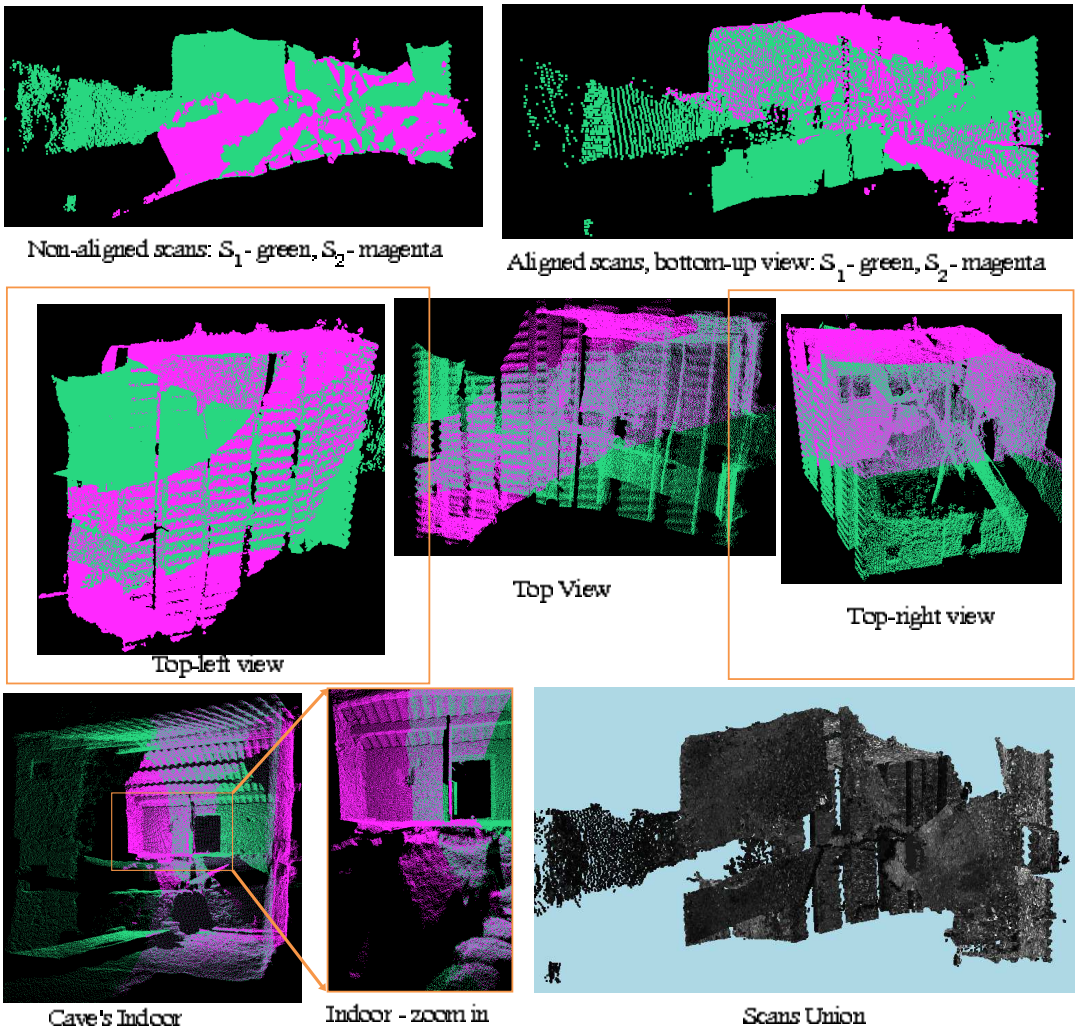


Figure B.2: Pair-wise scan matching results on data sets acquired in Moulin de Languenay prehistoric cave (France). Operating mode: intensity, total number of points: $7,57260 \times 10^5$ runtime: 4 min 25 s on a 1.66 GHz Linux machine.

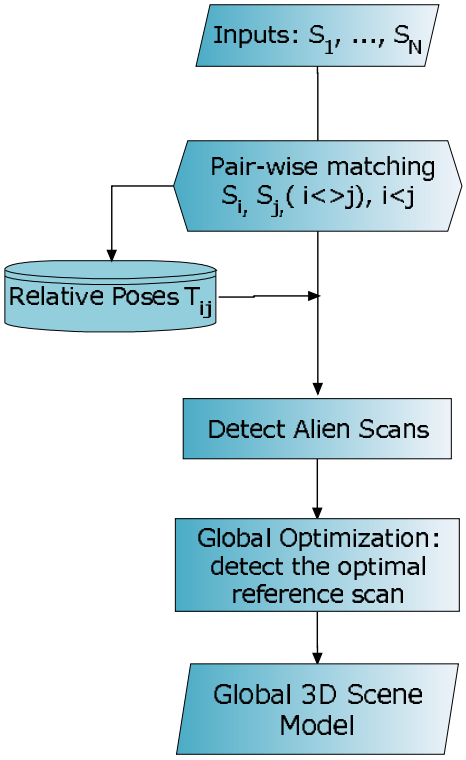


Figure B.3: The global pipeline of the multi-view scans alignment process.

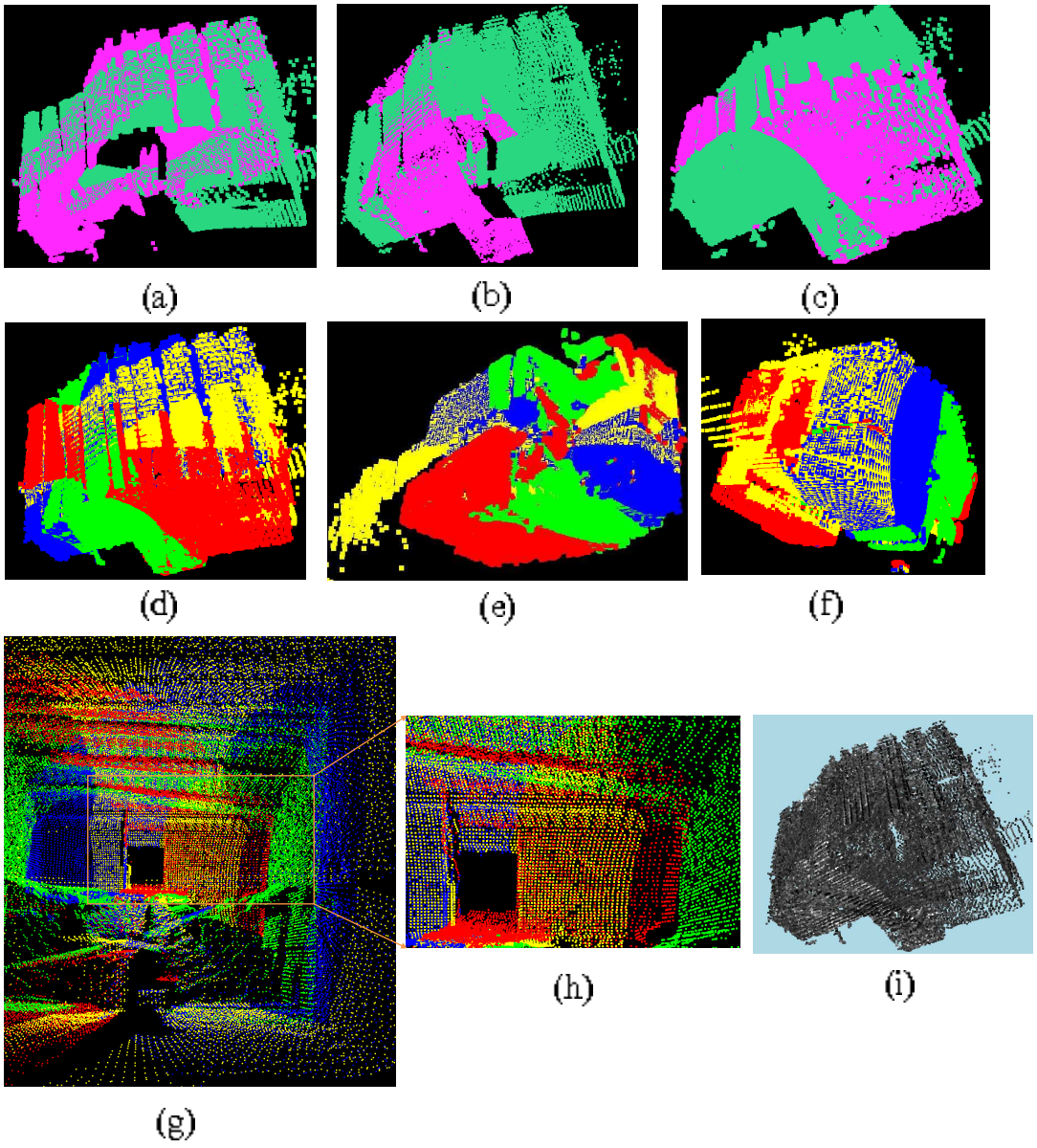


Figure B.4: Multiview Scan Matching results on data sets acquired in Moulin de Languenay prehistoric cave, France - Trial 1.(a) S_1 - green, S_2 - magenta, (b) S_{12} - green, S_3 - magenta, (c) S_{123} - green, S_4 - magenta, (d) Multiview scan alignment - Top-down view, S_1 - yellow, S_2 - blue, S_3 - green, S_4 - red, (e) Bottom-up view, (f) Front-left view, (g) Cave's indoor, (h) Cave's Indoor - zoom in, (i) Cave's outdoor rendering using the intensities acquired by the scanning device.

Appendix C

Complements to Chapter 5

C.1 Complement to Section 5.2.1

Table C.1 resumes the camera motions and their corresponding 2D image transformations.

Motion Models	Image 2D Transformations	3D Camera Motions
Euclidian	$\mathbf{T_{2D}} = \begin{bmatrix} \cos\theta & -\sin\theta & t_x \\ \sin\theta & \cos\theta & t_y \end{bmatrix}$	$\mathbf{T_{3D}} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix}$
D.O.F.	$\{\theta, t_x, t_y\}$	$\{\theta, \varphi, \psi, t_x, t_y, t_z\}$
Similarity	$\mathbf{T_{2D}} = \begin{bmatrix} s\cos\theta & -s\sin\theta & t_x \\ s\sin\theta & s\cos\theta & t_y \end{bmatrix}$	$\mathbf{T_{3D}} = \begin{bmatrix} sr_{11} & sr_{12} & sr_{13} & t_x \\ sr_{21} & sr_{22} & sr_{23} & t_y \\ sr_{31} & sr_{32} & sr_{33} & t_z \end{bmatrix}$
D.O.F.	$\{s, \theta, t_x, t_y\}$	$\{s, \theta, \varphi, \psi, t_x, t_y, t_z\}$
Affine	$\mathbf{T_{2D}} = \begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \end{bmatrix}$	$\mathbf{T_{3D}} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & t_x \\ a_{21} & a_{22} & a_{23} & t_y \\ a_{31} & a_{32} & a_{33} & t_z \end{bmatrix}$
D.O.F.	$\{a_{ij}, t_x, t_y \vert i, j = 1, 2\}$	$\{a_{ij}, t_x, t_y, t_z \vert i, j = 1, 2, 3\}$
Projective	$\mathbf{T_{2D}} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$	$\mathbf{T_{3D}} = \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14} \\ h_{21} & h_{22} & h_{23} & h_{24} \\ h_{31} & h_{32} & h_{33} & h_{34} \\ h_{41} & h_{42} & h_{43} & h_{44} \end{bmatrix}$
D.O.F.	$\{h_{ij} \vert i, j = 1, 2, 3\} \setminus \{h_{33}\}$	$\{h_{ij} \vert i, j = 1, .., 4\} \setminus \{h_{44}\}$

Table C.1: Parametric camera motions (3D) and image transformations (2D).

C.2 Perspective Geometry and Camera Calibration

Without loss of generality, let us now drop the subscript referring to each camera coordinate frame and consider the case of the imaging process for one camera with its coordinate frame centered in O_C, as shown in Figure C.1.

The image alignment process requires to model the image formation process in order to find out at which pixel location $\mathbf{u}$ gets mapped a 3D point $\mathbf{p}$. The first thing to do is to trace-back the 3D rigid transformation relating the world $\mathcal{O}_{(x,y,z)}$ and the camera

coordinate frame O_C in order to recover the 3D coordinates of $\mathbf{p}$ expressed in the camera coordinate frame, noted $\mathbf{p}_C$, using Equation C.1.

$$\mathbf{p}_C = \left[\mathbf{R}|\mathbf{t}\right] \mathbf{p} \tag{C.1}$$

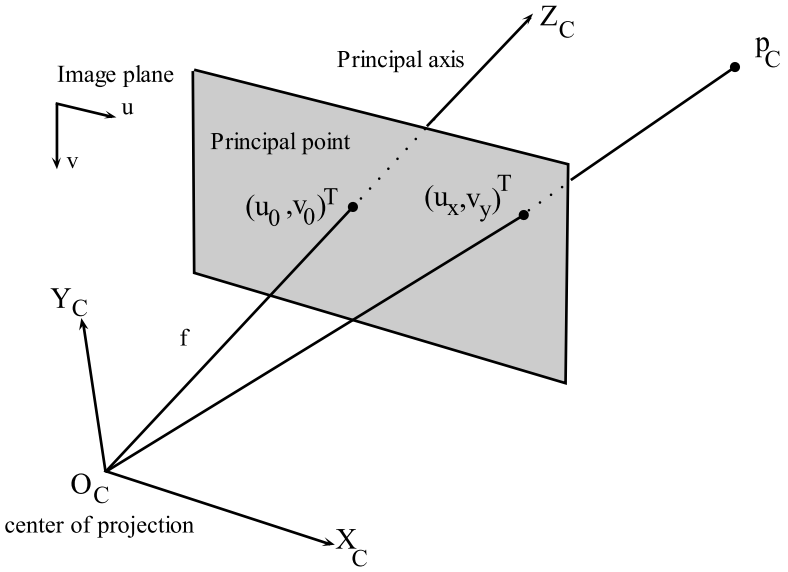


Figure C.1: The pinhole camera anatomy.

Under the pinhole camera model, a point in space $\mathbf{p}_C = (p_x, p_y, p_z)^T_C$ gets mapped at a 2D pixel location $\mathbf{u} = (u_x, u_y)^T$ through the central projection process. The perspective projection scales the coordinates by depth to produce image coordinates which are further converted to pixel values using the camera focal length f expressed in pixels and the coordinates of the principal point $(u_0, v_0)^T$. The mapping process takes the following form:

$$(p_x, p_y, p_z)^T_C \mapsto (f\frac{p^C_x}{p^C_z} + u_0, f\frac{p^C_x}{p^C_z} + v_0)^T \tag{C.2}$$

The entire perspective projection can be parameterized up to a scale factor and expressed as a chain of matrix transformations using the following expression:

$$\begin{bmatrix} u_x \\ u_y \\ 1 \end{bmatrix} \cong \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} p_x \\ p_y \\ p_z \\ 1 \end{bmatrix}_C \tag{C.3}$$

and by noting with $\mathbf{K}_{3\times3} = \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix}$ the camera calibration matrix and with $\mathbf{P}_{3\times4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$ the canonical perspective projection matrix, we can write in a concise form the entire perspective projection relating a world 3D point $\mathbf{p}$ and an image point $\mathbf{u}$, as expressed in the equation hereafter:

$$\mathbf{u} \cong \mathbf{K}\mathbf{P}\mathbf{p}_C = \mathbf{K}\mathbf{P}[\mathbf{R}|\mathbf{t}]\mathbf{p} = \mathbf{M}\mathbf{p} \tag{C.4}$$

where $\mathbf{M}$ denotes the 3×4 perspective projection which encodes the imaging process of a world point $\mathbf{p}$ under the pinhole camera model.

CCD cameras. Equation (C.3) assumes that pinhole camera model scales equally in both axial directions leading to a 9-parameters camera model (6 D.O.F. for the external parameters, and 3 intrinsic camera parameters, f and the principal point location). When it comes to CCD cameras, a more general camera matrix models non-square pixels and the skew factor [Hartley and Zisserman, 2004], leading to a 10-parameter camera model. However, [Szeliski, 2006] for image stitching purposes it was shown that the focal length and the variable optic center location yields high-quality results.

Nonlinear distortion. Up to now the imaging process was assumed to be linear and accurate, meaning that the world point $\mathbf{p}$, the image point $\mathbf{u}$ and the optical center O_C are collinear. In a real case study, all imaging devices introduce an amount of nonlinear distortion and non-pinhole lenses must be modeled in order to take into account deviations from the ideal pinhole camera model. As stated in [Lenz and Tsai, 1988], for accurate 3D-measurements it is very important to model lens distortions and correct them.

Under such nonlinear distortions, the actually observed image point is $\mathbf{u}^d$. As stated in [Tsai, 1987], [Devernay and Faugeras, 2001], the most severe deviation from the pinhole camera model is the radial distortion, which is performed along the radial direction from the center of distortion, increasing as the focal length decreases. Such deviation displaces an image point from its ideal location $\mathbf{u}$ either inward (*pincushion* distortion) or outward (*barrel* distortion) the image center with an amount proportionally to their radial distance [Juyang Weng and Herniou, 1992]. There are cameras for which the distortion varies with color, being refereed to as a chromatic aberration.

A commonly used technique for the alleviation of the radial distortion is performed in three steps: (i) by first applying a parametric radial distortion model, (ii) by estimating the distortion parameters and finally (iii) by correcting the distortion.

Radial distortion parametrization. When taking into account such deviations, the goal is to correct image measurements to those that would have been produced under an ideal (non-distorted) pinhole projection. Since $\mathbf{u}$ denotes the image coordinate of a point produced by an perfect imaging process, we note $\mathbf{u}^d$ the distorted image point which is related to $\mathbf{u}$ by a radial displacement. Assuming that the center of the distortion is at the principal point, the relationship between the distorted and the undistorted radial distance r is given by:

$$r^d = r + \delta_r \tag{C.5}$$

where δ_r is the radial distortion. Early studies in photogrammetry [Slama, 1980] modeled the radial distortion using a Taylor expansion and the simplest radial distortion models use low-order polynomials:

$$r^d = rf(r) = r(1 + k_1 r^2 + k_2 r^4 + k_3 r^6 + ...) \tag{C.6}$$

which settles the coefficients of the radial correction to $\{k_1, k_2, k_3, ...\}$ which are usually considered as part of the interior camera calibration [Brown, 1971], [Slama, 1980]. In practice, the principal point is usually set as the center of the radial distortion, although is not always the case. Expressed in pixel coordinates, the correction is written as following:

$$\begin{cases} u_x^d - u_0 = (u_x - u_0) f(r) \\ u_y^d - v_0 = (u_y - v_0) f(r) \end{cases} \tag{C.7}$$

Computing the radial distortion function. It is difficult to compute analytically the inverse polynomial function, being generally computed numerically via iterative schemes. When estimating the radial distortion model in (C.6), the dominant parameters is k_1 and various studies have shown that when too higher order polynomial may cause numerical instability [Tsai, 1987], [Lenz and Tsai, 1994], [Zhang, 1999], [Lee, 2000]. A widely employed technique for computing specific radial distortion models exploits feature correspondences (corners [Zhang, 1999], circles [Heikkila, 2000]). The most commonly used due to its simplicity is the plumb line algorithm [Brown, 1971] which iteratively adjusts the radial distortion parameters until all of the lines present in the image are straight [Brown, 1971], [Kang, 2001], [El-Melegy and Farag, 2003]. The bind removal technique [Farid and Popescu, 2001] which relies on the fact that lens distortion introduce high-order correlations in the frequency domain. Another approach is to estimate the radial distortion parameters within the image alignment process [Sawhney and Kumar, 1999], [Stein, 1997]. More recent approaches solve simultaneously for intrinsic and radial distortion parameters using higher order temps or non-parametric forms [Claus and Fitzgibbon, 2005], [Sturm, 2005], [Tardif et al., 2006].

The camera calibration matrix and the radial distortion specifies the mapping of an image point to a ray in the camera coordinate system. Although radial distortion suffice for designing consumer-level image stitching algorithms, more accurate results can be obtained by modeling the *tangential* distortion can be modeled in order to increase the accuracy [Slama, 1980].

C.3 Basic rendering

The mosaic rendering process starts by choosing a compositing surface. The second main step is the mapping between the input images and the output pixel coordinates wrt the compositing surface.

Choosing the compositing surface. A natural approach gives rise to what is usually called a *flat panorama* consisting in choosing one image as the reference and warping all the other images into the chosen reference coordinate frame. This approach keeps the perspective projection attributes for wide-angle panoramas with a FOV inferior than 90°. Since for larger FOV panoramas this method stretches pixels near the image border, cylindrical [Szeliski, 1994], [Chen, 1995] or spherical projections [Szeliski and Shum, 1997] are used instead. Cube maps are also used for *environment mapping* purposes in Computer Graphics [Szeliski and Shum, 1997] and other methods for representing the globe were introduced by cartographers [Bugayevskiy and Snyder, 1995].

In [Shum and Szeliski, 2000] authors first proposed a texture mapping algorithm for spherical surfaces. Since these models do not employ a polyhedrone-shape representation, they do not exploit the hardware texture-mapping acceleration, requiring instead specialized viewer. Moreover, spherical representation results in distorted rendering at poles. In order to overcome these shortcomings, authors propose the use of environment maps [Greener, 1986] and allow to the user to choose the model shape, which can range from a cube to subdivided dodecahedron or even a latitude-longitude tessellated globe. In all cases, the choice is somehow hardware-dependent, being also strongly related to the desired quality (distortion minimization, etc.). Authors propose a texture mapping algorithm for any geometry and choice of texture map coordinates. In its general formulation, this algorithm allow to project a collection of images onto an arbitrary model, implying non-convex models - which do not surround the viewer.

The surface parametrization of the final image involves a tradeoff between keeping the local appearance undistorted under uniform sampling. Since the choice of the parametrization surface is strongly related to the application type, recent researches are directed toward methods capable to perform automatically the surface selection to allow smooth navigation between panoramic views.

Texture maps. The mapping between the inputs and the output pixels coordinates wrt the compositing surface requires a *coordinate transformation* process. In the flat panorama case, the coordinate transformation is a homography [Szeliski, 2006]. The warping is performed in graphics hardware by setting the texture mapping coordinates and rendering a single quadrilateral. In the cylindrical or spherical case, every pixel is converted into a viewing 3D ray and then back-projected into each image, including eventually the radial distortion. In other cases, the final compositing can take the form of a texture-mapped polyhedron. In this case, the 3D and texture map coordinates must be properly handled and ensure that the texture pixels interpolated during the 3D rendering have valid values [Szeliski and Shum, 1997].

Since the coordinate transformations can yield fractional pixel locations, special attention must be given to *sampling issues*. When the final compositing has a lower resolution than the input images, the input images must be passed through a pre-filtering step in order to avoid aliasing. Researches on this directions were reported in Computer Graphics community [Wiliam, 1983], [Greener, 1986], dealing with the problem of computing an appropriate pre-filter wrt the distance between the neighboring samples in a source image. High-visual quality can be obtained using a cubic interpolator used jointly with a spatially adaptive pre-filter [Wang et al., 2001]. Higher resolution than the input images can be obtained through a process called super-resolution [Szeliski, 2006].

C.4 Mosaicing "make-up"

When the images are perfectly registered and identically exposed, every combination of pixels will lead to correct rendering. However, this is not always the case for a real application, which includes casually acquired images using the automatic mode of the camera. In this situation, special attention must be given to exposure differences, blurring caused by mis-registration and ghosting caused by dynamic scenes. Consequently, one must decide which pixel are to be used and how to weight or blend them. Several spatially varying weighting (*feathering*), pixel selection and blending strategies can be found in [Szeliski, 2006].

Feathering. The simplest technique takes the average value at each pixel. However, this does not work well in presence of exposure difference, mis-registrations and scene movement. The latter issue is tackled in [Irani and Anandan, 1998] which proposed the use of median filter to eliminate moving objects.

Blending. When using feathering, it is difficult to balance between smoothing out low-frequency exposure variations and retaining sharp enough transitions to prevent blurring. Blending strategies for compensating moderate exposure differences include laplacian pyramid blending [Burt and Adelson, 1983] and gradient domain blending [Agarwala et al., 2004], [Perez et al., 2003]. In order to handle larger amounts of exposure differences between images, alternative approaches were reported in [Uyttendaele et al., 2001] which are capable of handling local variations in exposure due to lens vignetting effects.

Alternative methods were introduced for exposure compensation, such as *high dynamic range imaging* which consists in estimating a single *high dynamic range* (HDR) from differ-

ently exposed images [Mann and Picard, 1995], [Devebec and Malik, 1997], [Mitsunaga and Nayar, 1999], [Reinhard et al., 2005]. When casually acquired images are employed, it is difficult to design a blending method capable to avoid sharp transitions and deal with scene motion. This problem has recently been attacked in [Eden et al., 2006] by first finding a consensus mosaic and then selectively computing radiances in under- and over-exposed regions. Such a mosaic rendering framework is usually employed for casually acquired images, which may not be perfectly registered due to different exposures.

In [Shum and Szeliski, 2000] and [Brown and Lowe, 2007] the target is to provide an artistic mosaic rendering pipeline, easy to use which can eventually require user-input, whereas in [Garcias and Santos-Victor, 2000] a more industrial-oriented problem is attacked. The latter addresses the issue of automatic creation of video mosaics in underwater environments and vehicle-self localization relying on mosaics as visual maps. The rendering pipeline is focused on selecting a unique intensity value from multiple contributions in the overlapping regions in order to compose the output image. The proposed rendering technique operates in the time domain, being introduced as a *temporal operator*. The contributions of each image for the mosaic output are seen as lying on a line parallel with the time axis. Possible solutions include use-first, use-last, mean and median. As mentioned before, the average value has been proved to be effective for removing temporal noise inherent in video. The median operator removes both noise and transient data, such as moving objects whose intensity patterns are stationary for less than half of the frames. This is particularly useful for underwater sequences, when moving fish or algae may be captured.

C.5 Complement to Section 5.7.4.2

Figures C.2 and C.3 present the experimental results of the pair-wise motion estimation process obtained on a second trial performed in the 12^{th} district of Paris.

C.6 Proposed Closed-from solution for Optimal Unit Quaternion Computation

When solving for the optimal quaternion which minimizes the cross product between a set of N homologous 3D vectors, the criterion to minimize is expressed under the form:

$$\hat{\mathbf{q}} = \arg\min_{\mathbf{q}} Q_{12-\times}(\mathbf{R}(\mathbf{q})) = \arg\min_{\mathbf{q}} \sum_{k=0}^{N-1} \|\mathbf{v}_1^k \times \mathbf{R}(\mathbf{q})\mathbf{v}_2^k\|^2 \tag{C.8}$$

where $\mathbf{R}(\mathbf{q})$ is given in Equation (4.4). The minimum of Equation (C.8) is obtained for the quaternion $\hat{\mathbf{q}}$ which annuls its first derivative wrt the its 4-parameters, noted $\frac{\partial Q_\times}{\partial \mathbf{q}}$ in which we dropped the subscript 12 for the sake of clarity. By posing $r = \|\mathbf{v}_1^k \times \mathbf{R}\mathbf{v}_2^k\|$, the partial derivative can be written under the following form:

$$\frac{\partial Q_\times}{\partial \mathbf{q}} = 2r\frac{\partial r}{\partial \mathbf{q}} \tag{C.9}$$

and by developing the partial derivatives $\frac{\partial r}{\partial \mathbf{q}}$ we obtain:

$$\begin{cases} \frac{\partial r}{\partial \mathbf{q}} = \|\frac{\partial \mathbf{v}_1^k}{\partial \mathbf{q}} \times \mathbf{R}\mathbf{v}_2^k + \mathbf{v}_1^k \times \frac{\partial(\mathbf{R}\mathbf{v}_2^k)}{\partial \mathbf{q}}\| \\ \frac{\partial \mathbf{v}_1^k}{\partial \mathbf{q}} = 0 \end{cases} \tag{C.10}$$

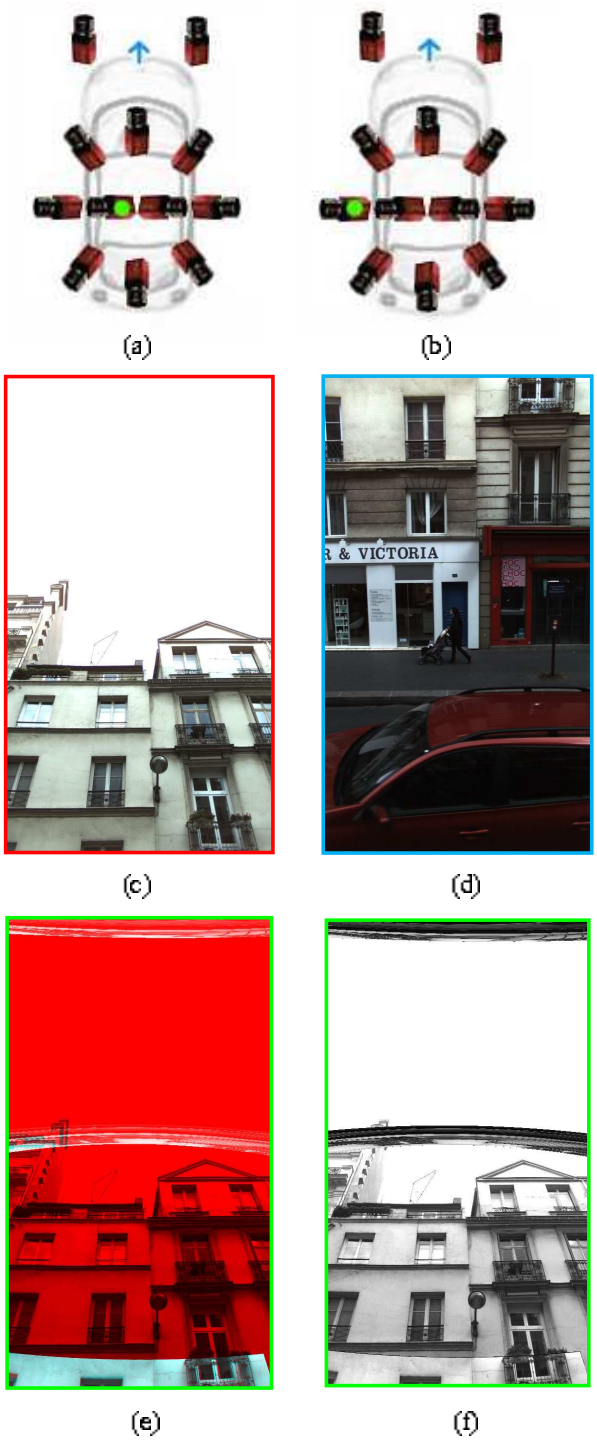


Figure C.2: Trial 2 outdoor structured environments - 12^{th} district of Paris. Global rotation estimation results. (a) camera no. 32, (b) camera no. 31, (c) I_1 - reference image, (d) I_2 - image to align, (e) global maximum localization at level $L_{max} = 3$, (f) rotationally aligned images at level $l = 0$: I_1-red channel, the warped image $I_2(\mathbf{u}; \hat{\mathbf{R}})$-green channel, $\hat{\mathbf{R}}_{(\theta,\varphi,\psi)} = (70.5306^\circ = -0.5462^\circ, 0.1556^\circ)$.

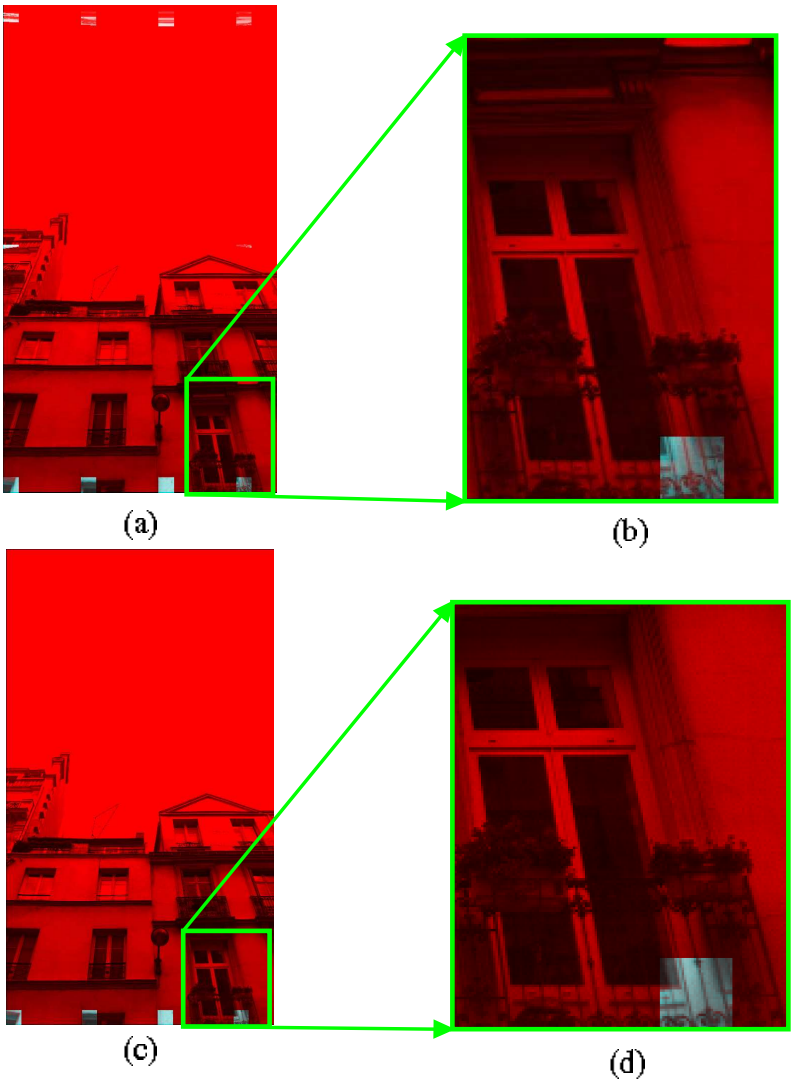


Figure C.3: Trial 2 outdoor structured environments - 12^{th} district of Paris. Non-rigid motion estimation via local patching. (a) Patch extraction in I_2 and warping in the I_1's image space using the global rotation $\hat{\mathbf{R}}$: I_1 - red channel, warped patches in I_2 using the estimated rotation $\mathcal{P}(\mathbf{u}_2^k)$ - green channel. (b) zoom-in of the warped patches of (a) emphasizing mis-registrations caused by parallax, (c) I_1 - red channel, locally matched patches in $\mathcal{P}(\hat{\mathbf{u}}_2^k)$ and warped in I_1's image space - green channel, (d) zoom-in of the locally matched patches from (c) exibiting a fairly good local patch matching.

From Equation C.10 we have:

$$\frac{\partial r}{\partial \mathbf{q}} = \|\mathbf{v}_1^k \times \frac{\partial(\mathbf{R}\mathbf{v}_2^k)}{\partial \mathbf{q}}\| \tag{C.11}$$

The partial derivatives of $\mathbf{R}(\mathbf{q})$ wrt the unit quaternion $\mathbf{q} = [q_0, q_x, q_y, q_z]^T$ form a $3\times4\times3$-tensor and by multiplying it with the vector $\mathbf{v}_2$ we obtain:

$$\frac{\partial \mathbf{R}}{\partial \mathbf{q}}\mathbf{v}_2 = \begin{bmatrix} a & d & -c & b \\ b & c & d & -a \\ c & -b & a & d \end{bmatrix} \tag{C.12}$$

with

$$\begin{cases} a = q_0 v_{2x} + q_z v_{2y} - q_y v_{2z} \\ b = -q_z v_{2x} + q_0 v_{2y} + q_x v_{2z} \\ c = q_y v_{2x} - q_x v_{2y} + q_0 v_{2z} \\ d = q_x v_{2x} + q_y v_{2y} + q_z v_{2z} \end{cases} \tag{C.13}$$

in which we dropped the superscript k for clarity. The derivative is zero if at least one of two conditions below are fulfilled:

$$2r = 2\|\mathbf{v}_1^k \times \mathbf{R}(\mathbf{q})\mathbf{v}_2^k\| = 0 \tag{C.14}$$

or,

$$\frac{\partial r}{\partial \mathbf{q}} = \|\mathbf{v}_1^k \times \frac{\partial(\mathbf{R}\mathbf{v}_2^k)}{\partial \mathbf{q}}\| = 0 \Leftrightarrow \mathbf{v}_1^k \times \frac{\partial(\mathbf{R}\mathbf{v}_2^k)}{\partial \mathbf{q}} = 0 \tag{C.15}$$

By developing the condition expressed in Equation C.15, we obtain a system of form $\mathbf{A}\mathbf{q} = \mathbf{0}$ where $\mathbf{A}_{12\times4}$ is the measures matrix defined by its terms a_{ij} obtained by developing the cross product $\mathbf{v}_1^k \times \frac{\partial(\mathbf{R}\mathbf{v}_2^k)}{\partial \mathbf{q}}$ and by expressing it in terms of $\mathbf{v}_1^k$ and $\mathbf{v}_2^k$.

We develop hereafter the cross product:

$$\mathbf{v}_1^k \times \frac{\partial(\mathbf{R}\mathbf{v}_2^k)}{\partial \mathbf{q}} = \begin{pmatrix} v_{1x}^k \\ v_{1y}^k \\ v_{1z}^k \end{pmatrix} \times \begin{bmatrix} a & d & -c & b \\ b & c & d & -a \\ c & -b & a & d \end{bmatrix} = \tag{C.16}$$

$$= \left[\begin{pmatrix} v_{1x}^k \\ v_{1y}^k \\ v_{1z}^k \end{pmatrix} \times \begin{pmatrix} a \\ b \\ c \end{pmatrix} \quad \begin{pmatrix} v_{1x}^k \\ v_{1y}^k \\ v_{1z}^k \end{pmatrix} \times \begin{pmatrix} d \\ c \\ -b \end{pmatrix} \quad \begin{pmatrix} v_{1x}^k \\ v_{1y}^k \\ v_{1z}^k \end{pmatrix} \times \begin{pmatrix} -c \\ d \\ a \end{pmatrix} \quad \begin{pmatrix} v_{1x}^k \\ v_{1y}^k \\ v_{1z}^k \end{pmatrix} \times \begin{pmatrix} b \\ -a \\ d \end{pmatrix} \right] \tag{C.17}$$

The above matrix multiplication leads to the matrix $\mathbf{D}_{3\times4}$ with each of its terms being expressed as a linear combination between the corresponding measures vectors $\vec{\mathbf{v}}_1^k$, $\vec{\mathbf{v}}_2^{\,k}$ and the quaternion, i.e.:

$$\mathbf{D} = \begin{bmatrix} d_{00} & d_{01} & d_{02} & d_{03} \\ d_{10} & d_{11} & d_{12} & d_{13} \\ d_{20} & d_{21} & d_{22} & d_{23} \end{bmatrix} \tag{C.18}$$

with $d_{ij} = \mathbf{f}(\mathbf{v}_1^k, \mathbf{v}_2^k)\mathbf{q} = f_0 q_0 + f_x q_x + f_y q_y + f_z q_z$, and $\mathbf{f} : \mathbb{R}^{2\times3} \rightarrow \mathbb{R}^4$. After arranging terms, we get the measure matrix $\mathbf{A}_{12\times4}$ of form:

$$\mathbf{A}_{12\times4} = \begin{bmatrix} d_{00}^0 & d_{00}^x & d_{00}^y & d_{00}^z \\ d_{01}^0 & d_{01}^x & d_{01}^y & d_{02}^z \\ d_{02}^0 & d_{02}^x & d_{02}^y & d_{02}^z \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ d_{22}^0 & d_{22}^x & d_{22}^y & d_{22}^z \\ d_{23}^0 & d_{23}^x & d_{23}^y & d_{23}^z \end{bmatrix} \tag{C.19}$$

After obtaining the measures matrix, the resolution follows the same technical solution employed for the case when the cost function was the sum of the squared residual errors for which a description is provided in Appendix C.8.

By replacing $\mathbf{A}_{12\times 4}$ with the matrix relation $[\mathbf{Q}(\mathbf{v}_1^k) - \mathbf{W}(\mathbf{v}_2^k)]$ from Appendix C.8, the criterion to minimize becomes:

$$\|\mathbf{v}_1^k \times \mathbf{R}(\mathbf{q})\mathbf{v}_2^k\|^2 = \mathbf{q}^T\mathbf{A}_k^T\mathbf{A}_k\mathbf{q} = \mathbf{q}^T\mathbf{U}_k\mathbf{q} \tag{C.20}$$

where $\mathbf{U}_k$ is a symmetrical, squared and 4-dimensional matrix.

Consequently, the optimal quaternion minimizing the angle between a set of N homologous 3D vectors relating two adjacent images, initially given in Equation (C.8) takes the following form:

$$\hat{\mathbf{q}} = \arg\min_{\mathbf{q}} \sum_{k=0}^{N-1} \mathbf{q}^T\mathbf{U}_k\mathbf{q} = \arg\min_{\mathbf{q}} \mathbf{q}^T(\sum_{k=0}^{N-1} \mathbf{U}_k)\mathbf{q} = \arg\min_{\mathbf{q}} \mathbf{q}^T\mathbf{V}\mathbf{q} \tag{C.21}$$

with $\mathbf{V} = \sum_{k=0}^{N-1} \mathbf{U}_k = \sum_{k=0}^{N-1} \mathbf{A}_k^T\mathbf{A}$ being a symmetric and positive definite matrix. By constraining the system to produce a unit quaternion, we obtain:

$$\hat{q} = \arg\min_{\mathbf{q}} Q_\times = \arg\min_{q}(q^T\mathbf{V}\mathbf{q} + \lambda(1 - \mathbf{q}^T\mathbf{q})) \tag{C.22}$$

which provides us with a similar solution as the one obtain in Appendix C.8, i.e.: the unit quaternion which minimize $Q_\times$ is given by the eigen vector of $\mathbf{V}$ corresponding to the smallest eigen value of $\mathbf{V}$, noted λ.

C.7 Generalization to the Multi-view Case

This section generalizes the optimal quaternion computation for n images composing a mosaic node by minimizing the sum of squared angular errors enclosed by the corresponding 3D vectors. The global registration computes the best set of unit quaternions $\mathbf{q}_{node} = \{\mathbf{q}_{ij} | w_{ij} = 1\}$ to be applied to each vector set $\mathbf{v}_j^k, j = 0, ..., n-1, i \neq j, k = 0, .., N_{ij}$ such that the sum of all squared angular errors expressed in Equation (C.23) is minimized.

$$Q_{mosaic-\times} = \sum_{i=0}^{i=n-1} \sum_{j=0, i\neq j, i<j}^{j=n-1} w_{ij}(\sum_{k=0}^{k=N_{ij}-1} \delta_{ij}^k \|\mathbf{v}_i^k \times \mathbf{R}(\mathbf{q}_{ij})\mathbf{v}_j^k\|^2) \tag{C.23}$$

w_{ij} and δ_{ij}^k denote the adjacency and the visibility weights defined in Equation (5.23) and (5.24) respectively. The number N_{ij} denote the total number of features found between I_i and I_j. Instead of having a set of two-corresponding features coming from two overlapped images, for the multi-view image alignment case we have a collection of N_{node} features, with $N_{node} = \sum_{i=0}^{i=n-1} \sum_{j=0, i\neq j, i<j}^{j=n-1} N_{ij}$. The multi-view criterion from Equation (C.23) can be expressed using the pair-wise cost function from Equation(C.8) under the following from:

$$Q_{mosaic-\times} = \sum_{i=0}^{i=n-1} \sum_{j=0, i\neq j, i<j}^{j=n-1} w_{ij} Q_{ij-\times} \tag{C.24}$$

and by introducing the result from Equation (C.21) we obtain:

$$Q_{mosaic-\times} = \sum_{i=0}^{i=n-1} \sum_{j=0,i\neq j,i<j}^{j=n-1} w_{ij}(\mathbf{q}_{ij}^T(\sum_{k=0}^{N} \delta_{ij}^k \mathbf{U}_k)\mathbf{q}_{ij}) = \sum_{i=0}^{i=n-1} \sum_{j=0,i\neq j,i<j}^{j=n-1} w_{ij}\mathbf{q}_{ij}^T \mathbf{V}_{ij}\mathbf{q}_{ij} \tag{C.25}$$

where,

$$[\mathbf{V}_{ij}]_{4\times 4} = \sum_{k=0}^{N_{ij}-1} \delta_{ij}^k \mathbf{U}_k = \sum_{k=0}^{N_{ij}-1} \delta_{ij}^k \mathbf{A}_k^T \mathbf{A}_k = \sum_{k=0}^{N-1} \delta_{ij}^k [\mathbf{Q}(\mathbf{v}_i^k) - \mathbf{W}(\mathbf{v}_j^k)]^T [\mathbf{Q}(\mathbf{v}_i^k) - \mathbf{W}(\mathbf{v}_j^k)] \tag{C.26}$$

For n images, the total number of relative quaternions required to be computed is given by $N_{\mathbf{q}} = C_n^2$. Therefore, the unknown parameter vector is given by the upper triangular part of the following $n \times n \times 4$ symmetric tensor:

$$\Gamma = \begin{bmatrix} \mathbf{q}_{11} & \mathbf{q}_{12} & \mathbf{q}_{13} & \cdot & \mathbf{q}_{1i} & \cdot & \mathbf{q}_{1n} \\ \mathbf{q}_{21} & \mathbf{q}_{22} & \mathbf{q}_{23} & \cdot & \mathbf{q}_{2i} & \cdot & \mathbf{q}_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathbf{q}_{n1} & \mathbf{q}_{n2} & \mathbf{q}_{n3} & \cdot & \mathbf{q}_{ni} & \cdot & \mathbf{q}_{nn} \end{bmatrix} \tag{C.27}$$

The unknown parameter vector is given by

$$\mathbf{q}_{node} = \bigtriangledown(\Gamma) = [\mathbf{q}_{12}^T, \mathbf{q}_{13}^T, \cdot, \mathbf{q}_{1n}^T, \mathbf{q}_{23}^T, \mathbf{q}_{24}^T, \cdot, \mathbf{q}_{2n}^T, \cdot, \mathbf{q}^{ij}, \cdot, \mathbf{q}_{in}^T, \cdot, \mathbf{q}_{(n-1,n)}^T]^T \tag{C.28}$$

It worths noting that since not all images are adjacent, the number of relative quaternions to compute will be considerably reduced, leading to sparse matrices.

Absolute quaternions computation. In order to compute simultaneously the absolute quaternions relating all views wrt a global reference frame, one can fix the reference frame to an arbitrary image, let's say I_1, and set its associated unit quaternion $\mathbf{q}_1$ to identity. This resumes to the minimization of the criterion $Q_{mosaic-\times}$ wrt $4(n-1)$ parameter vector $\mathbf{q}_{node-abs}^T = \{\mathbf{q}_{12}^T, .., \mathbf{q}_{1n}^T\} \equiv \{\mathbf{q}_2^T, .., \mathbf{q}_n^T\}$. It can also can be observed that $\mathbf{q}_{node-abs}^T \subset \mathbf{q}_{node}^T$. Therefore, the absolute quaternions for an arbitrary reference system fixed to an image I_i is given by the i^{th} row of the matrix Γ.

Case n=3. Let us now see what becomes the result deduced in Equation (C.25) for the 3-image case. The optimal quaternions to compute form the upper-triangular part of the $3 \times 3 \times 4$ tensor expressed as:

$$\mathbf{\Gamma}_{3\times 3} = \begin{bmatrix} \mathbf{q}_{11} & \mathbf{q}_{12} & \mathbf{q}_{13} \\ \mathbf{q}_{21} & \mathbf{q}_{22} & \mathbf{q}_{23} \\ \mathbf{q}_{31} & \mathbf{q}_{32} & \mathbf{q}_{33} \end{bmatrix} \tag{C.29}$$

Therefore, the unknown parameter vector of size $4N_q \times 1 = 12 \times 1$ is given by the concatenation of the $N_{\mathbf{q}} = C_n^2 = C_3^2 = 3$ unit quaternions :

$$\mathbf{q}_{node} = [\mathbf{q}_{12}^T, \mathbf{q}_{13}^T, \mathbf{q}_{23}^T] \tag{C.30}$$

For $n = 3$ the cost function becomes:

$$Q_{mosaic-\times} = \sum_{i=0}^{2} \sum_{j=0}^{2} w_{ij} Q_{ij-\times} \tag{C.31}$$

Supposing that all images are adjacent, i.e. $\forall i, j = 0, ..2, i \neq j, w_{ij} = 1$ and by developing the previous equation we obtain:

$$\sum_{k=0}^{N-1}(Q_{12-\times} + Q_{13-\times} + Q_{23-\times}) = \mathbf{q}_{node}^T \begin{bmatrix} \mathbf{V}_{12} & 0 & 0 \\ 0 & \mathbf{V}_{13} & 0 \\ 0 & 0 & \mathbf{V}_{23} \end{bmatrix} \mathbf{q}_{node} \tag{C.32}$$

By posing the matrix measures

$$\mathbf{E}_{12\times 12} = \begin{bmatrix} \mathbf{V}_{12} & \mathbf{0}_{4\times 4} & \mathbf{0}_{4\times 4} \\ \mathbf{0}_{4\times 4} & \mathbf{V}_{13} & \mathbf{0}_{4\times 4} \\ \mathbf{0}_{4\times 4} & \mathbf{0}_{4\times 4} & \mathbf{V}_{23} \end{bmatrix} \tag{C.33}$$

we obtain a more compact form of the final cost function:

$$Q_{mosaic-\times} = \mathbf{q}_{node}^T \mathbf{E} \mathbf{q}_{node} \tag{C.34}$$

whose minimization provides us with the optimal unit quaternions.

Case for an arbitrary number of images n. We recall that total number of quaternions to compute is given by $N_\mathbf{q} = C_n^2$. The size of the matrix $\mathbf{E}$ is then $4N_\mathbf{q} \times 4N_\mathbf{q}$ and the size of the vector to estimate is $4N_\mathbf{q} \times 1$. The squared matrix $\mathbf{E}$ of size $4N_\mathbf{q}$ and the corresponding unknown parameters vector take the following form:

$$\mathbf{E} = \begin{bmatrix} \mathbf{V}_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \cdot & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{V}_{1n} & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & \mathbf{V}_{23} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdot & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{V}_{2n} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdot & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{V}_{(n-1)n} \end{bmatrix} \tag{C.35}$$

$$\mathbf{q}_{node} = \begin{bmatrix} \mathbf{q}_{12} \\ \cdot \\ \mathbf{q}_{1n} \\ \mathbf{q}_{23} \\ \cdot \\ \mathbf{q}_{2n} \\ \cdot \\ \mathbf{q}_{(n-1)n} \end{bmatrix} \tag{C.36}$$

By constraining the system so that the estimated quaternions must be unitary, we obtain the following constrained minimization problem:

$$\mathbf{q}_{ij}^T \mathbf{V}_{ij} \mathbf{q}_{ij} + \lambda_{ij}(1 - \mathbf{q}_{ij}^T \mathbf{q}_{ij}) = 0 \tag{C.37}$$

yielding,

$$\frac{\partial(\mathbf{q}^T \mathbf{V}_{ij} \mathbf{q} + \lambda_{ij}(1 - \mathbf{q}^T \mathbf{q}))}{\partial \mathbf{q}} = 0 \Leftrightarrow \mathbf{q}^T \mathbf{V}_{ij} - \lambda_{ij} \mathbf{q}^T = 0 \Leftrightarrow \mathbf{V}_{ij} = \lambda_{ij} \tag{C.38}$$

The total number of unit quaternions $N_\mathbf{q}$ which minimize the total cost $Q_{mosaic-\times}$ are therefore given by the $N_\mathbf{q}$ eigen-vectors of the $\mathbf{V}_{ij}$ matrices associated to their smallest eigen values λ_{ij}.

C.8 Optimal Rigid Transformation using Sum of the Squared Residual Errors

Let us now consider that we have access to the 3D point coordinates of a set of N image matches previously established using a feature extraction and matching algorithm. The following description recalls the technique reported in [Horaud and Monga, 1995] which solves for the optimal rigid transformation reported by minimizing the following cost function:

$$[\mathbf{R}^*, \mathbf{t}^*] = \arg\min_{[\mathbf{R},\mathbf{t}]} \sum_{k=0}^{N-1} \|\mathbf{p}_1^k - \mathbf{R}\mathbf{p}_2^k - \mathbf{t}\|^2 \tag{C.39}$$

If $[\mathbf{R}^*, \mathbf{t}^*]$ denote the optimal rotation and translation resulted from Equation C.39, then we have the following property relating the corresponding centers of gravity of the two clusters noted by $\bar{\mathbf{v}}_1$ and $\bar{\mathbf{v}}_2$ [Lin et al., 1986]:

$$\bar{\mathbf{p}}_1 = \mathbf{R}^*\bar{\mathbf{p}}_2 + \mathbf{t}^* \tag{C.40}$$

and by using centered coordinates $\breve{\mathbf{p}}_i^k = \mathbf{p}_i^k - \bar{\mathbf{p}}_i$ with $i = 1, 2$ we obtain:

$$\mathbf{p}_1^k - \mathbf{R}^*\mathbf{p}_2^k - \mathbf{t}^* = \breve{\mathbf{p}}_1^k - \mathbf{R}^*\breve{\mathbf{p}}_2^k \tag{C.41}$$

This allows to split the optimal rigid transformation criterion in expressed in Equation (C.39) in two different ones solving separately for the optimal rotation and translation.

$$\begin{cases} \mathbf{R}^* = \arg\min_{\mathbf{R}} \sum_{k=0}^{N-1} \|\breve{\mathbf{p}}_1^k - \mathbf{R}\breve{\mathbf{p}}_2^k\|^2 \\ \mathbf{t}^* = \bar{\mathbf{p}}_1 - \mathbf{R}^*\bar{\mathbf{p}}_2 \end{cases} \tag{C.42}$$

This methods solves for the rigid transformation in two steps. The rotation is first computed, which is exploited within the second stage for translation estimation, being seen as an optimization step. As stated in [Horaud and Monga, 1995], the proposed method has its advantages and its inconveniences. Its main positive is related to the rotation separation which allows to design an elegant and numerically stable solution for the rotation estimation. On the downside, the translation estimation is subject to the accuracy of the rotation estimate.

Let us now turn back to our mosaicing problem by recalling that in this case study, the images were acquired from the same optical center and even if small amounts of parallax are susceptible of being introduced, there can be compensated by the local motion estimation procedure integrated within the pair-wise alignment process described in Section 5.7. Consequently, the proposed method can be applied to estimate a pure rotation, without being biased by eventual parallax or subject to the residual translation.

Optimal quaternion computation for a couple of adjacent images. We describe hereafter the solution reported in [Horaud and Monga, 1995] which solves for rotation estimation using unit quaternions. The main advantage of this technique stands in the choice of the rotation parametrization. Is more easy to constraint the system to produce an unit quaternion, instead of imposing 6 constraints to ensure the orthogonality of the 3×3 rotation matrix.

We recall hereafter the notation of a unit quaternion describing a rotation of angle θ around the $\vec{\mathbf{n}}$:

$$\mathbf{q} = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(in_x + in_y + kn_z) \tag{C.43}$$

The criterion allowing to compute the optimal rotation for N homologous 3D vectors is given by:

$$Q_{12} = \min_{\mathbf{R}} \sum_{k=0}^{N-1} \|\mathbf{v}_1^k - \mathbf{R}\mathbf{v}_2^k\|^2 \tag{C.44}$$

Using unit quaternions, the rotation of an arbitrary 3D vector $\mathbf{v}$ is written as:

$$\mathbf{v}_1 = \mathbf{R}\mathbf{v}_2 = \mathbf{q} * \mathbf{v}_2 * \bar{\mathbf{q}} \tag{C.45}$$

where $\bar{\mathbf{q}}$ represent the conjugate quaternion with the property $\mathbf{q} * \bar{\mathbf{q}} = \mathbf{q} \cdot \mathbf{q} = \|\mathbf{q}\|^2 = 1$. The criterion from Equation (C.44) becomes:

$$\min_{\mathbf{q}} \sum_{k=0}^{N-1} \|\mathbf{v}_1^k - \mathbf{q} * \mathbf{v}_2^k * \bar{\mathbf{q}}\|^2 \tag{C.46}$$

Using quaternions' properties, from Equation (C.46) we have successively:

$$\|\mathbf{v}_1^k - \mathbf{q} * \mathbf{v}_2^k * \bar{\mathbf{q}}\|^2 = \|\mathbf{v}_1^k - \mathbf{q} * \mathbf{v}_2^k * \bar{\mathbf{q}}\|^2 \|\mathbf{q}\|^2 = \tag{C.47}$$

$$\Leftrightarrow \|\mathbf{v}_1^k * \mathbf{q} - \mathbf{q} * \mathbf{v}_2^k * \bar{\mathbf{q}} * \mathbf{q}\|^2 = \|\mathbf{v}_1^k * \mathbf{q} - \mathbf{q} * \mathbf{v}_2^k\|^2 \tag{C.48}$$

By using matrix notations we can write the product of two quaternions in the following form:

$$\mathbf{v}_1 * \mathbf{q} = \mathbf{Q}(\mathbf{v}_1)\mathbf{q} \text{ and } \mathbf{q} * \mathbf{v}_2 = \mathbf{W}(\mathbf{v}_2)\mathbf{q} \tag{C.49}$$

where, $\mathbf{Q}(\mathbf{v})$ and $\mathbf{W}(\mathbf{v})$ are the associated antisymmetrical matrices of a pure-imaginary quaternion $\mathbf{v} = [0, v_x, v_y, v_z]^T$ of the form:

$$\mathbf{Q}(\mathbf{v}) = \begin{bmatrix} 0 & -v_x & -v_y & -v_z \\ v_x & 0 & -v_z & v_y \\ v_y & v_z & 0 & -v_x \\ v_z & -v_y & v_x & 0 \end{bmatrix} \tag{C.50}$$

$$\mathbf{W}(\mathbf{v}) = \begin{bmatrix} 0 & -v_x & -v_y & -v_z \\ v_x & 0 & v_z & -v_y \\ v_y & -v_z & 0 & v_x \\ v_z & v_y & -v_x & 0 \end{bmatrix} \tag{C.51}$$

with the properties:

$$\mathbf{Q}(\mathbf{v})^T = -\mathbf{Q}(\mathbf{v}) \text{ and } \mathbf{W}(\mathbf{v})^T = -\mathbf{W}(\mathbf{v}) \tag{C.52}$$

Using the aforementioned properties, the right term in Equation C.48 becomes:

$$\|\mathbf{v}_1^k * \mathbf{q} - \mathbf{q} * \mathbf{v}_2^k\|^2 = [[\mathbf{Q}(\mathbf{v}_1^k) - \mathbf{W}(\mathbf{v}_2^k)]\mathbf{q}]^T [[\mathbf{Q}(\mathbf{v}_1^k) - \mathbf{W}(\mathbf{v}_2^k)]\mathbf{q}] = \mathbf{q}^T \mathbf{A}_k \mathbf{q} \tag{C.53}$$

where $\mathbf{A}_k$ is a 4×4 symmetric matrix containing the measures:

$$\mathbf{A}_k = [\mathbf{Q}(\mathbf{v}_1^k) - \mathbf{W}(\mathbf{v}_2^k)]^T \mathbf{Q}(\mathbf{v}_1^k) - \mathbf{W}(\mathbf{v}_2^k) \tag{C.54}$$

For a set of N point correspondences, the criterion from Equation C.44 becomes:

$$\hat{\mathbf{q}} = \arg\min_{\mathbf{q}} \sum_{k=0}^{N-1} \mathbf{q}^T \mathbf{A}_k \mathbf{q} = \arg\min_{\mathbf{q}} (\mathbf{q}^T (\sum_{k=1}^{N-1} \mathbf{A}_k)\mathbf{q}) = \arg\min_{\mathbf{q}} (\mathbf{q}^T \mathbf{B}\mathbf{q}) \tag{C.55}$$

where $\mathbf{B} = \sum_{k=1}^{N-1} \mathbf{A}_k = \sum_{k=1}^{N-1} [\mathbf{Q}(\mathbf{v}_1^k) - \mathbf{W}(\mathbf{v}_2^k)]^T [\mathbf{Q}(\mathbf{v}_1^k) - \mathbf{W}(\mathbf{v}_2^k)]$ is a symmetric and positive definite matrix. By introducing the constraint that the quaternion must have an unit norm, the criterion takes the following form:

$$\hat{\mathbf{q}} = \arg\min_{\mathbf{q}} Q = \arg\min_{q} (\mathbf{q}^T \mathbf{B} \mathbf{q} + \lambda(1 - \mathbf{q}^T \mathbf{q})) \tag{C.56}$$

By deriving Q wrt $\mathbf{q}$, we get:

$$\frac{\partial Q}{\partial \mathbf{q}} = 0 \Leftrightarrow \mathbf{B}\mathbf{q} - \lambda \mathbf{q} = 0 \Rightarrow (\mathbf{B} - \lambda)\mathbf{q} = 0 \tag{C.57}$$

from which can deduce

$$Q = \lambda \tag{C.58}$$

meaning that the unit quaternion minimizing Q is the eigen vector of $\mathbf{B}$ corresponding the smallest eigen value of $\mathbf{B}$, λ.

Proof: A symmetric matrix has real eigen values and a symmetric positive definite matrix has its eigen values positive. Let then $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4)$ be the eigen vectors of $\mathbf{B}$ forming an orthogonal base. If vectors are unitary, then the base is ortho-normalized. The quaternion $\mathbf{q}$ can then be written as a linear combination of the eigen vectors $\mathbf{e}$, i.e. :

$$\mathbf{q} = \sum_{i=1}^{4} \mu_i \mathbf{e}_i \tag{C.59}$$

and for all $i = 1, 2, 3, 4$ we have:

$$\mathbf{B}\mathbf{e}_i = \lambda_i \mathbf{e}_i, \lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 \tag{C.60}$$

and by exploiting eigen vectors orthonormality, we get:

$$\mathbf{q}^t \mathbf{B} \mathbf{q} = \sum_{i=1}^{4} \mu_i^2 \lambda_i \tag{C.61}$$

which reaches its minimum for

$$\begin{cases} \mu_1 = 1, \\ \mu_2 = \mu_3 = \mu_4 = 0 \end{cases} \tag{C.62}$$

Hence, we have:

$$\begin{cases} \mathbf{q} = \mathbf{e}_1 \\ \mathbf{q}^T \mathbf{B} \mathbf{q} = \lambda_1 \end{cases} \tag{C.63}$$

The optimal quaternion is therefore given by the eigen vector of $\mathbf{B}$ associated to its smallest eigen value.

q.e.d$\triangledown$

Appendix D

Complements to Chapter 7

The following two sections are concerned with the the autonomous site digitization and exploration problem. We start this section by presenting the autonomous site digitization capability and by revisiting several systems which are heavily relying on it through semi-autonomous methods. Section D.2 states the site digitization and exploration problem emphasizing the complementarity of the existing solutions.

D.1 Space's and Earth's Needs for Autonomous Exploration

Designing unmanned mobile platforms capable to drive themselves in previously unknown environments while gathering rich information about their surroundings represents a key aspect standing behind the feasibility of various military and civilian missions undertaken in hostile environments, where human presence is highly undesirable.

The great potential of unmanned systems stands in their capability to reach and explore terrains which are inaccessible or considered too dangerous for humans. Mostly, the deployment of such systems is driven by practical constraints. Although remotely-controlled by human experts, such endeavors have already been demonstrated in Space and on Earth to supply inspection, monitoring, exploration, searching and rescuing (SAR) operations.

Producing digital copies of large-scale and complex archeological sites, whose preservation is highly subject to weather conditions which cause high risk of collapse in any time, has led to a considerable research work in Remote Sensing, Computer Vision and Robotics research communities. Such environments preclude the access of human surveyors and unmanned mobile systems are highly desirable to supply photorealist site surveys. In addition, the digitization of prehistorical sites allows paleontologist to perform geomorphological studies revealing human evolution. Figure D.1 illustrates several complex sites requiring digitization.

The use of 3D dense mapping for underwater exploration of mineral resources on the ocean floor is reported in [Jasiobedzki and Jakola, 2007]. Generally, these systems are controlled from the surface and removing the human operator from the loop poses several technical challenges, such as low situational awareness in unknown environments, complexity of the operated machines and the need for high band-width and real-time communication systems. Currently, precise maps of underwater worksites are not available and accurate positioning after several kilometer of navigation is still an open issue. Moreover, tele-operated underwater campaigns are extremely expensive, requiring for highly skilled operator, being strongly related to the cost of an error [Hainsworth, 2001], [Liu et al.,

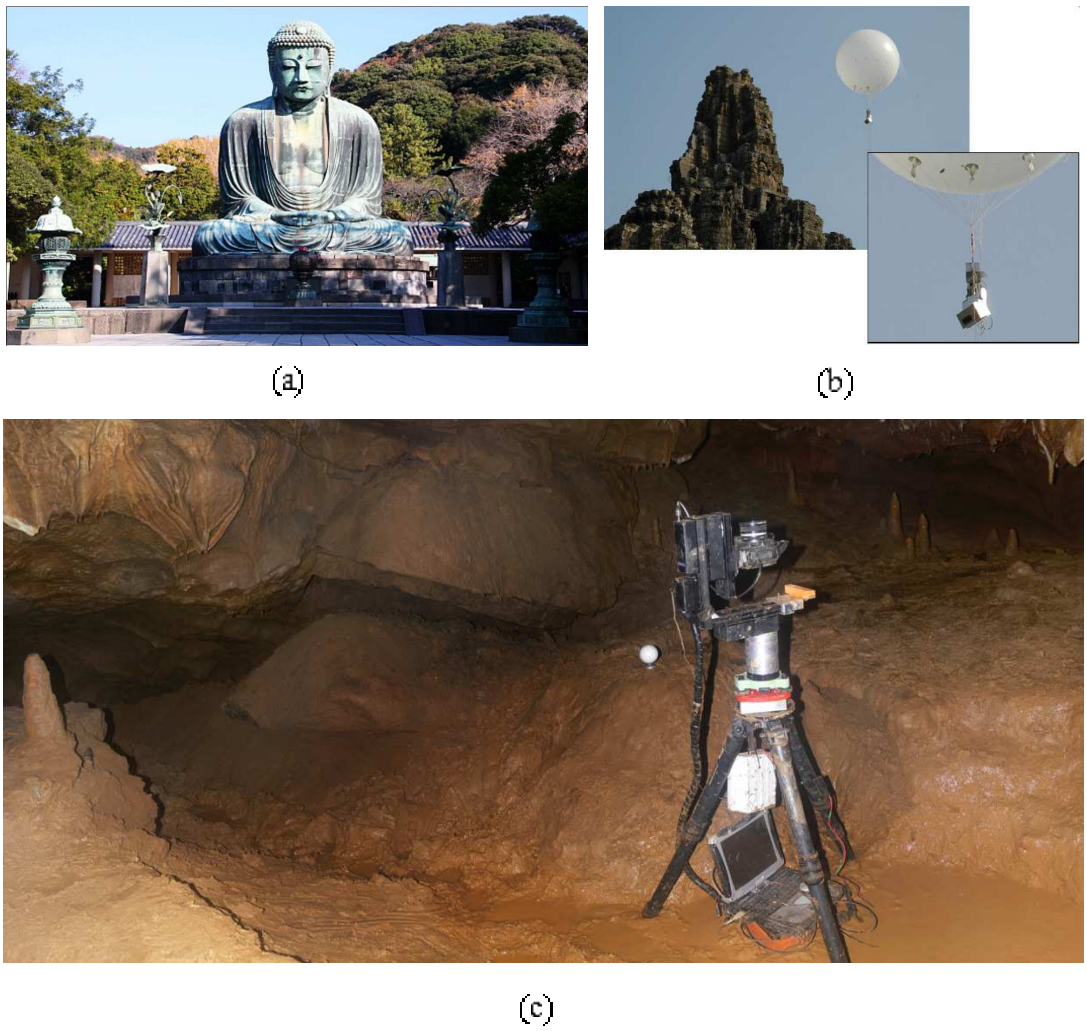


Figure D.1: Exemples of cultural heritage applications. (a) The Great Buddha Statue 13.35m (Kamakura, Japan). Cast in 1252, the statue has stood the test of time, including a tsunami which hit Kamakura in 1945. (b) the remotely control Flying Laser Range Sensor (FLRS) introduced in [Banno et al., 2008] digitizing the Bayon Temple (Angkor, Siem Reap, Cambodia) built around 1190 by King Jayavarman VII. (c) Data acquisition in Mayenne Science prehistoric cave (France) discovered in 1967 by Roger Bouillon ©IGN-MATIS 2009 - photo provided by Jean-Pierre Papelard.

2001], [Bailey, 2002], [Ralston et al., 2005]. Due to all these aspects, it is difficult to exploit the underwater resources economically and without environmental damage.

The operation of on-orbit robotic systems or of those designed for planetary exploration present similar challenges. In both contexts, recent research works addressed the tele-operation difficulty by presenting virtual models of worksites to the human expert and by increasing the autonomy of the remote system. For instance, for the recently site survey missions performed on Mars, data acquisition is performed automatically, transmitted on Earth and processed by computer vision experts. Furthermore, rovers receive commands from human operator for obstacles' avoidance and path planning. Such human dependency leads to inefficient and delayed missions, representing therefore a major bottleneck in performing rapidly complex tasks in critical situations. Figure D.2 (a) illustrates the

concept for a robotic explorer introduced in [Deans et al., 1998] designed for lunar south pole exploration. The platform employs a hybrid tele-operated/autonomous navigation approach. Figure D.2 (b) shows the rock modeling and matching presented in [Li et al., 2007] for autonomous Mars rover localization.

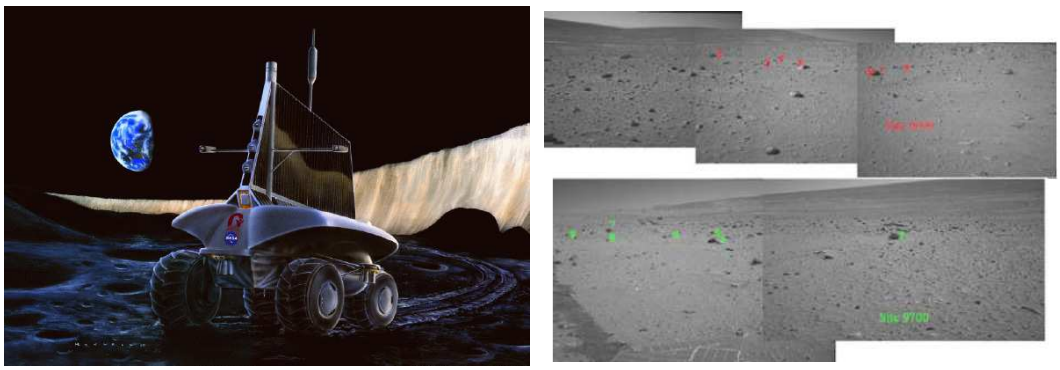


Figure D.2: (a) Icebreacker [Deans et al., 1998]. (b) Automatic rocks matching [Li et al., 2007].

The great advances in technology and the increasing number of natural catastrophes (high-magnitude earthquakes - Kobe 1995, Haiti 2010) and human-caused disasters (New York city 2001) lead to a high number of research works willing to provide robots for emergency response. These platforms are sent to explore unstructured terrains and to build a map that can be used by humans rescuers to retrieve victims.

In [Pilania and Chakravarty, 2008] a sense, communicate, plan and act paradigm is used along with a wireless visual sensor for a semi-autonomous mine navigation system designed to supply post-disaster rescue operation planning. Such a system aims at minimizing the chances of accidents during the extraction of the existing mineral resources by providing navigation and regular monitoring based on semi-autonomous robots. A robotic vehicle can explore the mine and provide valuable help for human rescuers for planning SAR missions. Figure D.3 illustrates the unmanned system of University of Technology Sydney designed to supply semi-autonomous searching and rescuing mission in mines.

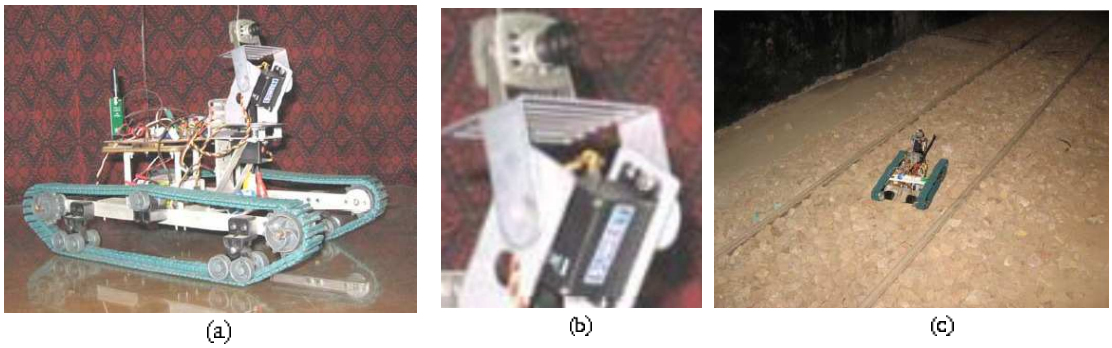


Figure D.3: The remotely-controlled unmanned system presented in [Pilania and Chakravarty, 2008] equipped with a pan-tilt mechanism (b) for improved navigation. (c) system's deployment in a testing mine illustrating the unstructured terrain.

The autonomy of unmanned mobile systems. All the aforementioned missions are

nowadays subject to the system's autonomy feasibility which has not been yet demonstrated and currently relies on heavy human operator intervention. Unmanned mobile platform's autonomy is related to the real world perception and consequently, it must be embedded with environment modeling functionalities. This calls for visual perception through multi-sensor fusion for producing a rich 3D reconstruction of the real world, allowing thereby the development of automatic reasoning and decisional resources.

Unmanned platforms are usually embodied with proprioceptive sensors (odometry, IMU) to localize themselves and exteroceptive devices (vision and tactile) to build local environment models in which they can localize themselves using the first class of sensors. It opposite to the latter one, localization devices are subject to accumulation errors for long term trajectories and traditional approaches are build on a combination of both to reduce drift. For missions undertaken in GPS-denied and unstructured environments, one of the most employed proprioceptive senses - odometry, is estimated through purely visual means [Nister and Stewenius, 2006], [Craciun et al., 2006].

Active perception and visual autonomy paradigm. Unmanned mobile platforms equipped with suitable vision sensors allows recording rich digital representations of the real world. This gives rise to an artificial vision engine based on which onboard intelligence can be developed, providing unmanned systems with significant autonomy, enabling them to execute the required tasks. The unmanned systems' autonomy deals with the system's capacity to learn characteristics without assistance and to adapt dynamically to unexpected environment changes.

Before conducting task-specific actions, the platform must become familiar with the environment and build an internal model of its surroundings. Active perception paradigm presents promising results in this direction. Instead of planning the required tasks using poor description of the environment, the robot interacts first with its surroundings in order to gather as much information as possible. This is an active research topic which has recently been attacked by researchers aiming to build mobile robots able to explore previously unknown environments and conduct task-specific actions without a central control system, special landmarks or human interaction. In this context, the exploration phase is of primer interest as it provides knowledge about the environment and situational awareness, allowing to the platform to see, detect and act in order to fulfill the required tasks.

D.2 The Visual-based Autonomous Site Exploration Problem

The main goal of our research work is concerned with the automatic 3D digitization of *complex environments* through the jointly use of image and laser data. The term *complex* denotes the accidental terrain which leads to a high number of occluded areas. It also denotes the inaccessibility of the site which precludes human surveyors access and consequently, the 3D modeling process requires for unmanned systems to be deployed.

Moreover, since the environment is considered to be previously unknown, no map is assumed to be available. Therefore, the system must be able to explore autonomously the site and build dynamically 3D models in order to generate an occlusions-free 3D model of the entire site. This is usually refereed to as the *autonomous site digitization and exploration problem* which is willing to answer to the following question: *given the currently build 3D model, where the system should move to gain as much new information as possible?* Additional constraints are introduced by the in-situ aspect, i.e. avoid cost's displacement

and minimizing data redundancy.

Global solutions. Several research works have attacked the exploration problem in its entirety [Kuipers and Byun, 1991], [Bolduc et al., 1996], [Yamauchi et al., 1998], [Baker et al., 2004b], [Nüchter et al., 2004]. Reported methods are based on the wall-following principle which is too simple to be applied for the exploration of complex environments [Mataric, 1997], [Duckett and Nehmzow, 1997]. In [Lee, 1996] authors exploit parallel and perpendicular walls and validate the proposed method in simple environments (three corridors). In [Thurn and Bucken, 1996] authors combine evidence grids with topological maps and demonstrate the feasibility of the system in a large-scale building. The technique assumes that walls are observable, without being obstructed by obstacles. At INRIA, AROBAS research team is currently focusing to address both autonomous navigation and SLAM problems jointly to provide perception, modeling and platform's control. In [Victorino et al., 2003], simulation results showed that a sensor-based control approach adds constraints on the relative pose and its local environment.

Partial solutions. Researchers split the autonomous exploration problem into smaller tasks and introduce solutions for several ingredients of the visual-based autonomous navigation: SLAM, visual-odometry, path planning and obstacles' detection. Unfortunately, none of these procedures include visual servoing schemes nor motion control. In addition, the autonomous navigation stage supposes that a map is available from a prior processing step and no exploration is performed on line.

3D modeling of large-scale scenes. Solving for the 3D digitization problem provides unmanned systems with rich visual information, giving the possibility to see, detect and act in order to explore the environments. A considerable amount of research work reported photorealist 3D modeling of large-scale scenes from image-laser data fusion (Appendix A.3 presents a survey on the existing 3D modeling systems).

All the aforementioned systems do not provide visual feedback for on-line exploration in order to make possible the complete digitization of the site. Generally, the main problem is raised by the capacity to generate dynamically textured 3D models along with visual servoing procedures. This provides sensor's control to supply decisions and actions contributing to the 3D scene model completeness. Such systems have been reported for supplying 3D modeling of small scale objects by humanoids in [Foissotte et al., 2008a], [Foissotte et al., 2008b] and we believe that a similar approach represent be a promising research direction for automatic 3D modeling and exploration of large-scale environments.

Bridging between Robotics and Computer Vision world modeling methods. Field robotic applications require rich 3D maps in order to allow visual-based autonomy. Their result provide localization and mapping to be exploited along with visual servoing procedures. On the other hand, computer vision researchers aims at producing 3D digital copies of complex and difficult to access sites [Banno et al., 2008], [Craciun et al., 2008]. To this end, researchers employ remotely-controlled unmanned platforms embodied with different vision sensing devices [Banno et al., 2008] and build rich 3D models of complex environments.

Robotics and Computer Vision research communities report real world modeling methods but without exploiting them in a exhaustive way along with visual servoing procedures to provide visual-autonomy. This an important issue which must be addressed which allows to solve active research topics in both research communities, allowing (1) to provide autonomous site digitization and exploration for producing rich and complete maps of difficult to access environments and (2) based on these maps, visual-autonomy can be build in order to allow the execution of complex tasks.

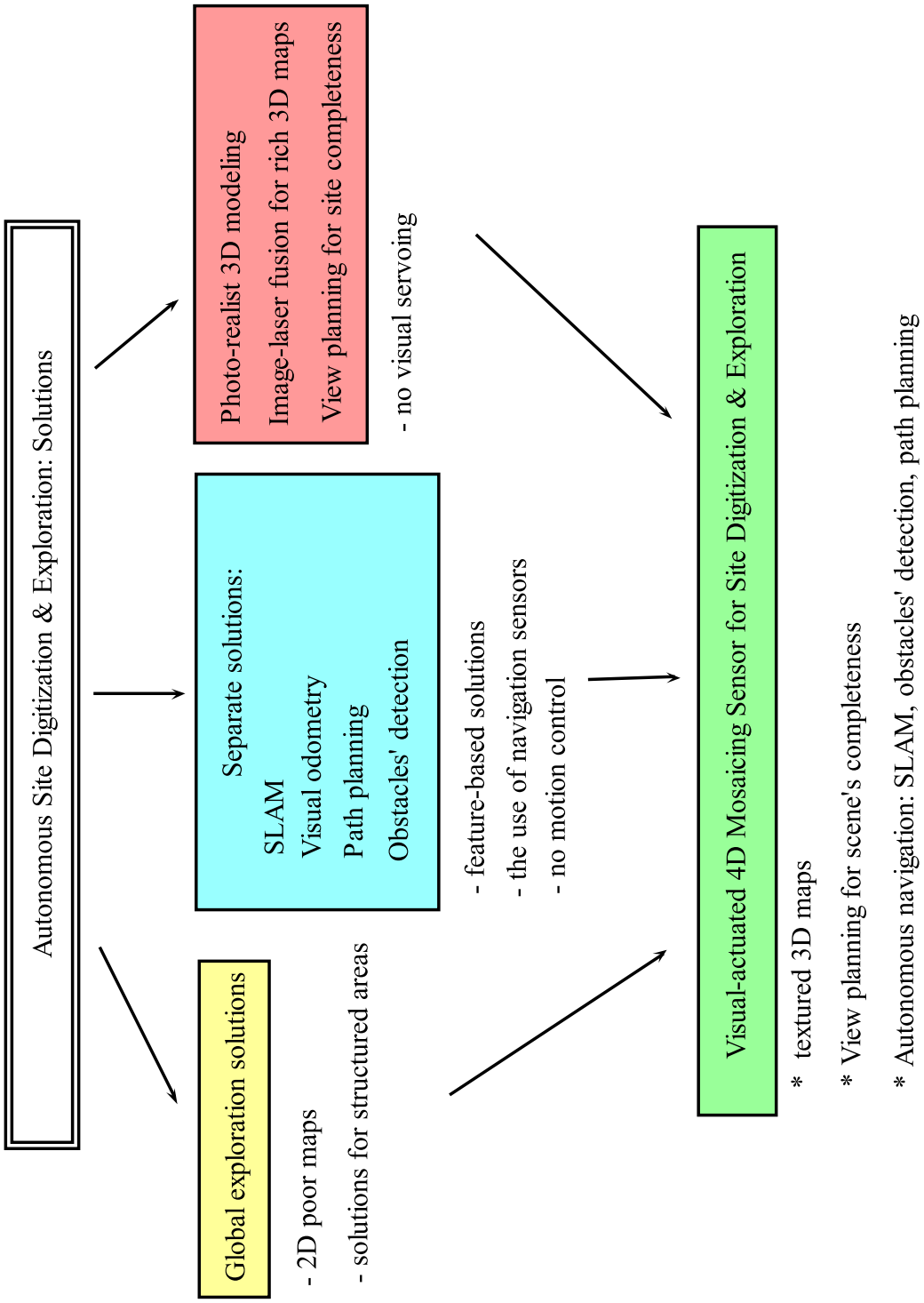


Figure D.4: The available site exploration and digitization methods. The proposed strategy is highlighted in green.

D.3 Complement to Section 7.4

Figure D.5 illustrates Harris corners' extraction of 2D mosaic views obtained on a data set acquired in the Tautavel prehistoric cave. Figure D.6 shows the result obtained when SIFT descriptors are extracted on the same data set.

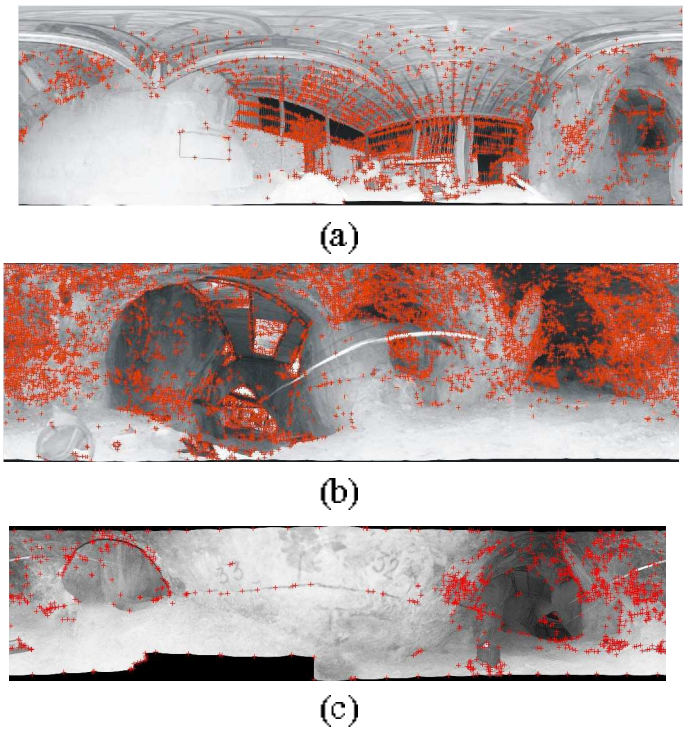
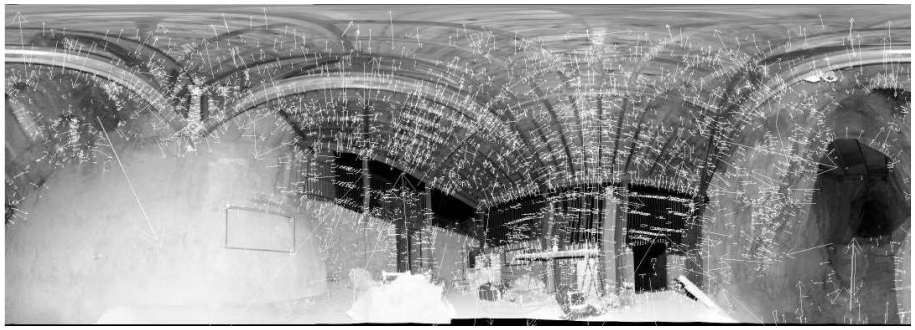
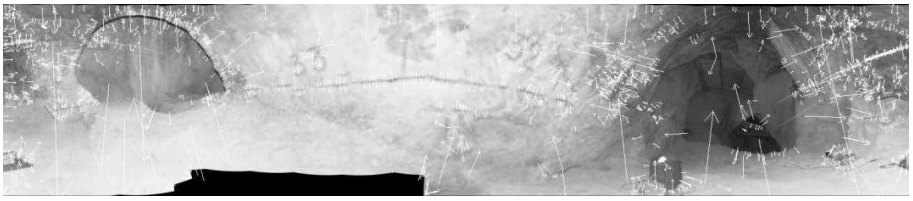


Figure D.5: Harris features extraction on three partially overlapped mosaic views acquired in Tautavel prehistoric cave. (a) M^1_{2D} size 2710×816,♯Harris: 3299 (b) M^2_{2D} size 2771×811 ♯ Harris: 7648, (c) M^3_{2D} size 2775×652,♯Harris: 1100.



(a)

(b)



(c)

Figure D.6: SIFT [Lowe, 2004] extraction on the 2D panoramic views illustrated in Figure 7.6. (a) ♯SIFT on M^1_{2D}: 6694, (b) ♯SIFT on M^2_{2D}: 11816, (c) ♯SIFT on M^3_{2D}: 1442.

Thesis Publications

Journal

- **Multi-view Scans Alignment for 3D Spherical Mosaicing in Large-Scale Unstructured Environments.**

  D. Craciun, N. Paparoditis and F. Schmitt.

  To appear in *Computer Vision and Image Understanding Volume 114, Issue 11, November 2010, pages 1248-1263 - Special issue on Embedded Vision.*

Conferences

- **Automatic Gigapixel Mosaicing in Large Scale Unstructured Underground Environments.**

  D. Craciun, N. Paparoditis and F. Schmitt.

  In IAPR Conference on Machine Vision Applications, Yokohama, Japan, May 2009.

- **Automatic Pyramidal Intensity-based Laser Scan Matcher for 3D Modeling of Large Scale Unstructured Environments.**

  D. Craciun, N. Paparoditis and F. Schmitt.

  In IEEE Fifth Canadian Conference on Computer and Robot Vision, Windsor, Ontario, Canada, May 2008.

- **Bundle adjustment and pose estimation of images of a multiframe panoramic camera.**

  B. Cannelle, D. Craciun, N. Paparoditis, D. Boldo.

  In 9th Conference on Optical 3-D, Vienna, Austria, July 2009.

Bibliography

A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. *In Proceedings on ACM Transactions on Graphics*, 23(3):292–300, 2004.

B. Allen, B. Curless, and Z. Popovic. The space of human body shapes: reconstructino and parametrization from range scans. *In ACM Transactions on Graphics*, 22(3):587–594, 2003a.

P. Allen, I. Stamos, A. Gueorguiev, E. Gold, and P. Blaer. Avenue: Automated site modeling in urban environments. *In Proceedings of the 3rd International Conference on 3D Digital Imaging and Modeling (3DIM'01)*, 2001.

P. Allen, P. K. Allen, A. Troccoli, B. Smith, I. Stamos, and S. Murray. The Beauvais cathedral project. *Workshop on Applications of Computer Vision in Archeology, IEEE International Conference of Computer Vision and Pattern Recognition*, pages 731–737, 2003b.

P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *In International Journal of Computer Vision*, 2(3):283–310, 1989.

P. Anandan. Ieee workshop on visual scenes. *In IEEE Computer Society Press, Cambridge, Massachusetts*, 1995.

S. Argamon-Engelson. Using image signatures for place reconition. *In Pattern Recognition Letters*, 19(4):941–951, 1998.

K. S. Arun, T. S. T. S. Huang, and S. D. Blostein. Least square fitting of two 3-d point sets. *In IEEE Journal on Pattern Analysis and Machine Intelligence*, 9(5):698–700, 1987.

K.-H. Bae. Automated registration of unorganized point clouds from terrestrial laser scanners. *Ph. D. thesis, Curtin University of Technology, Perth, Australia*, 2006.

T. Bailey. Mobile robot localisation and mapping in extensive outdoor environments. *PhD thesis, University of Syndey*, 2002.

T. Bailey. Constrained initialization for bearing-only SLAM. *In Proceedings of IEEE International Conference on Robotics and Automation (ICRA'04)*, 2003.

T. Bailey and H. Durrant-White. Simultaneous localization and mapping: Part II. *In Proceeding of IEEE Robotics and Automation Magazine*, 13(2):99–110, 2006.

C. Baker, A. Morris, D. Ferguson, S. Thayer, C. Whittaker, Z. Omohundro, C. Reverte, W. Whittaker, D. Hahnel, and S. Thurn. A campaign in autonomous mine mapping. *In Proceedings of International Conference on Robotics and Automation (ICRA'04)*, 2004a.

C. Baker, A. C. Morris, D. Ferguson, S. Thayer, C. Whittaker, Z. Omohundro, C. Reverte, W. L. Whittaker, D. Haehnel, and S. Thrun. A campaign in autonomous mine mapping. *In Proceedings of the IEEE Conference on Robotics and Automation (ICRA)*, 2, 2004b.

S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework: Part i: The quantity approximated, the warp update rule, and the gradient-descent approaximation. *In International Journal on Computer Vision*, 56(3):221–255, 2004a.

S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework: Part 4. *Technical Report CMU-RI-TR-04-14, The Robotics Institute, Carnegie Mellon Unversity*, 2004b.

S. Baker, R. Gross, and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *Technical Report CMU-RI-TR-03-35, The Robotics Institute, Carnegie Mellon University*, 2003.

A. Banno, T. Masuda, T. Oishi, and K. Ikeuchi. Flying Laser Range Sensor for Large-Scale Site-Modeling and Its Applications in Bayon Digital Archival Project. *In International Journal of Computer Vision*, 78(2-3):207–222, 2008.

A. Bartoli, M. Coquerelle, and P. Sturm. A framework for pencil-of-points structure-from-motion. *In Eighth European Conference on Computer Vision (ECCV'04)*, pages 28–40, 2004.

B. G. Baumgart. Geometric modeling for computer vision. *PhD thesis, Standford University*, 1974.

D. Baumgartner, P. Rossler, and W. Kubinger. Performance benchmark of dsp and fpga implementations of low-level vision algorithms. *In IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *In Journal Computer Vision and Image Understanding (CVIU)*, pages 346–359, 2008.

S. Becker. Vision-assisted 3d modeling from model-based video representation. *Ph. D. thesis, Massachusetts Institute of Technology*, 1997.

S. Becker and V. M. J. Bove. Semi-automatic 3d model extraction from uncalibrated 2-d camera views. *In SPIE Visual Data Exploration and Analysis II*, 2410:447–461, 1995.

J. S. Beis and D. G. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. *In Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pages 1000–1006, 1997.

R. Benjemaa and F. Schmitt. A solution for the registration of multiple 3d point sets using unit quaternions. *In Proceedings of the European Conference on Computer Vision (ECCV'98)*, pages 34–50, 1997.

J. Bentely. Multidimensional binary search trees used for associative searching. *In Communications of the ACM*, 18(9):509–517, 1975.

J.-A. Beraldin and L. Cournoyer. Object modeling creation from multiple range images: Acquisition, calibration, model building and verification. *In Proceedings of International on Recent Advances on 3-D Digital Imaging and Modeling*, pages 326–333, 1997.

J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. *In Proceedings of 2nd European Conference on Computer Vision (ECCV'92)*, pages 237–252, 1992.

R. Bergevin, M. Soucy, H. Gagnon, and D. Laurendeau. Toward a general multi-view registration technique. *In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 18(5):540–547, 1996.

F. Bernardini and H. Rushmeier. The 3D Model Acquisition Pipeline. *In Computer Graphics Forum*, 21(2):149–172, Octobre 2002.

F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin. The ball pivoting algorithm for surface reconstruction. *In IEEE Transactions on Visualizaiton and Computer Graphics*, 5(4):349–359, 1999.

P. J. Besl. Active, optical imaging sensors. *Machine Vision Applications*, 1:127–152, 1988.

P. J. Besl and N. D. McKay. A method for registration of 3d-shapes. *In IEEE Transactions on Pattern Recognition and Machine Intelligence*, 14(2):239–256, 1992.

P. Biber and W. Strasser. The normal distributions transform. A new approach to laser scan matching. *In Proceedings of International Conference on Intelligent Robots and Systems*, 3:2743–2748, 2003.

C. M. Bishop. *Neural Networks and Machine Learning*. Springer in cooperation with NATO Scientific Affairs Division, 1998.

M. J. Black and A. Rangarajan. On the unification of the line processes, outlier rejection and robust statistics with applications in early vision. *In International Journal of Computer Vision*, 19:57–91, 1996.

G. Blaise and M. Levine. Registering Multiview Range Data to Create 3D Computer Objects. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8): 820–824, 1995.

M. Bolduc, E. Bourque, G. Dudek, N. Roy, and R. Sim. Autonomous exploration: an integrated system approach. *In Proceedings of AAAI Conference*, pages 779–780, 1996.

D. Borrmann, J. Elseberg, K. Lingemann, A. Nuchter, and J. Hertzberg. Globally consistent 3d mapping with scan matching. *In International Journal on Robotics and Autonomous Systems*, 56:130–142, 2008.

J.-Y. Bouguet. Pyramidal implementation of the lucas-kanade feature tracker. *Technical Report, Intel Corporation, Microprocessor Research Labs*, 2000.

A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for the space carving algorithm. *In Proceedings of the Eight IEEE International Conference on Computer Vision (ICCV'01)*, pages 388–393, 2001.

D. C. Brown. Close-range camera calibration. *Photogrammetric Engineering*, 37(8):855–866, 1971.

H. Brown, A. Kim, and R. Eustice. Development of a multi-AUV SLAM testbed at the university of michigan. *In Proceedings of the IEEE/MTS OCEANS Conference and Exhibition*, pages 1–6, 2008.

L. G. Brown. A survey of image registration techniques. *In ACM Computing Surveys*, 24 (4):326–376, 1992.

M. Brown and D. Lowe. Recognising panoramas. *In Proceedings of Ninth International Conference on Computer Vision (ICCV'03)*, pages 1218–1225, 1983.

M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *In International Journal on Computer Vision*, 74:59–73, 2007.

M. Brown, R. Szeliski, and S. Winder. Multi-image matching using multi-scale oriented patches. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517, 2005.

X. Brun and F. Goulette. Modeling and calibration of coupled fish-eye CCD camera and laser range scanner for ourdoor ennvironments reconstruction. *In Proceedings of the Sixth International Conference on 3-D Digital Imaging and Modeling*, pages 320–327, 2007.

C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured lumigraph rendering. *In Proceedings of the 28th annual conference on Computer graphics and interactive techniques (SIGGRAPH '01)*, pages 425–434, 2001.

L. M. Bugayevskiy and J. P. Snyder. *Map Projections: A Reference Manual*. CRC Press, 1995.

P. J. Burt and E. H. Adelson. A multiresolution spline with application to image mosaicing. *In ACM Transactions on Graphics*, 2(4):217–236, 1983.

B. Cannelle, D. Craciun, N. Paparoditis, and D. Boldo. Bundle adjustement and pose estimation of images of a multiframe panoramic camera. *In Proceedings of 9th Conference on Optical 3-D*, 2009.

M. C. Chaing and T. E. Boult. Efficient image warping and super-resolution. *In IEEE Workshop on Applications of Computer Vision (WACV'96)*, pages 56–61, 1996.

K. Chaiyasarn, T.-K. Kim, F. Viola, R. Cipolla, and K. Soga. Image mosaicing via quadric surface estimation with priors for tunnel inspection. *In Proceedings of International Conference on Image Processing*, 2009.

S. E. Chen. Quicktime vr: An image-based approach to virtual environment navigation. *In Computer Graphics ACM SIGGRAPH '95*, pages 29–38, 1995.

S. E. Chen and L. Williams. View interpolation for image synthesis. *In SIGGRAPH'93*, 1993.

Y. Chen and G. Medioni. Object modeling by registration of multiple range images. *In Image and Vision Computing Journal*, 10(3):145–155, 1992.

Y. Chen, Y. Hung, and J. Cheng. Ransac-based darces: A new approach to fast automatic registration of partially overlapping range images. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1229–1243, 1999.

O. Chum and J. Matas. Matching with prosac - progressive sample consensus. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 220–226, 2005.

D. Claus and A. Fitzgibbon. A rational function lens distortion for more general cameras. *In Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR'2005)*, pages 213–219, 2005.

L. Clemente, A. Davison, I. D. Reid, J. Niera, and J. D. Tardos. Mapping large loops with a single hand-held camera. *In Proceedings of Robotics, Science and Systems Conference*, 2007.

D. M. Cole and P. M. Newman. Using laser range data for 3d SLAM in outdoor environments. *In Proceedings of IEEE International Conference on Robotics and Automation (ICRA'06)*, 2006.

D. M. Cole, A. R. Harrison, and P. M. Newman. Using naturally salient regions for SLAM with 3d laser data. *In IEEE International Conference on Robotics and Automation (ICRA'06) Workshop on SLAM*, 2005.

G. Conte, D. Scaradozzi, S. M. Zanoli, L. Gambetta, and G. Marani. Underwater SLAM with ICP localization and neural network object classification. *In Proceedings of the Eighteenth (2008) International Offshore and Polar Engineering Conference*, pages 351–359, 2008.

S. Coorg. Pose imagery and automated three-dimensional modeling of urban environments. *PhD thesis, Massachusetts Institute of Technology*, 1998.

D. Craciun, G. L. Besnerais, P. Fabiani, R. Mampey, and L. Macaire. Real-time vision-based odometry for safe landing of an unmanned aerial vechicle. *Master of Science Thesis, ONERA, DCSD-Toulouse, Universite de Sciences et Technologies de Lille 1*, 2006.

D. Craciun, N. Paparoditis, and F. Schmitt. Automatic pyramidal intensity-based laser scan matcher for 3d modeling of large scale unstructured environments. *In Proceedings of the Fifth IEEE Canadian Conference on Computer and Robots Vision*, pages 18–25, 2008.

D. Craciun, N. Paparoditis, and F. Schmitt. Automatic Gigapixel mosaicing in large scale unstructured underground environments. *In Tenth IAPR Conference on Machine Vision Application*, pages 13–16, 2009.

D. Craciun, N. Paparoditis, and F. Schmitt. Multi-view scans alignment for 3d spherical mosaicing in large scale unstructured environments. *Computer Vision and Image Understanding*, 2010.

G. Cross and A. Zisserman. Surface reconstruction from multiple views using apparent contours and surface texture. *In A. Leonardis and F. Solina and R. Bajcsy editors, NATO Advanced Research Workshop on Confluence on Computer Vision and Computer Graphics*, pages 25–47, 2000.

B. Curless and M. Levoy. A volumetric method for building complex models from range images. *In Proceedings of ACM Computer Graphics SIGGRAPH'96*, pages 303–312, 2000.

G. Dalley and P. Flynn. Pair-wise range image registration: A study in outlier rejection. *In Computer Vision and Image Understanding*, 87(1-3):104–115, 2002.

A. Danesi, D. Fontanelli, P. Murrieri, V. G. Scordio, and A. Bicchi. Cooperative visual SLAMS by homography. *In ROBOCUP2003*, 2003.

J. Davis. Mosaics of scenes with moving objects. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 98)*, pages 354–360, 1998.

A. Davison. Mobile robot navigation using active vision. *Ph. D. disseration, University of Oxford*, 1998.

A. Davison and D. W. Murray. Simultaneous localization and map-building using active vision. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7): 865–880, 2002.

A. Davison and D. W. Murray. Mobile robot localization and map-building using monocular active vision. *In Proceedings of Fifth European Conference on Computer Vision (ECCV'98)*, pages 809–825, 1998.

A. Davison, Y. G. Cid, and N. Kita. Real-time 3D SLAM with wide-angle vision. *In Proceedings of IFAC/EURON Symp. Intell. Auton. Vehicles*, 2004.

A. J. Davison. Real-time simultaneous localization and mapping with a single camera. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV'03)*, 2003.

A. J. Davison, I. Reid, N. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6): 1052–1067, June 2007.

M. Deans and M. Herbert. Experimental comparation techniques for localization and mapping using bearing only sensor. *In Proceedings of the ISER'00 Seventh International Symposium on Experimental Robotics*, 2000.

M. Deans, S. Moorehead, B. Shamah, and W. R. W. Kimberly Shillcutt. A concept for robotic lunar south pole exploration. *Proceedings of the Sixth International Conference and Exposition on Engineering, Construction, and Operations in Space (Space '98)*, pages 333–339, 1998.

M. C. Deans. Bearing-only localization and mapping. *Ph. D. thesis Carnegie Mellon University*, 2002.

M. Deveau, M. Pierrot-Deseilligny, N. Paparoditis, and X. Chen. Relative laser scanner and image pose. estimation from points and segments. *In International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 35, 2004.

P. Devebec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs; a hybrid geometry- and image- based approach. *In Proceedings of ACM SIGGRAPH*, pages 11–20, 1996.

P. E. Devebec and J. Malik. Recovering high dynamic range radiance maps from photographs. *In Proceedings of SIGGRAPH 97*, pages 369–378, 1997.

F. Devernay and O. Faugeras. Straght lines have to be straight. *In Proceeding of Machine Vision Applications*, 13(1):14–24, 2001.

P. Dias, V. Sequeira, F. Vaz, and J. Goncalves. Registration and Fusion of Intensity and Range Data for 3D Modelling of Real World Scenes. *In Proceedings of the Four IEEE International Conference on 3-D Digital Imaging and Modeling (3DIM'03)*, pages 418–425, 2003.

M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte, and M. Csorba. A solution to the simultaneous localization and map building (SLAM) problem. *In IEEE Transactions on Robotics and Automation*, 17(3):229–241, 2001.

C. Dorai, G.Wang, A. Jain, and C. Mercer. Registration and Integration of Multiple Objects Views for 3D Model Reconstruction. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):83–89, 1998.

T. Duckett and U. Nehmzow. Experiments in evidence-based localization for a mobile robot. *In Proceedings of AISB Workshop on Spatial Reasoning in Mobile Robots and Animals*, 1997.

R. Dupont, R. Keriven, and P. Fuchs. An improved calibration technique for coupled single row telemeter and CCD camera. *In Proceedings of the International Conference on 3-D Digital Imaging and Modeling*, pages 89–94, 2005.

H. Durrant-White and T. Bailey. Simultaneous localization and mapping: Part I. *In Proceeding of IEEE Robotics and Automation Magazine*, 13(2):99–110, 2006.

H. F. Durrant-Whyte. Sensor models and multi-sensor integration. *In International Journal of Robotics Research*, 7(6):97–113, 1988a.

H. F. Durrant-Whyte. Uncertain geometry in robotics. *In IEEE Transactions on Robotics and Automation*, 4(1):23–31, 1988b.

M. Eck, T. DeRose, T. Duchamp, H. Hoppe, M. Lounsbery, and W. Stuetzle. Multiresolution analysis of arbitrary meshes. *In R. Cook (ed) Proceedings of SIGGRAPH'95*, pages 173–182, 1995.

H. Edelbrunner. *Shape reconstruction with delaunay complex.* In A. V. Moura and C. L. Lucchesi (eds), LATIN'98: Theoretical Informatics. Thirs American Symposium, Campinas, Brazil, Lecture Notes in Computer Science, LNCS 1380. New York: Springer, pp. 119-132, 1998.

A. Eden, Uyttendaele, and R. Szeliski. Seamless image stitching of scenes with large motion and exposure differences. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 2498–2505, 2006.

D. Eggert, A. Fitzgibbon, and R. Fisher. Simultaneous registration of multiple range views for use in reverse engineering of cad models. *In Computer Vision and Image Understanding*, 69(3):253–272, 1998.

S. El-Hakim, J. Fryer, M. Picard, and E. Whiting. Digital recording of aboriginal rock art. *In Proceedings of the 10th International Conference on Virtual Systems and Multimedia (VSMM'04)*, pages 344–353, 2004.

S. El-Hakim, L. Gonzo, F. Voltolini, S. Girardi, A. Rizzi, and F. Remondino. Detailed 3d modeling of castles. *In International Journal of Architectural Computing*, 5(2):199–220, 2007.

S. F. El-Hakim, P. Boulanger, F. Blais, and J.-A. Beraldin. A system for indoor 3-Dmapping and virtual environments. *In Proceedings of Videometrics V*, 1997.

M. El-Melegy and A. Farag. Non-metric lens distortion calibration: Closed-form solution, robust estimation and model selection. *In Proceedings on International Conference on Computer Vision (ICCV'03)*, pages 554–559, 2003.

L. P. Ellekilde, J. V. Miro, and G. Dissanayake. Fusing range and intensity images for generating dense models of three-dimensional environments. *In Proceedings of IEEE International Conference on Man-Machine Systems*, pages 698–700, 2005.

T. Estlin, J. Yen, R. Petras, D. Mutz, R. Castao, G. Rabideau, R. Steele, A. Jain, S. Chien, E. Mjolsness, A. Gray, T. Mann, S. Hayati, and H. Das. An integrated architecture for cooperating rovers. *In Proceedings of the 1999 International Symposium on Art Intelligence, Robotics and Automation for Space*, pages 255–262, 1999.

C. Estrada, J. Neira, and J. D. Tardos. Hierarchical SLAM: Realtime accurate mapping of large environments. *In IEEE Transasctions on Robotics*, 21(4), 2005.

R. Eustice, H. Singh, J. Leonard, M. Walter, and R. Baillard. *Visually Navigating the RMS Titanic with SLAM information filter.* In Robotics: Science and Systems Cambridge MA: MIT Press, 2005.

M. Farid and A. C. Popescu. Blind removal of lens distortion. *Journal of Optical Society of America, Optics, Image Science and Vision*, 18(9):2072–2078, 2001.

O. Faugeras and M. Herbert. The representation, recognition and locating of 3D objects. *In International Journal of Robotic Research*, 5(3):27–52, 1986.

O. Faugeras and R. Keriven. Variational principles, surface evolution, pdes, level set method and stereo problem. *IEEE Transactions on Image Processing*, 7(3):336–344, 1998.

M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *In Communications of the ACM 24*, pages 381–395, 1981.

A. Fitzgibbon. Robust registration of 2d and 3d point sets. *In Image and Vision Computing Journal*, 21(4):1145–1153, 2003.

A. W. Fitzgibbon and A. Zisserman. Automatic 3d model acquisition and generation of new images from video sequences. *In Proceedings of European Signal Processing Conference (EUSIPCO'98)*, pages 1261–1269, 1998.

T. Foissotte, O. Stasse, A. Escande, and A. Kheddar. Next-best-view algorithm for autonomous 3d object modeling by a humanoid robot. *In Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, 2008a.

T. Foissotte, O. Stasse, A. Escande, and A. Kheddar. Towards a next-best-view algorithm for autonomous 3d object modeling by a humanoid robot. *In Proceedings of the 26-Annual Conference of the Robotics Society Japan (RSJ)*, 2008b.

W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

P. Fua and Y. G. Leclerc. Taking advantage of image-based and geometry-based constraints to recover 3d surfaces. *In Computer Vision and Image Understanding*, 64(1):111–127, 1996.

P. Fua and Y. Leclers. Object-centered surface reconstruction:combining multi-image stereo and shading. *International Journal of Computer Vision*, 16:35–56, 1995.

N. Garcias and J. Santos-Victor. Underwater video mosaics as visual navigation maps. *Journal of Computer Vision and Image Understanding*, 79:66–91, 2000.

M. A. Gennert. Brightness-based stereo-matching. *In Proceedings of 2nd International Conference on Computer Vision (ICCV'88)*, pages 139–143, 1988.

A. Georgiev and P. K. Allen. Localization methods for a mobile robot in urban environments. *In IEEE Transaction on Robotics and Automation (TRO)*, pages 851–864, 2004.

P. Giacon, S. Saggin, G. Tommasi, and M. Busti. Implementing dsp algorithms using spartan-3 fpgas. *DSP Magazine*, pages 16–19, 2005.

GigaPan. Gigapan. *http://gigapan.org*, 2009.

G. Giralt and L. Boissier. The French Planetary Rover VAP: Concept and Current Developments. *In Proceedings on IEEE/RSJ International Conf. on Intelligent Robots and Systems*, pages 1391–1398, 1992.

S. B. Goldberg, M. W. Maimone, and L. Matthies. Stereo vision and rover navigation software for planetary exploration. *In IEEE Aerospace Conference Proceedings*, 5:2025–2036, 2002.

J. Gomes. Level sets and distance functions. *In Proceedings of the 6th European Conference on Computer Vision (ECCV'00)*, pages 588–602, 2000.

M. Gopi, S. Krishnan, and C. T. Silva. Surface reconstruction based on lower dimensional localized delaunay triagulation. *In Proceedings of Eurographics 2000*, pages 467–478, 2000.

S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. *In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques (SIGGRAPH '96)*, pages 43–54, 1996.

V. M. Govindu. Revisiting the birghtness constraint: probabilistic formulation and algorithms. *In Proceedings on European Conference on Computer Vision (ECCV'96)*, pages 177–188, 2006.

N. Greener. Environment mapping and other applications of world projections. *In IEEE Conference on Computer Graphics and Applications*, 6(11):21–29, 1986.

M. Greenspan and G. Godin. A nearest neighbour method for efficient icp. *In Proceedings of the 3th International Conference on 3-D digital Imaging and Modeling*, pages 161–168, 2001.

A. D. Griffiths, A. J. Coates, R. Jaumann, H. Michaelis, G. Paar, D. Barnes, J.-L. Josset, and the PanCam team. Context for the ESA ExoMars Rover: the Panoramic Camera (PanCam) Instrument. *In International Journal of Astrobiology*, 5(3):269–275, 2006.

E. Guivant and E. M. Nebot. Optimization of the simultaneous localization and map-building for real-time implementation. *In IEEE Transactions on Robotics and Automation*, 17(3):242–257, 2001.

J. Gutmann and K. Konolige. Incremental mapping for large cyclic environments. *In Proceedings of IEEE International Symposium on Comput. Intell. Robot. Autom.*, pages 318–325, 1999.

G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry illumination. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, 1998.

D. W. Hainsworth. Teleoperation user interfaces for mining robotics. *Autonomous Robots, Kluwer Academic Publishers*, 11:19–28, 2001.

M. Hansen, P. Anandan, K. Dana, van der Wal C., and P. Burt. Real-time scene stabilization and mosaic construction. *In IEEE Workshop on Applications of Computer Vision (WACV'94)*, pages 54–62, 1994.

C. Harris and M. Stephens. A combined corner and edge detector. *In Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988.

C. Harris and M. Stephens. A combined corner and edge detector. *In Proceedings of Fourth Alvey Vision Conference*, pages 147–151, 1998.

A. Harrison and P. Newman. Image and sparse laser fusion for dense scene reconstruction. *In Proceedings of the International Conference on Field and Service Robotics (FSR)*, pages 214–219, 2009.

R. I. Hartley. Self-calibration from multiple views of a rotating camera. *In Proceedings of Third European Conference on Computer Vision (ECCVŠ94)*, pages 471–478, 1994.

R. I. Hartley and A. Zisserman. *Multiple View Geometry*. Cambridge University Press, Cambridge, UK, 2000.

R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Second edition, 2004.

M. Hebert, C. Caillas, E. Krotknov, I. Kweon, and T. Kanade. Terrain mapping for rover planetary explorer. *In IEEE International Conference on Robotics and Automation (ICRA)*, 2:997–1002, 1989.

P. Heckbert. Fundamentals of texture mapping and image warping. *Master Thesis, University of California at Berkeley*, 1989.

D. J. Heeger and A. D. Jepson. Subspace methods for recognizing rigid motion i: Algorithm and implementation. *In International Journal of Computer Vision*, 7(2):95–117, 1992.

J. Heikkila. Geometric camera using singular central points. *In IEEE Transactions on Patter Analysis and Machine Intelligence*, 22(10):1066–1077, 2000.

C. E. Hernandez. Silouette and stereo fusion for 3d objet modeling. *PhD Dissertation ENST Telecom ParisTech*, 2004.

A. Hilton, A. Stoddart, J. Illingworth, and T. Windeatt. Reliable surface reconstruction from multiple range images. *In IEEE Sixth International Conference on Computer Vision (ICCV'96)*, pages 117–126, 1996.

R. Horaud and O. Monga. *Vision par Ordinateur - Outils Fundamentaux*. Hermes, II edition, 1995.

B. Horn and M. Brooks. *Shape from Shading*. The MIT Press, 1989.

B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *In Journal of the Optical Society of America*, 4(4):629–642, 1987.

A. M. Howard and e. E. W. Tunstel. *Intelligence for Space Robotics*. TSI Press, 2006.

T. D. Howell and J.-C. Lafon. The complexity of the quaternion product. *Technical Report TR 75-245; Departement of Computer Science, Cornell University, Ithaca, N. Y.*, 1975.

D. Huber. Automatic Three-dimensional Modeling from Reality. *Ph. D. thesis, Robotics Institute, Carnegie Mellon University,Pittsburgh, PA*, Decembre 2002.

D. Huber and M. Herbert. A New Approach to 3D Terrain Mapping. *In IEEE/RSJ International Conference on Intelligent Robotics and Systems (IROS99)*, pages 1121–1127, Octobre .

D. Huber and N. Vandapel. Automatic 3d underground mine mapping. *The 4th International Conference on Field and Service Robotics*, 2003a.

D. Huber and N. Vandapel. Automatic 3D underground mine mapping. *In 4th International Conference on Field and Service Robotics*, 2003b.

P. J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

L. Ikemoto, N. Gelfand, and M. Levoy. A hierarchical method for aligning warped meshes. *In Proceedings of the 4th International Conference on 3-D digital Imaging and Modeling*, 2003.

K. Ikeuchi and Y. Sato. *Modeling from Reality*. Kluwer Academic Publisher, 2001.

K. Ikeuchi, T. Oishi, J. Takamatsu, R. Sagawa, A. Nakazawa, R. Kurazume, K. Nishino, M. Kamakura, and Y. Okamoto. The Great Buddha Project: Digitally Archiving, Restoring, and Analyzing Cultural Heritage Objects. *In International Journal of Computer Vision*, 75(1):189–208, 2007.

M. Irani and P. Anandan. Video indexing based on mosaic representation. *In Proceedings of the IEEE*, 86(5):905–921, 1998.

M. Irani and S. Peleg. Improving resolution by image registration. *Graphical Models and Image Processing*, 53(3):231–239, 1991.

M. Irani, P. Anandan, and S. Hsu. Mosaic-based representations of video sequences and their applications. *In Proceedings of Fifth International Conference on Computer Vision (ICCV'95)*, pages 605–611, 1995a.

M. Irani, S. Hsu, and P. Anandan. Video compressing using mosaic representations. *In Proceedings of Signal Processing: Image Communication*, pages 529–552, 1995b.

J. Isidoro and S. Sclaroff. Stochastic refinement of the visual hull to satisfy photometric and silhouette consistency constraints. *In Proceedings of International Conference on Computer Vision*, pages 1335–1342, 2003.

iTowns. itowns. *http://www.itowns.fr/*, 2008.

P. Jasiobedzki and R. Jakola. From space robotics to underwater mining. *In Underwater Mining Institute 2007*, 2007.

J. Jia and C. K. Tang. Image registration with global and local luminance alignment. *In Proceedings of 9th International Conference on Computer Vision (ICCV'03)*, pages 156–163, 2003.

H. Jin, A. Yezzi, and S. Soatto. Stereoscopic shading: integrating shape cues in a variational framework. *In Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 169–176, 2000.

A. Johnson. Spin-images: A representation for 3-d surface matching. *PhD thesis, Robotics Institute, Carnegie Mellon University*, 1997.

A. Johnson and M. Herbert. Using spin images for efficient object recognition in cluttered 3d scenes. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5): 433–449, 1999.

A. Johnson, R. Willson, Y. Cheng, J. Goguen, C. Leger, M. Sanmartin, and L. Matthies. Design through operation of an image-based velocity estimation system for Mars landing. *In International Journal of Computer Vision*, 74(3):319–341, 2007.

I. Jung and S. Lacroix. High resolution terrain mapping using low altitude aerial stereo imagery. *In Proceedings of IEEE Ninth International Conference on Computer Vision (ICCV'03)*, 2003.

F. Jurie and M. Dhome. Hyperplane approximation for template matching. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):996–1000, 2002.

P. C. Juyang Weng and M. Herniou. Camera calibration with distortion models and accuracy evaluation. *In IEEE Transactions Pattern Analysis and Machine Intelligence*, 14(10):965–980, 1992.

T. Kanade, H. Kato, S. Kimura, A. Yoshida, and K. O. D. of a video-rate stereo machine. Development of a video-rate stereo machine. *In Proceeding of the IEEE/RSJ International Conference on Intelligent Robotics and Systems (IROS)*, 3:95–100, 1995.

S. B. Kang. Radial distortion snakes. *In Proceedings on IEICE Transactions on Information & Systems*, pages 1603–1611, 2001.

R. Katz, N. Melkumyan, J. Guivant, T. Bailey, J. Nieto, and E. Nebot. Integrated sensing framework for 3d mapping in outdoor navigation. *In Proceedings of IEEE International Conference on Intelligent Robots and Systems (IROS'06)*, 2006.

Y. Ke and Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pages 506–513, 2004.

A. Kim and R. Eustice. Pose-graph visual graph with geometric model selection for autonomous underwater ship hull inspection. *In Proceedings of IEEE International Conference on Robotics and Intelligent Systems (IROS'09)*, 2009.

J. Kim, V. Kolmogorov, and R. Zabih. Visual correspondance using energy minimization and mutual information. *In Proceedings of Ninth International Conference on Computer Vision (ICCV'03)*, pages 1033–1040, 2003.

J. H. Kim and S. Sukkarieh. Airborne simultaneous localization and map building. *In Proceedings of the IEEE International Conference on Robotics and Automation*, January 2003.

S. Kim, H. Kim, and T.-K. Yang. Increasing SLAM performances by integrated grid and topology map. *In Journal of Computers*, 4(7):601–609, 2009.

S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):617–680, 1983.

K. Klein and V. Sequeira. View planning for the 3d modelling of real world scenes. *In Proceedings of the IEEE International Conference on Robotics and Intelligent Systems (IROS'00)*, pages 943–948, November 2000.

Kolor. Autopano pro. *http://www.autopano.net/en/*, 2005.

Kolor. Autopano giga. *http://www.autopano.net/en/photo-stitching-solutions/autopano-giga.html*, 2009.

K. Konolige. Large-scale map-making. *In Proceedings of National Conference on Artificial Intelligence (AAAI)*, pages 457–463, 2004.

K. Konolige. Constraint-maps: A general least squares method for SLAM. *Submitted for publication*, 2005.

K. Konolige and J. S. Gutmann. Incremental mapping of large cyclic environments. *In Proceedins of IEEE International Conference on Robotics and Automation*, 1999.

KrPano. Krpano.com. *http://www.krpano.com/*, 2009.

B. Kuipers and Y. T. Byun. A robot exploration and mapping strategy based on semantic hierarchy of spatial representations. *In International Journal on Robotics and Autonomous Systems*, 8(1-2):47–63, 1991.

C. Kunz, C. Murphy, H. Singh, S. Singh, T. Sato, C. Roman, and K. Nakamura. Toward extraplanetary under-ice exploration: Robotic steps in the arctic. *In Journal of Field Robotics*, 1(19):1–19, 2009.

K. Kutulakos and S. Seitz. A theory of shape by space carving. *In International Journal of Computer Vision*, 3(38):199–219, 2000.

N. M. Kwok and G. Dissanayake. An efficient multiple hypothesis filter for bearing-only SLAM. *In Proceedings of International Conference on Intelligent Robots and Systems (IROS'04)*, 2004.

N. M. Kwok, G. Dissanayake, and Q. P. Ha. Bearing-only SLAM using a sprt based gaussian sum filter. *In Proceedings of the IEEE International Conference on Robots and Automation (ICRA'05)*, 2005.

F. Labrosse. Visual compass. *In Proceedings of Toward Autonomous Robotic Systems*, 2004.

A. Laurentini. The visual hull concept for silhouette based image understanding. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(16):150–162, 1994.

C. Lee. The radial undistortion and calibration on an image array. *Ph. D. thesis, MIT*, 2000.

M.-C. Lee, W.-G. Chen, C. Lin, C. Gu, T. Markoc, S. Zabinsky, and R. Szeliski. A layered video object coding system using sprite and affine motion model. *In IEEE Transactions on Circuits and Systems for Video Technology*, 7(1):130–145, 1997.

W. Y. Lee. Spatial semantic hierarchy for a physical robot. *Ph. D. Thesis, University of Texas at Austin*, 1996.

T. Lemaire. Robotique mobile: Localization et cartographie simultanées. *In Colloque Ecole Doctorale Informatique et Telecommunications (EDIT)*, 2004.

T. Lemaire and S. Lacroix. SLAM with panoramic vision. *In IEEE International Journal on Field Robotics*, 24(1):91–111, 2007.

T. Lemaire, S. Lacroix, and J. Sola. A practical 3d bearing only SLAM algorithm. *In Proceedings of the IEEE International Conference on Intelligent Robotos (IROS'05)*, 2005.

T. Lemaire, C. Berger, and I.-K. J. S. Lacroix. Vision-based SLAM: stereo and monocular approaches. *In IEEE International Journal on Computer Vision*, 74(3):343–363, 2007.

R. K. Lenz and R. Y. Tsai. Techniques for calibration of the scale factor and image center for high-accuracy 3-d machine vision metrology. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5):713–720, 1988.

R. K. Lenz and R. Y. Tsai. Implicit and explicit camera calibration: theory and experiments. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5): 469–480, 1994.

M. Levoy and P. Hanrahan. Light-field rendering. *In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques (SIGGRAPH '96)*, 1996.

M. Levoy, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. Anderson, J. Davis, J. Ginsberg, J. Shade, and D. Fulk. The Digital Michelangelo Project: 3D Scanning of Large Statues. *In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 131–144, July 2000.

H. Li, M. Yang, and H. Qian. Camera and laser scanner co-detection of pedestrians. *In Proceedings of the International Conference on Robotics and Automation - Workshop on Safe Navigation in Open and Dynamic Environments*, 2009.

R. Li, K. Di, J. Wang, S. He, A. Howard, and L. Matthies. Rock modeling and matching for autonomous mars rover localization. *In Annual NASA Science Technology Conference*, 2007.

C. E. Liedtke, H. Busch, and R. Koch. Shape adaptation for 3d modeling of 3d objects in natural scenes. *In Proceedings of Conference on Computer Vision and Pattern Recognition*, 40(3):704–705, 1991.

Z.-C. Lin, T. S. Huang, S. D. Blostein, H. Lee, and E. A. Margerum. Motion estimation from 3-d point sets with and without correspondances. *In Proceedings on IEEE Computer Vision and Pattern Recognition Conference*, pages 194–201, 1986.

Y. Liu and M. Rodrigues. Geometrical analysis of two sets of 3d correspondance data patterns for the registration of free-form shapes. *In Journal of Intelligent and Robotic Systems*, 33(4):409–436, 2002.

Y. Liu, R. Emery, D. Chakrabarti, W. Burgard, and S. Thrun. Using EM to learn 3D models with mobile robots. *In Proceedings of the International Conference on Machine Learning (ICML)*, pages 329–336, 2001.

M. B. Lopez, J. Hannuksela, O. Silven, and M. Vehvilanen. Graphics hardware accelerated panorama builder for mobile phone. *In Proceedings of SPIE Conference on Multimedia on Mobile Devices 2009*, 2009.

W. Lorensen and H. Cline. Marching cubes: a high resolution 3d surface contruction algorithm. *In International Journal on Computer Graphics*, 21(4):163–170, 1987.

M. I. A. Lourakis and A. A. Argyros. The design and implementation of a generic sparse bundle adjustement software package based on levenberg-marquardt algorithm. *Technical Report - 340*, 2004.

K.-L. Low. View planning for range acquisition of indoor environments. *Ph. D. Dissertation, Unversity of North Carolina at Chapel Hill*, 2006.

D. Lowe. Distictive image features from scale invariant keypoints. *International Journal of Computer Vision*, 60(1):91–110, 2004.

D. Lowe. Autostitch. *http://people.cs.ubc.ca/ mbrown/autostitch/autostitch.html*, 2007.

F. Lu and E. Milios. Globally consistent range scan alignment for environment mapping. *In International Journal on Robotics and Autonomous Systems*, 5:333–349, 1997.

B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *In Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

L. Lucchese, G. Doretto, and G. Cortelazzo. A frequency domain technique for range data registration. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 (11):1468–1484, 2002.

M. Magnusson and T. Duckett. A comparation of 3D registration algorithms for autonomous underground mining vehicles. *In Proceedings of the Second European Conference on Mobile Robotics (ECMR'05)*, pages 86–91, 2005.

M. Magnusson and T. Duckett. Scan registration for autonomous mining vehicles using 3D-NDT. *In International Journal of Field Robotics*, pages 803–827, 2007.

I. Mahon and S. Williams. Three dimensional robot mapping. *In Proceedings of Australian Conference on Robotics and Automation*, 2003.

A. Makadia, A. Patterson, , and K. Daniilidis. Fully automatic registration of 3d point clouds. *In Proceedings of Compute Vision and Pattern Recognition CVPR'06*, pages 1297–1304, 2006.

A. Mallet, S. Lacroix, and L. Gallo. Position estimation in outdoor environments using pixel tracking and stereovision. *In Proceedings of IEEE Conference on Robotics and Automation - ICRA'00*, 2000.

T. K. Man and S. Kwong. Genetic algorithms: concepts and applications. *In IEEE Transactions on Industrial Electronics*, 43(5):519–534, 1996.

S. Mann and R. Picard. Virtual bellows: constructing high-quality images from video. *In First IEEE International Conference on Image Processing (ICIP'94)*, pages 363–367, 1994.

S. Mann and R. W. Picard. On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. *In IS & T 48th Anual Conference*, pages 422–428, 1995.

J. Martin and J. Crowley. Comparation of correlation techniques. *In Proceedings of International Conference on Intelligent Autonomous Systems*, pages 86–93, March 1995.

T. Masuda. Registration and integration of multiple range images by matching signed distance fields for object shape modeling. *In Computer Vision and Image Understanding*, 87(1-3):51–65, 2002.

T. Masuda and N. Yokoya. A robust method for registration and segmentation of multiple range images. *In Computer Vision and Image Understanding*, 61(3):295–307, 1995.

M. Mataric. Integration and representation into goal-driven behaviour-based robots. *In IEEE Transactions on Robotics and Automation*, 8(3):304–312, 1997.

L. Mathies, M. Maimone, A. Johnson, Y. Chieng, R. Willson, C. Villalpando, S. Goldberg, and A. Huertas. Computer vision on Mars. *In International Journal of Computer Vision*, 75(1):67–92, 2007.

Y. Matsumoto, K. Fujimura, and T. Kitamura. A portable three-dimensional digitizer. *In International Conference on Recent and Advances on 3D Imaging and Modeling*, pages 197–205, 1997.

L. Matthies and S. Shafer. Error modeling in stereo navigation. *In IEEE International Journal of Robotics and Automation*, 3:239–250, 1987.

W. Matusik. The visual hull concept for silhouette based image understanding. *In Proceedings of ACM SIGGRAPH*, pages 369–374, 2000.

L. McMillan and G. Bishop. Plenoptic modeling: An image-based rendering system. *In SIGGRAPH'95*, 1995.

J. Meehan. *Panoramic Photography*. Watson-Guptill, 1990.

C. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *In International Journal of Computer Vision*, (1):63–86, 2004.

K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

D. L. Milgram. Adaptive techniques for photomosaicking. *In IEEE Tranactions on Computers*, 26(11):1175–1180, 2006.

J. V. Miro and G. Dissanayake. Robotic 3D visual mapping for augumented situational awareness in unstructured environments. *In Proceedings of the International Workshop on Robotics for Risky Interventions and Surveillance of the Environment (RISE 2008)*, pages 1–13, 2008.

MIT. Mit city scanning project. *http://city.csail.mit.edu/*, 2000.

H. B. Mitchell. *Multi-sensor Data Fusion An Introduction*. Springer, 2007.

N. J. Mitra, N. Gelfand, H. Pottmann, and L. J. Guibas. Registration of point cloud data from a geometric optimization perspective. *In Symposium of Geometry Processing*, pages 23–32, 2004.

T. Mitsunaga and S. K. Nayar. Radiometric self calibration. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'99)*, pages 374–380, 1999.

J. Miura and S. Ikeda. A simple modeling of complex environments for mobile robots. *In International Journal of Systems Technologies and Applications*, 6(1-2):166–177, 2009.

F. H. Moffit and E. M. Mikhail. *Photogrammetry*. Harper & Row, New York, 3rd edition, 1980.

M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. Fast-SLAM: A factored solution to the simultaneous localization and mapping problem. *In Proceedings of American Association on Artificial Intelligence (AAAI'02)*, pages 593–598, 2002.

J. M. M. Montiel, J. Civera, and A. J. Davison. Unified depth inverse parametrization for monocular SLAM. *In Proceedings of Robotics, Science and Systems Conference*, 2006.

H. Moravec. The stanford cart and the cmu rover. *In Proceedings of the IEEE*, 71(7): 872–884, 1983.

H. P. Moravec. Obstacle avoidance and navigation in the real world by a seeing robot rover. *Ph. D. thesis, Stanford University, Stanford, California*, May 1980.

J. J. Moré. *Numerical Analysis*, volume 630. Springer Berlin, 2006.

R. A. Moreno and A. S. de Miguel. A machine consciousness approach to autonomous mobile robotics. *In 5th International Cognitive Robotics Workshop (AAAI-06)*, 2006.

R. A. Moreno and A. S. de Miguel. Applying machine consciousness models in autonomous situated agents. *In Pattern Recognition Letters archive*, 29(8):1033–1038, 2008.

E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Monocular vision for mobile robot localization and autonomous navigation. *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:363–370, 2006.

A. C. Murillo, J. J. Guerrero, and C. Sagues. Surf features for efficient robot localization with omnidirectional images. *In Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3901–3907, 2007.

R. M. Murray. Recent research in cooperative control of multi-vehicle systems. *In International Conference on Advances in Control and Optimization of Dynamical Systems*, pages 1–27, 2007.

S. Nayar. Catadioptric omnidirection camera. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pages 482–488, 1997.

A. Nüchter, H. Surmann, K. Lingemann, J. Hertzberg, and S. Thrun. 6D SLAM with application in autonomous mine mapping. *In Proceedings IEEE 2004 International Conference Robotics and Automation (ICRA 2004)*, 2004.

P. Newman, D. Cole, and K. Ho. Outdoor SLAM using visual appearance and laser ranging. *In Proceedings on International Conference on Robotics and Automation*, 2006.

F. Nielsen and N. Yamashita. Clairvoyance: A fast and robust precision mosaicing system for gigapixel images. *In IEEE 32 Anual Conference on Industrial Electronics IECON'06*, pages 3471–3476, 2006.

D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 2161–2168, 2006.

D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. *In Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, pages 652–659, 2004.

A. Nuchter, H. Surmann, and S. Thrun. 6D SLAM with an application in autonomous mine mapping. *In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'04)*, 2004.

A. Nuchter, K. Lingemann, J. Hertzberg, H. Surmann, K. Pervolz, and M. Henning. Mapping of rescue environments with kurt3d. *In Proceedings of the International Workshop on Safety, Security and Rescue Robotics (SSRR '05), Kobe, Japan*, 2005.

C. Olson, L. Matthies, M. Schoppers, and M. Maimone. Robust stereo ego-motion for long distance navigation. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

E. Olson, J. Leonard, and S. Teller. Fast iterative alignment of pose graphs with poor initial estimates. *In Proceedings of IEEE International Conference on Robotics and Automation (ICRA'06)*, 2006.

S. Ono and K. Ikeuchi. Development of ladder-type laser scanning system for 3-D modeling of vertical and narrow areas by space-time analysis. *In Proceedings of IAPR Conference on Machine Vision Applications (MVA'07)*, 2007.

G. Paar, A. Griffiths, A. Bauer, T. Nunner, N. Schmitz, D. Barnes, and E. Riegler. 3D vision ground processing workflow for the panoramic camera on ESA's ExoMars mission 2016. *In Proceedings of ISPRS Optical 3-D Measurement Techniques IX*, 2:161–170, 2009.

A. Pantland and S. E. Sclaroff. Close form solutions for physically based shape modeling and recovery. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13 (7):715–729, 1991.

P. Perez, M. Gangnet, and A. Blacke. Poisson image editing. *In ACM Transactions on Computer Graphics*, 22(3):313–318, 2003.

A. G. E. G. Peter K. Allen, Ioannis Stamos and P. Blaer. Avenue: Automated site modeling in urban environments. *In Proceedings of 3rd International Conference on Digital Imaging and Modeling*, 2001.

J. Peters. Constructing c^1 surfaces for arbitrary topology using biquadratic and bicubic splines. *In N. Sapidis (ed) Designing Fair Curves and Surfaces SIAM*, pages 277–293, 1994.

Y. Petillot, J. Salvi, and B. Batlle. 3d large-scale seabed reconstruction for uuv simultaneous localization and mapping. *In Proceedings on Guidance and Control of Underwater Vehicles (NGCUVŠ08)*, 2008.

V. K. Pilania and D. Chakravarty. Application of wireless visual sensor for semi-autonomous mine navigation systems. *In World Academy of Science and Technology*, (45):437–441, 2008.

R. L. Plackett. The discovery of the method of least squares. *Biometrika*, 59:239–251, 1972.

M. Pollefeys, R. Koch, M. Vergauwen, and L. van Gool. Metric 3d surface reconstruction from uncalibrated image sequences. *In Reinhard Koch and Luc van Gool, editors, 3D Structure from Multiple Images of Large Scale Environments European Workshop, SMILE'98*, pages 139–154, June 1998.

M. Pollefeys, D. Nister, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewenius, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3d reconstruction from video. *In International Journal of Computer Vision*, 78(2-3):143–167, 2008.

D. Poussart and D. Laurendeau. *Advances in Computer Vision, chapter 3D Snesing for Industrial Computer Vision*. Springer-Verlag, 1989.

K. Pulli. Multiview registration for large data sets. *In Proceedings of the 2nd International Conference on 3D Imaging and Modeling (3DIM'99)*, pages 160–168, 1999.

K. Pulli, H. Abi-Rached, T. Duchamp, L. G. Shapiro, and W. Stuetzle. Vizualizing real objects from scanned range and color data. *In 8th Eurographics Workshop on rendering*, 1997.

Z. Qi and J. R. Cooperstock. Overcoming parallax and sampling density issues in image mosaicing of non-planar scenes. *In Proceedings of British Machine Vision Conference (BMVC'07)*, 2007.

Z. Qi and J. R. Cooperstock. Depth-based image mosaicing for both static and dynamic scenes. *In Proceedings of International Conference on Pattern Recognition (ICPR'07)*, 2008.

L. H. Quam. Hierarchical warp stereo. *In Image Understanding Workshop*, pages 149–155, 1984.

P. Rademacher and G. Bishop. Multiple-center-of-projection images. *In SIGGRAPH'98*, 1998.

J. C. Ralston, C. O. Hargrave, and D. W. Hainsworth. Localisation of mobile underground mining equipment using wireless ethernet. *In Proceedings of the International Conference on Machine Learning (ICML)*, pages 225–230, 2005.

A. Ramisa, D. Aldavert, and R. Toledo. A panorama based localization system. *CVC Research and Development Workshop*, pages 36–42, 2006.

P. Rander. A multi-camera method for 3d digitization of dynamic, real-world events. *PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsurgh, PA*, May 1998.

M. Reed and K. Allen. 3D Modeling from Range Imagery: An Incremental Method with a Planning Component. *In Journal on Image and Vision Computing*, 17(2):99–111, 1999.

M. K. Reed and P. K. Allen. Constraint-based Sensor Planning for Scene Modeling. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1460–1467, 2000.

E. Reinhard, G. Ward, S. Pattanaik, and P. Devebec. High dynamic range imaging: Acquisition, display and image-based lighting. *Morgan Kaufmann*, 2005.

I. Rekleitis, J.-L. Bedwani, and E. Dupuis. Autonomous planetary exploration using lidar data. *In Proceedings of IEEE International Conference on Robotics and Automation*, pages 3025–3030, 2009.

A. Reshetov, A. Soupikov, and J. Hurley. Multi-level ray tracing algorithm. *In ACM Transactions on Graphics Proceeding of ACM SIGGRAPH*, 24(3):1176–1185, 2005.

R. Reulke, A. Wehr, and D. Griesbach. High resolution mapping using CCD-line camera and laser-scanner with integrated position and orientation system. *In Proceedings of International Society of Photogrammetry and Remote Sensing*, pages 322–329, 2004.

C. Robertson and R. Fisher. Parallel evolutionary registration of range data. *In Computer Vision and Image Understanding*, 87(1-3):39–50, 2002.

M. S. Robison and K. Edwards. A new digital image mosaic of mercury. *In Twenty-sixth Lunar and Planetary Science Conference*, 1995.

R. Rocha, J. Dias, and A. Carvalho. Cooperative multi-robot systems a study of vision-based 3-d mapping using information theory. *In Proceedings of IEEE International Conference on Robotics and Automation (ICRA'05)*, pages 384–389, 2005.

M. Rodrigues, R. Fisher, and Y. Liu. Special issue on registration and fusion of range images. *In Computer Vision and Image Understanding*, 87(1-3):1–7, 2002.

S. A. Rodrigues, V. Frémont, and P. Bonnifait. Extrinsic calibration of a multi-layer lidar and a camera. *In Proceedings of IEEE Multi-sensor Fusion and Integration for Intelligent Systems*, pages 214–219, 2008.

P. J. Rousseeuw. Least median of squares regression. *In Journal of the American Statistical Association*, (79):871–880, 1984.

E. Royer, M. Lhuillier, and M. D. J.-M. Lavest. Monocular vision for mobile robot localization and autonomous navigation. *International Journal of Computer Vision*, 74(3): 237–260, 2007.

Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. *In Proceedings on IEEE International Conference on Computer Vision*, 1998.

S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. *In Proceedings of the 3th Internatinal Conference on 3-D Digital Imaging and Modeling*, 1:145–152, 2001.

Y. Rzhanov, L. M. Linnett, and R. Forbes. Underwater video mosaicing for seabed mapping. *In Proceedings of International Conference on Image Processing*, pages 224–227, 2000.

J. Sàez, A. Hogue, F. Escolano, and M. Jenkin. Registration and fusion of intensity and range data for 3d modelling of real world scenes. *In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'06)*, pages 3562–3567, 2006.

H. Saito, S. Baba, M. Kimura, S. Vedula, and T. Kanade. Appereance-based virtual new generation of temporally-varying events from multi-camera images in the 3d room. *In Proceedings of the International Conference on 3D Digital Imaging and Modeling*, pages 516–525, 1999.

H. Sakai, T. Tanaka, T. Mori, S. Ohata, K. Ishii, and T. Ura. Underwater video mosaicing using auv and its application to vehicle navigation. *2004 International Symposium on Underwater Technology*, pages 405–410, 2004.

E. Salamin. Application of quaternions to computation with rotations. *Technical Report, Standford, AI Lab*, 1979.

H. Samet. *The design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading, Massachusetts, 1989.

A. Sanderson. A distributed algorithm for cooperative navigation among multiple robots. *In Journal on Advanced Robotics*, 12:335–349, 1998.

A. Sappa, A. Restrepo-Specht, and Y. Chen. Range image registration by using an edge-map representation. *In International Symposium on Intelligent Robotic Systems*, pages 167–176, 2001.

A. Sarti and S. Tubaro. Image based multiresolution implicit object modeling. *EURASIP Journal on Applied Signal Processing*, 40(3):1053–1066, 2002.

H. Sawhney, S. Hsu, and R. Kumar. Robust video mosaicing through topology inference and local to global alignment. *In Proceedings of 5th European Conference on Computer Vision*, 2:103–119, 1998.

H. S. Sawhney and S. Ayer. Compact representation of video thourgh dominant multiple motion estimation. *In IEEE Transactions on Pattern Recognition and Machine Intelligence*, 18(8):814–830, 1996.

H. S. Sawhney and R. Kumar. True multi-image alignment and its application to mosaicing and lens distortion correction. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(3):235Ű243, 1999.

D. Scaramuzza, A. Harati, and R. Siegwart. Extrinsic self calibration of a camera and a 3d laser range finder from natural scenes. *In Proceedings of International Conference on Intelligent Robots and Systems (IROS'07)*, pages 4164–4169, 2007.

F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image step, or how do i organize my holidays snaps? *In Seventh European Conference on Computer Vision (ECCV'02)*, pages 414–431, 2002.

C. Schmid, R. Mohr, and C. Bauckhage. Camparing and evaluating interest points. *In Proceedings of the IEEE International Conference on Computer Vision*, January 1998.

C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *In International Journal of Computer Vision*, 37(2):151–172, 2000.

S. Se, D. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *In International Journal of Robotics Research*, 21(8), 2002.

V. Sequeira, K. Ng, E. Wolfart, J. G. M. Gonçalves, and D. Hogg. Automated reconstruction of 3d models from real environments. *In ISPRS Journal of Photogrammetry and Remote Sensing*, 54:1–22, 1999.

M. Sezgin and B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *In Journal on Electronic Imaging*, 13(1):146–165, 2004.

J. Shade, J. S. Gortler, L.-W. he, and R. Szeliski. Layered depth images. *In SIGGRAPH'98*, 1998.

G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. *In Proceedings of Ninth International Conference on Computer Vision (ICCV'03)*, pages 750–757, 2003.

G. Sharp, S. Lee, and D. Wehe. ICP Registration Using Invariant Features. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):90–102, 2002.

J. Shi and C. Tomasi. Good features to track. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994a.

J. Shi and C. Tomasi. Good features to track. *In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994b.

H. Shum, M. H. K. Ikeuchi, and R. Reddy. An integral approach to free-form object modeling. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(12): 1366–1370, 1997.

H. Y. Shum and R. Szeliski. Construction of panoramic mosaics with global and local alignment. *In International Journal of Computer Vision*, 36(2):101Ű130, 2000.

L. Silva, O. Bellon, K. Boyer, and P. Gotardo. Low-overlap range image registration for archeological applications. *In Proceedings of IEEE/CVPR Workshop on Applications of Computer Vision in Archeology*, 2003.

L. Silva, O. R. P. Bellon, and K. L. Boyer. *Robust Range Image Registration using Genetic Algorithms and the Surface Interpenetration Measure*. World Scientific Co Pte Ltd, Series in Machine Perception and Artificial Intelligence, vol. 60, 2005.

E. P. Simoncelli, E. H. Adelson, and D. J. Heeger. Probability distribution of the optical flow. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recongnition (CVPR'91)*, pages 310–315, 1991.

C. C. Slama. Manual of photogrammetry. *American Society of Photogrammetry, Falls Church, Virginia, fourth edition*, 1980.

R. Smith and P. Cheesman. On the representation of spatial uncertainty. *In International Journal of Robotics Research*, 5(4):56–68, 1987.

N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. *In Proceedings of ACM SIGGRAPH'06*, 2006.

S. Soatto, A. J. Yezzi, and H. Jin. Tales of shape and radiance in multi-view stereo. *In Proceedings ICCV*, pages 974–981, 2003.

J. Sola, M. Devy, A. Monin, and T. Lemaire. Undelayed initialization in bearing only SLAM. *In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS'05)*, 2005.

M. Soucy and D. Laurendeau. A general surface approach to the integration of a set of range views. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17 (4):344–358, 1995.

I. Stamos. Geometry and texture recovery of scenes of large scale: Integration of range and intensity sensing. *PhD Dissertation Columbia University*, 2001.

I. Stamos and M. Leordeanu. Automated Feature-based Range Registration of Urban Scenes of Large Scale. *In Proceedings of International Conference on Computer Vision and Pattern Recognition 2003*, pages 555–561, 2003.

I. Stamos, L. Liu, C. Chen, G. Wolberg, G. Yu, and S. Zokai. Integrating Automated Range Registration with Multiview Geometry for the Photorealistic Modeling of Large-Scale Scenes. *In International Journal of Computer Vision*, 78(2-3):237–260, 2008.

F. Stein and G. Medioni. Structural indexing: Efficient 3-d object recognition. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):125–145, 1992.

G. Stein. Lens distortion calibration using rotation, with analysis of source of error. *In Proceedings on Computer Vision and Pattern Recognition (CVPR'97)*, pages 602–608, 1997.

T. F. Stepinski and C. Bagaria. Automatic mapping of martian physiography: Application to tharsis region. *In Proceedings of Lunar and Planetary Science Conference*, 2009.

C. V. Stewart. Robust parameter estimation in computer vision. *SIAM Reviews*, 41(3): 513–537, 1999.

A. Stoddart and A. Hilton. Registration of multiple point sets. *In Proceedings of IEEE International Conference on Pattern Recognition*, pages 40–44, 1996.

P. Sturm. Multi-view geometry for general camera models. *In Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR'2005)*, pages 206–212, 2005.

V. A. Sujan and M. A. Meggiolaro. Intelligent and efficient strategy for unstructured environment sensing using mobile robot agents. *In International Journal of Intelligent and Robotic Systems*, pages 217–253, 2005.

H. Surmann, A. Nuchter, and J. Hertzberg. An autonomous mobile robot with a 3d laser range finder for 3d exploration and digitization of indoor environments. *Robotics and Automation Systems*, 2003.

U. G. Survey. Usgs cmg infobank definitions. *http://walrus.wr.usgs.gov/infobank/programs/html/main/definitions.h* 2006.

R. Szeliski. Image alignment and stitching: A tutorial. *Foundation and Trends in Computer Graphics and Computer Vision*, 2, 2006.

R. Szeliski. Image mosaicing for tele-reality applications. *In IEEE Workshop on Applications of Computer Vision (WACVŠ94)*, pages 44–53, 1994.

R. Szeliski. Video mosaics for virtual environments. *In IEEE Computer Graphics and Applications*, 16(2):22Ű30, 1996.

R. Szeliski and H. Y. Shum. Creating full view panoramic image mosaics and texturemapped models. *In Proceedings of Computer Graphics (SIGGRAPHŠ97)*, page 251Ű258, 1997.

J.-P. Tardif, P. Sturm, and S. Roy. Self-calibration of general radially symmetric distortion model. *In Proceedings of Seventh European Conference on Computer Vision (ECCV'02)*, pages 186–199, 2006.

S. Teller and S. Coorg. Spherical mosaics with quaternios and dense correlation. *In International Journal Computer Vision*, 37(3):259–273, 2000.

D. Terzopoulos, A. Witkin, and M. Kass. Constraints on deformable models: recovering 3d shape and non-rigid motion. *In Artificial Intelligence Image and Vision Computing*, 36:91–123, 1988.

S. Thrun, W. Bugard, and D. Fox. A real-time algorithm for mobile robot mapping with application to mult-robot and 3d mapping. *In Proceedings on International Confenrence on Robotics and Automation*, pages 321–328, 2000.

S. Thrun, D. Koller, Z. Ghahramani, H.-F. Durrant-Whyte, S. Clark, and M. Csorba. A solution for the simultaneous mapping and localization with sparse extended information filters. *In Proceedings of the Fifth International Workshop on Algorithmic Fundations of Robotics*, 2002.

S. Thrun, D. Haehnel, D. Ferguson, M. Montemerlo, R. Triebel, W. Burgard, C. Baker, Z. Omohundro, S. Thayer, and W. R. L. Whittaker. A System for Volumetric Robotic Mapping of Abandoned Mines. *In IEEE International Conference on Robotics and Automation (ICRA'03)*, pages 4270–4275, May 2003.

S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics.* The MIT Press, 2005.

S. Thrun, M. Montemerlo, and A. Aron. Probabilistic terrain analysis for high-speed desert driving. *In Proceedings of Robotics: Science and Systems*, 2006.

S. Thurn and A. Bucken. Integrating grid-based and topological maps for mobile robot navigation. *In Proceedings of the 13th National Conference on Artificial Intelligence*, pages 944–950, 1996.

Q. Tian and M. N. Huhns. Algorithms for sub-pixel registration. *In Proceedings of Computer Vision, Graphics and Image Processing*, pages 220–233, 1986.

C. Tomasi and T. Kanade. Detection and tracking of point features. *Carnegie Mellon University Technical Report CMU-CS-91-132*, 1991.

C. Tomasi and T. Kanade. Shape and motion from from image streams under orthography: a factorization method. *In International Journal of Computer Vision*, 2(9):137–154, November 1992.

C. F. Touzet. Robot awareness in cooperative mobile robot learning. *In Autonomous Robots*, 8:87–97, 2000.

R. Triebel, P. Pfaff, and W. Burgard. Multi-level surface maps for outdoor terrain mapping and loop closing. *In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS'06)*, 2006.

B. Triggs. Detecting keypoints with stable position, orientation and scale under illumination changes. *In Proceedings of Eight European Conference on Computer Vision (ECCV'04)*, pages 100–113, 2004.

B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment - a modern synthesis. *In Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, pages 298–372, 1999.

R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv camera and lenses. *Robotics and Autonomous Systems*, 3(4):323–344, 1987.

G. Turk and M. Levoy. Zippered Polygon Mesh from Range Images. *In Proceedings of the 21st Annual Conference on Computer Graphics*, pages 311–318, May 1994.

I. Ulrich and I. Nourbakhsh. Appearance-based place recognition for topological localization. *In Proceedings of IEEE International Conference on Robotics and Automation*, pages 1023–1029, 2000.

M. Uyttendaele, A. Eden, and R. Szeliski. Eliminating ghosting and exposure artifacts in images mosaics. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'01)*, pages 509–519, 2001.

A. Victorino, P. Rives, and J.-J. Borelly. Safe navigation for indoor mobile robots, parti: A sensor-based navigation framework. *In International Journal of Robotics Research*, 22 (12):1005–1019, 2003.

R. Vidal, Y. Ma, S. Hsu, and Sastry. Optimal motion estimation from multiview normalized epipolar constraint. *In 8th IEEE International Conference on Computer Vision (ICCV'01)*, 2001.

P. Viola and W. Wells-III. Alignment by maximization of mutual information. *In Proceedings of Fifth International Conference on Computer Vision (ICCV'95)*, pages 16–25, 1995.

VIT. Visual information technology group, Canada. 2000.

L. Wang, S. B. Kang, R. Szeliski, and H. Y. Shum. Optimal texture map reconstruction from multiple views. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'01)*, pages 347–354, 2001.

J. Wasserman. *Michelangelo's Florence Pieta*. Princeton University Press, 2003.

J. Weijer and C. Schmid. Coloring local feature extraction. *In Proceedings of European Conference on Computer Vision (ECCV 06)*, pages 334–348, 2006.

J. Weingarten and R. Siegwart. EKF-based 3d SLAM for structured environments reconstruction. *In Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, pages 2089–2094, 2006.

T. Werner and A. Zisserman. New techniques for automated architectural reconstruction from photographs. *In Proceedings of the European Conference on Computer Vision (ECCV)*, pages 541–555, 2002.

R. T. Whitaker. A level-set approach to 3d reconstruction from range data. *In International Journal on Computer Vision*, 29(3):203–231, 1998.

L. Wiliam. Pyramidal parametrics. *In Proceedings of Computer Graphics*, 17(3):1–11, 1983.

C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys. 3d model matching with viewpoint-invariant patches (vip). *In Proceedings of IEEE International Conference on Intelligent Robots and Systems*, pages 1–8, 2008.

J. Wyngaerd and L. Gool. Automatic crude patch registration: Toward automatic 3d model building. *In Computer Vision and Image Understanding*, 87(1-3):8–26, 2002.

Y. Xiong and K. Turkowski. Creating image-based vr using self-calibrating fisheys lens. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pages 237–243, 1997.

S. Yamany and A. Farag. Free-form surface registration using surface signatures. *In Proceedings of International Conference on Computer Vision*, pages 1098–1104, 2001.

B. Yamauchi, A. Schults, and W. Adams. Mobile robot exploration and map building with continous localization. *In Proceedings of the 1998 IEEE International Conference on Robotics and Automation*, pages 3715–3720, 1998.

N. Yazawa, H. Uchiyama, and H. Saito. Image based view localization system retrieving from a panorama database by SURF. *In Proceedings of IAPR Conference on Machine Vision Applications (MVA'09)*, 2009.

A. Yezzi and S. Soatto. Stereoscopic segmentation. *In Proceedings of International Conference on Computer Vision*, pages 59–66, 2001.

A. Yezzi, G. Slabaugh, R. Cipolla, and R. Schafer. A surface evolution approach of propabilistic space carving. *In 3DPVT'02*, pages 618–621, 2002.

R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondaces. *In Third Eurpean Conference on Computer Vision*, May 1994.

J. Zhang and A. Rangarajan. Affine image registration using new information metric. *In Proceedings of International Conference on Computer Vision and Pattern Recognition*, 1:848–855, 2004.

L. Zhang and S. M. Seitz. Image-based multiresolution shape recovery by surface deformation. *In Proceedings of SPIE: Videometrics and Optical Methods for 3D Shape Measurement*, pages 51–61, 2001.

Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *In International Journal of Computer Vision*, 13(2):119–152, 1994.

Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. *In IEEE International Conference on Computer Vision*, pages 666–673, 1999.

Z. Zhang and O. Faugeras. Estimation of displacements from two 3-D frames obtained from stereo. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14 (2):1141–1156, 1992.

H. Zhao and R. Shibasaki. A system for reconstructing urban 3d objects using ground-based range and CCD sensors. *In Urban Multi-media/3D Mapping Workshop*, 1999.

W. Zhao, D. Nister, and S. Hsu. Alignment of Continuous Video onto 3D Point Clouds. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1308–1318, 2005.

I. Zoghlami, O. Faugeras, and R. Deriche. Using geometric corners to build a 2D mosaic from a set of images. *In Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pages 420–425, 1997.