Positionnement robuste et précis de réseaux d 'images

Pierre Moulon

▶ **To cite this version:**

Pierre Moulon. Positionnement robuste et précis de réseaux d 'images. Traitement du signal et de l'image [eess.SP]. Université Paris-Est, 2014. Français. NNT : . tel-00996935v1

HAL Id: tel-00996935

https://theses.hal.science/tel-00996935v1

Submitted on 27 May 2014 (v1), last revised 9 Jul 2015 (v2)

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.

LABORATOIRE D'INFORMATIQUE
GASPARD MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

UNIVERSITÉ
PARIS-EST
MARNE-LA-VALLÉE

École Doctorale Paris-Est
Mathématiques & Sciences et Technologies
de l'Information et de la Communication

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PARIS EST
Domaine : Traitement du Signal et des Images

présentée par **Pierre MOULON**
pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ PARIS EST

Positionnement robuste et précis de réseaux d'images.

Soutenue publiquement le 10 janvier 2014 devant le jury composé de :

Adrien BARTOLI	Université d'Auvergne Clermont1	Rapporteur
Julie DELON	Université Paris Descartes	Examinateur
David FOFI	Université de Bourgogne	Rapporteur
Marc PIERROT-DESEILLIGNY	École Nationale des Sciences Géographiques	Examinateur
Renaud MARLET	École des Ponts ParisTech	Directeur de Thèse
Benoît MAUJEAN	Mikros Image	Encadrant industriel
Pascal MONASSE	École des Ponts ParisTech	Co-Directeur de Thèse
Luc ROBERT	Autodesk	Examinateur

École des Ponts ParisTech
LIGM-IMAGINE
6, Av Blaise Pascal - Cité Descartes
Champs-sur-Marne
77455 Marne-la-Vallée cedex 2
France

Université Paris-Est Marne-la-Vallée
École Doctorale Paris-Est MSTIC
Département Études Doctorales
6, Av Blaise Pascal - Cité Descartes
Champs-sur-Marne
77454 Marne-la-Vallée cedex 2
France

Ici, mon cher, c'est adorable, et je découvre tous les jours des choses toujours plus belles. C'est à en devenir fou, tellement j'ai envie de tout faire, la tête m'en pète. [...] Eh bien, mon cher, je veux lutter, gratter, recommencer, car on peut faire ce que l'on voit et que l'on comprend, et il me semble, quand je vois la nature, que je vais tout faire, tout écrire, [...] quand on est à l'ouvrage [...] Tout cela prouve qu'il ne faut penser qu'à cela.
C'est à force d'observation, de réflexion que l'on trouve. Ainsi piochons et piochons continuellement [...].

Extrait d'une lettre de Claude Monet à Frédéric Bazille écrite en 1864.

Remerciements

Mes encadrants. Je remercie tout d'abord Benoît Maujean et Renaud Keriven pour m'avoir offert l'opportunité de réaliser ce travail de recherche au sein du laboratoire IMAGINE et de l'entreprise Mikros Image. C'est avec un immense plaisir que j'ai pu travailler sous la direction de Renaud Marlet, Benoît Maujean et Pascal Monasse pour leurs qualités pédagogiques, scientifiques et humaines. Profitant de leur infaillible soutien j'ai pu découvrir le monde de la recherche, de l'application de la recherche en industrie et en apprendre toujours plus sur la vision par ordinateur. Je les remercie pour leur disponibilité ainsi que leur patience face à mes nombreuses questions, ce qui m'a permis de réaliser avec confiance ce doctorat.

Comité de thèse. Je remercie Julie Delon, Marc Pierrot-Deseilligny et Luc Robert d'avoir accepté de faire partie du jury et je remercie tout particulièrement Adrien Bartoli et David Fofi pour avoir accepté d'être mes rapporteurs, en dépit du travail important que cela représente.

Mes collègues d'entreprise. Je remercie mes collègues pour les discussions techniques, les sujets aléatoires abordés, la passion partagée pour la technologie informatique et les langages de programmation : Lauren Agopian, Marc-Antoine Arnaud, Arnaud Chassagne, Guillaume Chatelet, Laurent Clavier, Julien Dubuisson, Adrien Dutertre, Michael Etienne, Thomas Eskenazi, Marie Fétiveau, Antonio Fiestas, Alexandra Lefève-Gourmelon, Guillaume Maucomble, Valentin Noël, Jules Pajot, Nicolas Provost, Michael Guiral, Nicolas Rondaud, Élodie Souton. Je remercie les personnes qui se reconnaîtront pour les nombreux traits d'humour partagés avec plus ou moins de succès. Je remercie tout particulièrement Bruno Duisit, Christophe Courgeau , Benoît Maujean et Guillaume Provôt pour avoir participé au projet MiMatte3D de sa genèse à sa réalisation concrète.

Mes collègues du laboratoire. Je remercie les membres permanents pour tous leurs conseils et suggestions qu'ils m'ont prodigués : Arnak Dalalyan et Guillaume Obozinski pour les discussions sur les optimisations convexes, Nikos Paragios pour ses précieux conseils pour l'écriture de '*rebuttal*', Bertrand Neveu pour toutes les références que tu as récupérées plus vite que l'éclair.

Je remercie également les post-doctorants, doctorants, futur doctorants et chercheurs du laboratoire pour la bonne humeur apportée au laboratoire : Martin De La Gorce, Alexandre Boulc'h, Amine Bourki, Raghudeep Gadde, Mateusz Kozinski, Zhe Liu, Francisco Vitor Suzano Massa, Yohann Salaün, Olivier Tournaire, Marina Vinyes, Zhongwei Tang. Je souhaite bonne continuation aux stagiaires que j'ai encadrés. Badis Djellab, Emmanuel Habbets, Tristan Faure, Luc Girod, Rafaël Marini Silva et Lucas Plaetevoet : Vous m'avez ouvert l'esprit sur de nouvelles problématiques. Je remercie aussi ceux qui sont partis vers d'autres horizons avant moi : Achraf Ben-Hamadou, Olivier Collier, Jamil Drareni, Ferran Espuny et Hoang-Hiep Vu. Je remercie David Ok, Victoria Rudakova et Pascal Monasse pour avoir fait de l'aventure PRoVisG Mars 3D Challenge un succès et une expérience inoubliable au *Jet Propulsion Laboratory* de la *NASA*. Enfin

je remercie Brigitte Mondou et Sylvie Cach pour leur disponibilité et réactivité qui nous facilitent le quotidien lors des missions et dossiers administratifs.

Mes anciens professeurs. Une pensée à tous mes professeurs qui grâce à leur pédagogie m'ont insufflé la passion du développement logiciel et de l'imagerie numérique.

Mes amis. Pour leur soutien et encouragements : Antonin P., Cyril L., Nicolas N., Philippe M., Michel T., Elvire et Ludovic T..

Ma famille. Je souhaite enfin exprimer ma gratitude envers mes proches qui m'ont toujours encouragé et mes parents pour m'avoir donné les moyens de réaliser mes études en adéquation avec mes passions. Enfin, mes plus profonds remerciements vont vers Fanny, ma chère et tendre, pour la patience et la compréhension dont elle a fait part durant ces trois dernières années et plus encore pour le bonheur que j'ai de vivre à ses côtés depuis notre rencontre.

Résumé

Calculer une représentation 3D d'une scène rigide à partir d'une collection d'images est aujourd'hui possible grâce aux progrès réalisés par les méthodes de stéréo-vision multi-vues, et ce avec un simple appareil photographique. Le principe de reconstruction, découlant de travaux de photogrammétrie, consiste à recouper les informations provenant de plusieurs images, prises de points de vue différents, pour identifier les positions et orientations relatives de chaque cliché. Une fois les positions et orientations de caméras déterminées (calibration externe), la structure de la scène peut être reconstruite.

Afin de résoudre le problème de calcul de la structure à partir du mouvement des caméras (Structure-from-Motion), des méthodes séquentielles et globales ont été proposées. Par nature, les méthodes séquentielles ont tendance à accumuler les erreurs. Cela donne lieu le plus souvent à des trajectoires de caméras qui dérivent et, lorsque les photos sont acquises autour d'un objet, à des reconstructions où les boucles ne se referment pas. Au contraire, les méthodes globales considèrent le réseau de caméras dans son ensemble. La configuration de caméras est recherchée et optimisée pour conserver au mieux l'ensemble des contraintes de cyclicité du réseau. Des reconstructions de meilleure qualité peuvent être obtenues, au détriment toutefois du temps de calcul.

Cette thèse propose d'analyser des problèmes critiques au cœur de ces méthodes de calibration externe et de fournir des solutions pour améliorer leur performance (précision, robustesse, vitesse) et leur facilité d'utilisation (paramétrisation restreinte).

Nous proposons tout d'abord un algorithme de suivi de points rapide et efficace. Nous montrons ensuite que l'utilisation généralisée de l'estimation robuste de modèles paramétriques *a contrario* permet de libérer l'utilisateur du réglage de seuils de détection, et d'obtenir une chaîne de reconstruction qui s'adapte automatiquement aux données. Dans un second temps, nous utilisons ces estimations robustes adaptatives et une formulation du problème qui permet des optimisations convexes pour construire une chaîne de calibration globale capable de passer à l'échelle. Nos expériences démontrent que les estimations identifiées *a contrario* améliorent de manière notable la qualité d'estimation de la position et de l'orientation des clichés, tout en étant automatiques et sans paramètres, et ce même sur des réseaux de caméras complexes. Nous proposons enfin d'améliorer le rendu visuel des reconstructions en proposant une optimisation convexe de la consistance colorée entre images.

Mots-clefs

calibration ; stéréovision multi-vue ; stéréovision ; estimation robuste ; programmation linéaire ; vision par ordinateur.

Abstract

To compute a 3D representation of a rigid scene from a collection of pictures is now possible thanks to the progress made by the multiple-view stereovision methods, even with a simple camera. The reconstruction process, arising from the photogrammetry consist in integrating information from multiple images taken from different viewpoints in order to identify the relative positions and orientations of each shot. Once the positions and orientations (external calibration) of the cameras are retrieved, the structure of the scene can be reconstructed.

To solve the problem of calculating the Structure from Motion (SfM), sequential and global methods have been proposed. By nature, sequential methods tend to accumulate errors. This provides most often trajectories of cameras that are subject to drift error. When pictures are acquired around an object it leads to reconstructions where the loops do not close. In contrast, global methods consider the network of cameras as a whole. The configuration of cameras is searched and optimized in order to to best preserve the constraints of the cyclical network. Reconstructions of better quality can be obtained, but at the expense of computation time.

This thesis aims to analyse critical issues at the heart of these methods of external calibration and provide solutions to improve their performance (accuracy , robustness and speed) and their ease of use (restricted parametrization).

We first propose a fast and efficient feature tracking algorithm. We then show that the widespread use of *a contrario* robust estimation of parametric models frees the user about choosing detection thresholds, and allows obtaining a chain of reconstruction that automatically adapts to the data. Then in a second step, we use the adaptive robust estimation and a series of convex optimization to build a scalable global calibration chain. Our experiments show that the *a contrario* identified estimates improve significantly the quality of the pictures's positions and orientations, while being automatic and without parameters , even on complex camera networks. Finally, we propose to improve the visual appearance of the reconstruction by providing a convex optimization of the color consistency between images.

Keywords

calibration ; multi-view stereovision ; stereovision ; robust estimation ; linear programming ; computer vision.

Sommaire

Chapitre 1

Avant propos

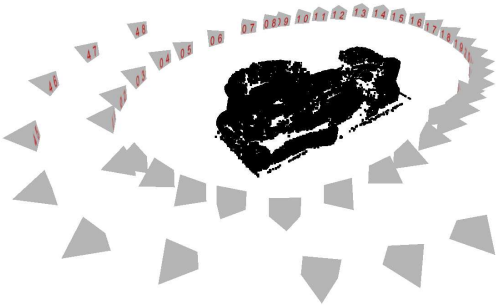
La reconstruction de l'espace tridimensionnel qui nous entoure à partir d'images est un des défis posés à la vision par ordinateur. Parmi les techniques possibles, la stéréo-vision est celle qui est la plus explorée. Son principe, découlant de travaux de photogrammétrie, est de recouper les informations provenant de plusieurs images prises de points de vue différents. Autrefois binoculaire et fournissant des informations partielles, la stéréo-vision est maintenant multi-vues et permet l'obtention de modèles complets de ce qui est observé. Des méthodes de reconstruction de *structure à partir du mouvement* (Structure-from-Motion) ont fait naître des nouvelles perspectives pour la photographie 3D. Ainsi avec un simple appareil photographique on peut désormais reconstruire un environnement en trois dimensions. Ce domaine porteur de la vision par ordinateur ouvre de nouveaux horizons et un champ d'application qui va bien au delà des besoins initiaux suscités par la robotique. Les applications possibles sont nombreuses : architecture et urbanisme (DE LUCA 2009), archéologie, métrologie, cartographie, divertissement (panoramas, visites virtuelles, jeux vidéo interactifs). Les retombées pour la production cinématographique et les effets spéciaux sont évidemment multiples.

Les travaux de cette thèse concernent l'application de la stéréo-vision pour la reconstruction la plus précise possible de décors à partir de photographies pour l'industrie audio-visuelle.

La captation du réel

Réaliser l'acquisition d'un environnement réel sur un support numérique comporte trois étapes principales (De Luca 2006) :

L'acquisition des données spatiales met en œuvre le relevé de la morphologie, des dimensions et des aspects de surface de l'environnement étudié. Cette phase peut utiliser différents dispositifs basés sur le principe de mesure avec ou sans contact sous différentes configurations. Dans le cas de la photogrammétrie le résultat de cette phase consiste en un nuage de points (la structure) représentant avec plus ou moins de densité l'environnement et une série d'images orientées et positionnées dans l'espace.



La reconstruction tridimensionnelle des surfaces est l'étape de modélisation qui permet de construire le modèle géométrique de l'édifice en s'appuyant sur les mesures issues de la phase de relevé. Plusieurs techniques permettent une reconstruction automatique, semi-automatique ou manuelle des surfaces à partir des nuages de points. Ces techniques diffèrent en fonction des données d'entrées qu'elles peuvent traiter et du type de représentations géométriques qu'elles peuvent générer.

La restitution de l'apparence visuelle s'intéresse à l'enrichissement de la géométrie issue de la phase de reconstruction. Des attributs capables de décrire les aspects de surface sont ajoutés sur la reconstruction. Il s'agit principalement d'associer au modèle 3D les informations photométriques acquises au moment du relevé.

L'acquisition des données spatiales sous une forme numérique est généralement réalisée par des méthodes dites de métrologie. Bien que cette thèse se concentre sur des méthodes de photogrammétrie il est important de citer les différentes méthodes de numérisation existantes. Il sera ainsi plus facile pour le lecteur de comprendre que la photogrammétrie est un choix privilégié dans le cadre de ce travail.

Les méthodes d'acquisition du réel peuvent être classifiées en deux catégories : les méthodes dites avec ou sans contact.

Avec contact. Les méthodes avec contact réalisent la numérisation d'un objet 3D grâce à un contact physique avec l'objet.

Palpeur

La numérisation est réalisée par le biais d'un palpeur et d'un bras articulé. Les mesures angulaires sur les articulations de l'arbre permettent de connaître précisément la position du palpeur et permettent ainsi de numériser des points de l'espace. Son usage intrusif envers les objets rend la numérisation d'objets fragiles périlleuse et l'acquisition d'objets de large dimension impossible. Un autre désavantage est la fréquence d'acquisition qui est limitée par l'opérateur lui-même, contraint par la vitesse de déplacement du bras mécanique. L'acquisition d'une surface dense n'est donc pas envisageable par un opérateur. Ce type d'acquisition est le plus souvent limité au milieu industriel et à la vérification de cotes sur des chaînes de production.

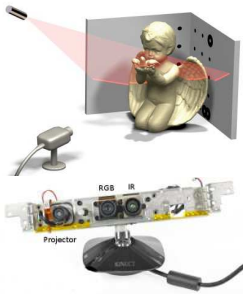
Sans contact. Les méthodes sans contact sont réalisées avec des appareils d'acquisition distants. On distingue deux méthodes d'acquisition, les méthodes actives et les méthodes passives.

Méthodes actives :

Télémétrie

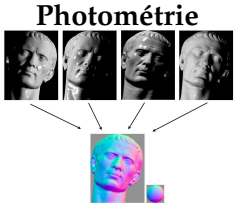
Les scanners actifs émettent un rayonnement et détectent sa réflexion afin de sonder un objet ou une scène. Différents types de source de rayonnement sont utilisés : lumière, ultrason ou rayon X. Les appareils de mesures les plus connus de cette catégorie sont les scanners LIDAR (dits à temps de vol) et les scanners 3D (dits à décalage de phase). Les scanners LIDAR ont une portée plus grande et une fréquence d'acquisition plus élevée (10 000 à 100 000 points par seconde) que les scanners à décalage de phase. Ces technologies ont un coût élevé et demandent une formation pour être utilisées. L'acquisition de larges volumes requiert plusieurs acquisitions avec la présence de marqueurs cibles à position fixe pour faciliter les recalages. Ils sont donc assez complexes à réaliser.

Photogrammétrie et triangulation



Les scanners dits à lumière structurée utilisent un appareil photo et une source de lumière contrôlée (un vidéo-projecteur). L'analyse de la déformation d'un motif lumineux projeté sur l'objet permet de déterminer le relief de la surface imagée. Selon le temps et la précision de la reconstruction souhaitées on utilise un ou plusieurs motifs (lignes, points). La démocratisation de ce type de scanner a été réalisée avec brio par Microsoft et son produit Kinect. La Kinect est un scanner 3D qui réalise l'analyse des déformations en temps réel d'un motif projeté en infrarouge afin de calculer une carte de profondeur et localiser les positions du squelette d'un ou plusieurs joueurs. Un inconvénient de ces scanners est que leur précision est limitée à la zone de netteté de la caméra et donc seuls de petits espaces peuvent être reconstruits. Ces méthodes utilisant une source de lumière infrarouge, les acquisitions se limitent à des espaces intérieurs dans un espace contrôlé afin de ne pas être perturbées par une source externe de lumière (soleil).

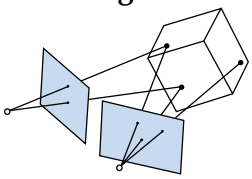
Stéréo-Photométrie



Ici on se place dans un cas similaire au précédent, on considère toujours un appareil photographique fixe, mais on considère désormais une source de lumière unique en mouvement. Le fait d'avoir différentes images avec des conditions d'illumination différentes permet de déterminer l'état de la surface de l'objet considéré. Des normales à la surface sont ainsi calculées et une phase d'intégration permet de déterminer une surface représentant l'objet observé. Image extraite de Wu et al. (2011b).

Méthode passive :

Photogrammétrie et triangulation



Des images sont capturées autour de l'objet à mesurer. Connaissant des points en correspondance entre les images, on peut identifier les positions des caméras et des points 3D correspondants par triangulation. On identifie ainsi le mouvement des caméras (orientation et translation) ainsi que la structure (points 3D) de la scène. Ce problème d'optimisation est résolu par des algorithmes de calcul de *structure à partir du mouvement*. Ces points peuvent être soit des points naturels détectés soit des points identifiés par des marqueurs cibles posés sur la scène imagée.

La photogrammétrie passive apparaît comme une solution particulièrement intéressante :

- Le pré requis matériel est faible, seul un appareil photographique est nécessaire,
- Le prix d'un appareil photographique numérique de bonne qualité est moindre que le prix d'un scanner de type LIDAR,
- Aucune formation particulière n'est nécessaire pour manipuler le matériel,
- La scène observée n'est pas manipulée ou détériorée,
- Aucune source de lumière projetée et aucuns contacts aux objets ne sont nécessaires.

1.1 La photogrammétrie

Le mot photogrammétrie apparaît comme une évolution du mot, «métrophotographie», apparu en 1850 par le biais d'Aimé Laussedat. Le terme se popularise ensuite à l'échelle européenne puis internationale en photogrammétrie sous l'impulsion allemande du photographe Otto Kersten et de l'ingénieur civil Albrecht Meydenbaeuer (WOCHENBLATT 1867) comme illustré sur la figure 1.1. L'idée originale est de réaliser des relevés métriques de bâtiments ou terrains à partir de photographies.

La photogrammétrie a ensuite évoluée en commençant sur des travaux basées sur de la stéréovision (stereoscopic viewing) et les travaux de Carl Pulfrich sur le *stereo-comparator* créé en 1901 (cf. FRITSCH (2006)). L'acquisition de données topographiques a été initiée par des pionniers comme Nadar en 1858 avec l'acquisition d'images aériennes en ballon à des fins militaires. Par la suite, d'autres techniques d'acquisitions ont suivi. Durant la première guerre mondiale, des cerf-volants, avions et même des pigeons (PHOTOGRAPHIQUES 1910) ont servi de moyens de transport pour l'acquisition et la reconnaissance de terrains. Cette période a permis de découvrir et de proposer des solutions de correction pour la rectification et l'utilisation d'images stéréographiques.

Dans une seconde phase, le développement de la géométrie algorithmique projective, la connaissance avancée du calcul matriciel et de l'algèbre linéaire ont donné naissance à la photogrammétrie analytique (KRUPPA 1913) et à la théorie de l'ajustement de faisceaux (BROWN 1976 ; SLAMA et al. 1980 ; TRIGGS et al. 2000). L'ajustement de faisceaux est le processus qui consiste à optimiser simultanément la trajectoire de la caméra et la structure de la scène. La photogrammétrie étant gourmande en calculs numériques son utilisation a été grandement facilitée par l'arrivée de l'ordinateur.

Le troisième fait marquant dans l'histoire de la photogrammétrie est l'apparition de la version moderne de la *camera obscura* (Aristote) : l'appareil photographique numérique (GARETH A. LLOYD et STEVEN J. SASSON 1978). L'acquisition numérique et l'accessibilité à des solutions de stockage de plus en plus grandes se sont tellement démocratisées que l'on a observé une scission des communautés de recherche en fonction du style d'acquisition et d'application (cf. figure 1.1). On note après 1970 la pleine croissance de la photogrammétrie aérienne et la télé-détection : (*remote sensing*), puis plus tard l'apparition du terme *Structure from Motion* (SfM) et *digital photogrammetry* dans les années 1980-90. Cette disparité temporelle a été observée car au début les appareils photographiques numériques n'étaient accessibles qu'à la communauté acquisition aérienne. Dans un second temps l'appareil photographique numérique est devenu grand public et la photogrammétrie au sol (SfM, *close-range photogrammetry*) alors a pu se démocratiser.

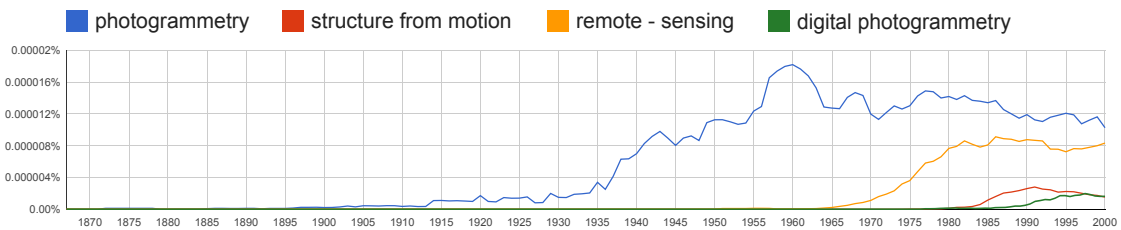


FIGURE 1.1 – Nombre d'occurrences des mots *photogrammetry*, *structure-from-motion* et *remote-sensing* dans les ouvrages référencés par Google© entre les années 1860 à 2000.

Nous sommes actuellement dans une quatrième phase de la photogrammétrie : une phase applicative. Nous pouvons observer que la communauté de la vision par ordinateur à fait mûrir des techniques et des applications qui sont désormais utilisables de

manière stable par tout un chacun. On notera que certaines applications permettent de nos jours :

De créer une image panoramique depuis nos téléphones mobiles,

De chercher de l'information en photographiant une pochette de disque ou une affiche : LTU Technologies [1], Kooaba [2],

De jouer de manière interactive avec un avatar virtuel imitant nos mouvements sur notre télévision (Microsoft Kinect).

Une utilisation concrète de la photogrammétrie et de la vision par ordinateur à très large échelle est le logiciel Google Maps. Cette application permet de visualiser la surface de notre planète à travers notre navigateur Internet (une couverture intégrale du globe en basse définition est disponible depuis 2005). La résolution en mode de visualisation aérienne est telle que l'on peut observer sa propre maison ou compter les piétons sur une place. L'inclusion récente des rues avec StreetView en 2007 permet de naviguer dans les rues d'une ville, de visualiser concrètement la situation d'un monument ou d'un magasin comme si on y était. Même si les informations actuellement proposées sont en majorité seulement en 2 dimensions, des représentations 3D sont d'ores et déjà en préparation ou visibles pour certaines villes du globe. La technologie de numérisation 3D de villes est en passe d'être mûre pour des applications concrètes comme l'ont montrés les sociétés "C3 Technologies" et Acute3D [3].

Le futur laisse entrevoir des solutions libres de partage et création de carte 3D à l'instar d'OpenStreetMap auxquelles des utilisateurs ordinaires peuvent contribuer pour apporter de l'information. Le fait que l'acquisition humaine soit limitée au sol est aussi en phase de changement. Les moyens de transport suivant l'évolution aéronautique au plus proche (ballons radiocommandés dirigeables, drones et UAV), le futur laisse imaginer que l'acquisition aérienne sera réalisable par tout un chacun dans quelques années avec un simple drone tel que le "Teeny, Tiny Crazyflie Nano Quadcopter" développé en 2012 (cf. figure 1.2).

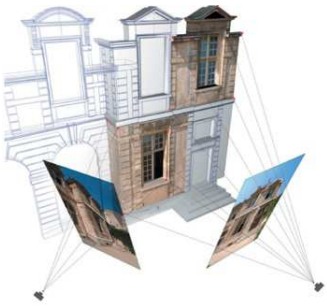
FIGURE 1.2 – De gauche à droite : Nadar et son ballon 1858, un pigeon photographe 1910, un drone Parrot 2010, un micro UAV 2012.

Cette obsession de recréer le réel pour en redéfinir l'usage est le but principal visé par l'industrie audiovisuelle. Voyons les usages de la photogrammétrie pour la création d'effets spéciaux.

1. LTU Technologies http://www.ltutech.com/fr/
2. Kooaba http://www.kooaba.com/
3. Acute3D http://www.acute3d.com

1.2 La photogrammétrie et les effets spéciaux

Pour le domaine des effets spéciaux, le terme photogrammétrie est interprété comme : une méthode pour acquérir une représentation manipulable d'un environnement. On cherche à acquérir le réel pour en détourner l'usage. L'intérêt est d'obtenir des copies numériques pour réaliser des trucages. Une collection de bâtiments (DE LUCA 2006) ou d'objets et personnages (BHAT et BURKE 2011) peut être ainsi créée et détournée (cf. figure 1.3). Les budgets étant souvent serrés la photogrammétrie s'impose comme un choix avant tout financier. Le prix d'un appareil photo numérique est bien moins élevé que celui d'un laser 3D d'acquisition LIDAR. Un autre avantage technique est le fait que le support photographique apporte la représentation photo-réaliste tandis que les LIDARs ne possèdent pas tous une caméra coaxiale pour acquérir avec précision la couleur de chaque point 3D numérisé. Cependant malgré le fait que les techniques de photogrammétrie semblent relativement mûres on réalise que l'usage de la technologie n'est pas encore aisée pour les besoins spécifiques des effets spéciaux.



(a)

(b)

FIGURE 1.3 – (a) La structure d'un bâtiment re-créée à partir de photographies (DE LUCA 2006). (b) Copie numérique d'un acteur par la société Agence de Doublure Numérique (image du Figaro).

Les objectifs principaux liés aux effets spéciaux mêlant réel et virtuel sont les suivants :

- l'estimation du mouvement d'une caméra vidéo, ***Match-moving*** :
  insertion d'éléments virtuels de manière réaliste sur une vidéo de tournage.
- La photo-modélisation, ***Image-Based-Modeling IBM*** :
  la création d'une copie numérique d'un environnement à partir de photographies.

1.2.1 *Le Match-moving*

Le *match-moving/motion-tracking* est une technique utilisée pour identifier la trajectoire d'une caméra à partir d'une séquence vidéo. Ayant la connaissance d'une caméra virtuelle il est possible de faire bouger des objets 3D qui auront un mouvement en cohérence avec la vidéo. La fusion de la scène réelle avec la scène virtuelle (*compositing*) donne alors l'impression qu'elles ont été filmées du même point de vue. On notera deux catégories de *match-moving* en fonction des dégrés de liberté du mouvement recherché : le suivi 2D dit bidimensionnel et le suivi 3D dit tridimensionnel.

Le suivi de mouvement bidimensionnel est disponible dans des logiciels tels que Adobe After Effects, Discreet Combustion et Shake. Cette technique se limite au suivi du mouvement de points particuliers choisis par l'utilisateur dans les images de la séquence. Une fois le mouvement de ces points identifié il est appliqué à de nouveaux objets venant occulter la vidéo avec un nouveau contenu. Cette technique est suffisante pour des surfaces planes, des mouvements de caméras simples et si il n'y a pas eu de changements importants des paramètres de la caméra. L'usage le plus classique est le remplacement d'un panneau publicitaire placé en arrière-plan d'une séquence vidéo par une autre image.

Le suivi de mouvement tridimensionnel va quant à lui extrapoler les informations tridimensionnelles (le mouvement de caméra) à partir de photographies bidimensionnelles (la séquence vidéo). Le processus d'estimation de la trajectoire de la caméra requiert l'estimation de contraintes de géométrie projective et l'application du processus d'ajustement de faisceaux. Les points suivis sont la plupart du temps précisés par l'opérateur. Des méthodes automatiques existent pour identifier certains points saillants, mais dans la plupart des cas des retouches manuelles sont nécessaires si la séquence vidéo présente des éléments en désaccord de mouvement. En effet les méthodes couramment utilisées considèrent en pré-requis que la scène observée est statique. L'opérateur vient alors supprimer les points qui ne sont pas sur la scène rigide : objets ou acteurs en mouvement.

Parmi les logiciels capables d'effectuer un match moving tridimensionnel on peut citer :

- 2d3 Boujou,
- Blender (depuis la version 2.61),
- Icarus (logiciel gratuit),
- Maya Live (Module de Maya Unlimited),
- PixelFarm PFTrack (réincarnation commerciale d'Icarus),
- Realviz MatchMover (racheté par Autodesk),
- Ssontech SynthEyes,
- Science.D.Visions 3DEqualizer,
- Voodoo (logiciel gratuit),
- VooCAT (logiciel commercial, réincarnation commerciale de Voodoo),
- VideoTrace.

Le marché du logiciel est assez diversifié. Des alternatives (commerciales, gratuites et open-source) existent et montrent que le marché est large et demandeur. Le suivi de mouvement de scène non rigide est quant à lui un domaine encore en évolution et de ce fait aucune solution logicielle commerciale n'est présente sur le marché.

1.2.2 La PhotoModélisation/*Image-Based-Modeling*

Dans le cas de l'*Image-Based-Modeling* on recherche à créer une représentation photo-réaliste 3D des éléments photographiés. Une solution logicielle doit être identifiée pour les trois étapes évoquées au début de ce chapitre : l'acquisition des données spatiales, la reconstruction tridimensionnelle des surfaces, la restitution de l'apparence visuelle (cf. figure 1.4).

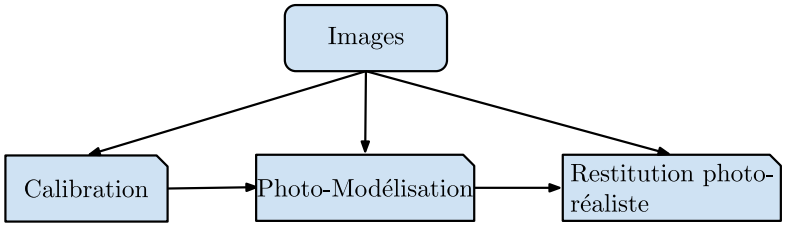


FIGURE 1.4 – Les étapes nécessaires pour la photo-modélisation : la calibration pour acquérir les données spatiales, la photo-modélisation pour reconstruire une surface et enfin le calcul de la restitution visuelle colorée.

L'application de la photogrammétrie pour la reconstruction de bâtiment comme élément de décor 3D pour le domaine des effets spéciaux a été initiée par DEBEVEC et al. (1996) avec son logiciel FAÇADE. Les auteurs proposent d'optimiser simultanément la reconstruction tridimensionnelle de surfaces planes et le placement des caméras dans l'espace. Les entrées de l'algorithme sont des primitives géométriques, comme des parallélépipèdes, placées manuellement par l'utilisateur dans les images sur les formes d'un bâtiment. En connaissant la projection d'une série de plan et de contraintes orthogonales, des blocs 3D et images sont ainsi orientés et placés dans l'espace. Dans un second temps, un raffinement manuel de la géométrie et une projection de texture permet d'enrichir le détail du modèle 3D (cf. figure 1.5).

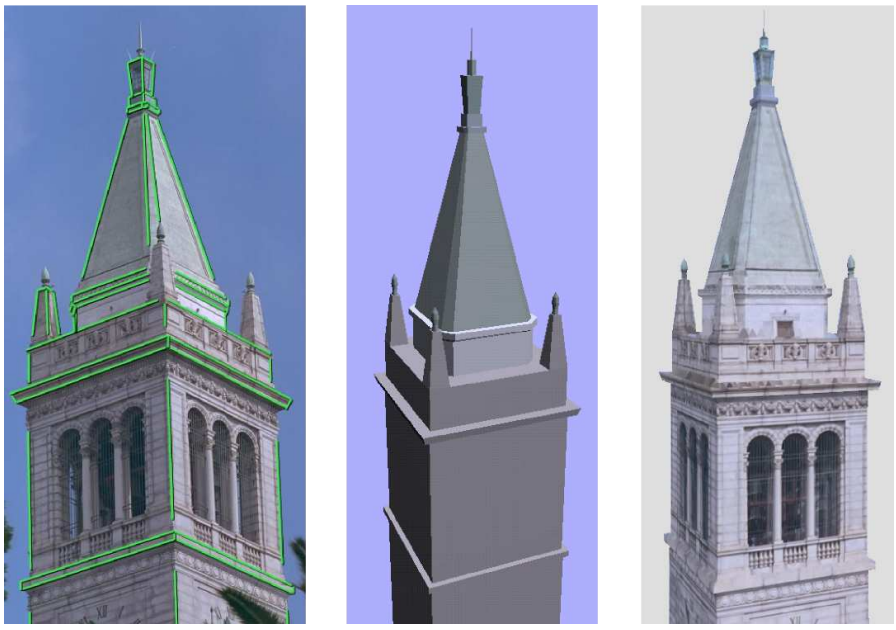


FIGURE 1.5 – Le logiciel FAÇADE (de gauche à droite) : les arêtes des parallélépipèdes utilisées, la reconstruction 3D obtenue et le rendu photo-réaliste.

A la suite de ce projet, trois classes de solutions ont émergé :

1. des solutions de reconstruction 3D par saisie manuelle,
2. des solutions automatiques avec le mûrissement des techniques de photogrammétrie,
3. des solutions semi-automatiques permettant d'intéragir avec les résultats obtenus de manière automatique.

1.Les solutions par saisie manuelle

Quatre solutions logicielles de reconstruction 3D par saisie manuelle ont émergé sur le marché :

- **Canoma 1999**
  Évolution commerciale de FAÇADE (DEBEVEC et al. 1996). Cette solution a disparu du marché suite au rachat mené par MetaCreations puis par Adobe Systems en 2000.
- **Eossystems PhotoModeler 1993**
  Précurseur sur le marché, le logiciel ne cesse d'évoluer depuis.
- **RealViz Image Modeler 2000**
  Transfert technologique issu de l'INRIA (Projet Robotvis). Racheté par Autodesk en 2009.
- **Banzai Pipeline Ltd Enwaii 2008**
  Conception d'un logiciel dédié pour les contraintes liées à la production des effets spéciaux. La solution s'intègre à un outil métier de la production visuelle : Autodesk Maya.

Ces solutions requièrent que l'utilisateur saisisse des informations en correspondance entre images (le plus souvent des points). Ces points sont utilisés pour la phase de calibration, des caméras sont ainsi positionnées dans l'espace et des points 3D très éparses sont reconstruits. Dans un deuxième temps, l'utilisateur peut réaliser à la main la photo-modélisation. Des faces sont ainsi saisies entre les points 3D (amers). Cette tâche reste un travail de longue haleine mais permet un contrôle précis sur les amers 3D utilisés et permet de guider la modélisation à faible nombre de polygones.

FIGURE 1.6 – ImageModeler : De gauche à droite, les images sources, les points et la géométrie saisies manuellement, la restitution colorée.

2.Les solutions automatiques

Par la suite, le développement des techniques de photogrammétrie a permis l'émergence de solutions automatiques. Ces solutions ont vu le jour grâce à :

- une évolution marquante de la stabilité de la recherche de points saillants communs entre images (SIFT : LOWE (1999)).
- l'évolution des algorithmes de *structure à partir du mouvement* (POLLEFEYS et al. 2000 ; BROWN et LOWE 2005a ; SNAVELY et al. 2006).

Ces améliorations notables ont permis de réaliser automatiquement :

- le calcul de la pose d'images dans l'espace,
- la création d'un nuage de points dense représentant la scène,
- la création d'une surface représentant la scène,
- la projection des images sources sur une surface pour une représentation photoréaliste.

Des résultats très réalistes peuvent être obtenus si les photographies acquises sont en adéquation en résolution et netteté pour les détails que l'on souhaite obtenir (cf. figure 1.7).

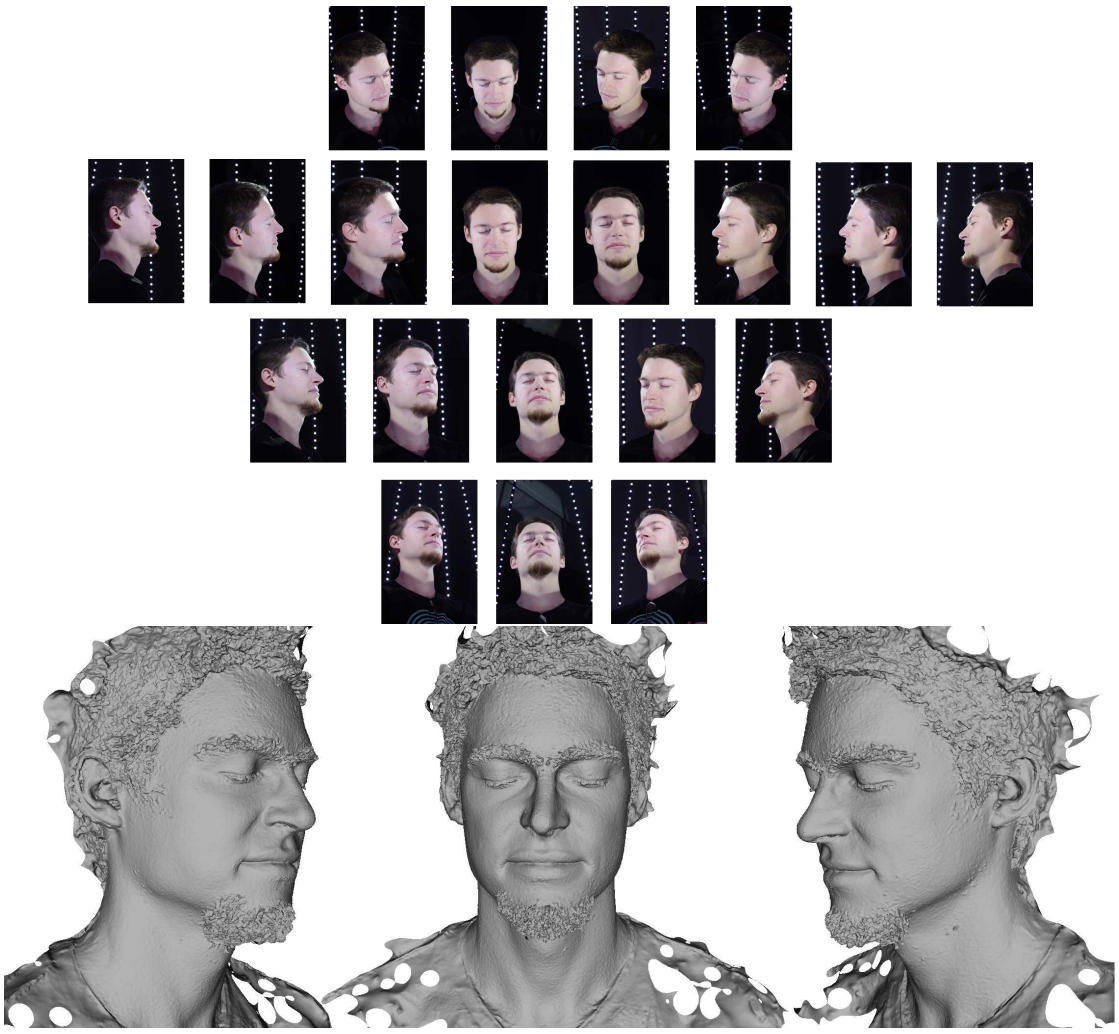


FIGURE 1.7 – Création d'une copie numérique d'un visage à partir de 20 images, merci à Cédric Guiard, Gilles Gambier et Pierre Lelièvre de ADN (Agence de Doublure Numérique) pour l'acquisition de ces images.

Les logiciels suivants sont apparus par la suite (liste non exhaustive) :

Société	Produit	lancement	Pays	Transfert de technologie
Metria	**Orthoware**	2007	Espagne	Université de Valence
Microsoft	**Photosynth**	2008	USA	Université de Washington
Agisoft	**Photoscan**	2010	Russie	?
Eossystems	**Photomodeler**	2010*	Canada	?
Autodesk	**PhotoFly**	2011	France	Realviz
Acute3D	**SmartCapture**	2011	France	ENPC (IMAGINE)
Pix4D	**PixelScanner**	2011	Suisse	EPFL (Cvlab)
3DFlow	**3DFZephyr**	2012	Italie	Université de Vérone
Aurvis	**PixelScanner**	2013	Turquie	Doctorant EPFL (Cvlab)

TABLE 1.1 – Listes des solutions commerciales pour la photogrammétrie automatique. *Intégration de composants automatiques depuis 2010.

On remarque que les 3/4 des solutions sont issues de savoir-faire académique et d'universités et que ces solutions automatiques émergent toutes dans les deux premières décennies des années 2000. Les ruptures significatives des dernières années citées au début de cette section montrent que la photogrammétrie est stable pour la réalisation d'applications concrètes et que le calcul de structure à partir du mouvement est donc en passe de devenir accessible pour le plus grand nombre.

Le fait le plus marquant qui démontre que la technologie est attractive est le projet Photosynth. Ce projet, basé sur une collaboration de Microsoft avec l'université de Washington et les travaux de SNAVELY et al. (2006) : "Photo tourism, exploring photo collections in 3D", permet d'explorer de manière interactive ses collections de photos personnelles en 3D. Le navigateur web est alors transformé en interface de navigation où l'on se promène en 3 dimensions d'image à image (cf figure 1.8). Microsoft a su mettre la technologie en accès libre via une plateforme de démonstration Internet pour la visualisation et un logiciel client pour réaliser les calculs de photogrammétrie. Malgré le fait que le résultat du calcul ne soit utilisable qu'à travers une interface Internet et non téléchargeable, cette application a fortement démocratisé la reconstruction 3D à partir de photographies.

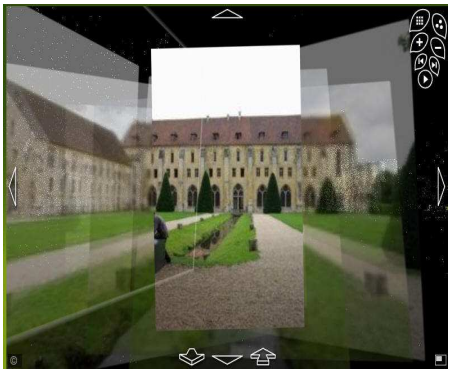


FIGURE 1.8 – Photosynth : une interface de navigation sur une collection d'images positionnée en 3 dimensions.

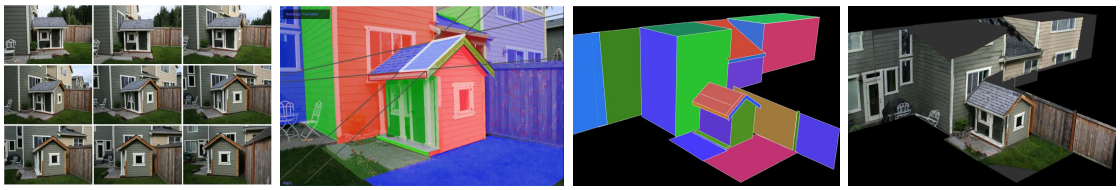
3.Les solutions semi-automatiques

Quelques solutions semi-automatiques ont été proposées par le milieu académique mais elles ne sont pas encore apparues dans des logiciels commerciaux. Elles proposent de faciliter les étapes de photo-modélisation en proposant des amers initiaux afin de faciliter la création de primitives géométriques. Deux solutions sont illustrées ici : **PhotoModel** et **O-Snap**.

PhotoModel. (SINHA et al. 2008) considère une scène calibrée. Le logiciel utilise les données suivantes :

– des caméras positionnées et orientées dans l'espace,
– un nuage de point initial,
– des lignes reconstruites en 3D (lignes de fuite détectées dans les images).

Lorsque l'utilisateur souhaite dessiner une facette 3D, il dessine les contours de cette face sur l'image de son choix. La position 3D de la face est alors interprétée automatiquement en fonction des données 3D visibles projetées à l'intérieur du polygone utilisateur. Une équation de plan 3D est ainsi déterminée automatiquement en ayant utilisé que peu d'interactions utilisateur, saisie dans une seule image (contrairement aux méthodes manuelles vues précédemment).



(a)Images (b) Saisie 2D (c) Modèle 3D (d) Modèle 3D texturé

FIGURE 1.9 – PhotoModel : une interface de photo-modélisation qui se base sur des interactions utilisateur et le support 3D de points et lignes de fuite.

O-Snap. (ARIKAN et al. 2013) propose une interface de modélisation qui utilise seulement le nuage de point 3D. La reconstruction 3D polygonale peut être interactivement raffinée par l'utilisateur. Un modèle initial est automatiquement créé via la génération d'hypothèses de polygones plans les plus probables. L'utilisateur guide ensuite la méthode automatique, vers le résultat qu'il souhaite, en indiquant des relations d'adjacences entre polygones pour former la géométrie désirée (exemple : les connections entre les murs et toits de la figure 1.10).

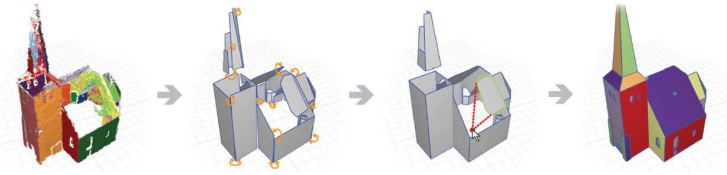


FIGURE 1.10 – Un aperçu de la chaîne semi-automatique de modélisation O-Snap. De gauche à droite : Un nuage de point bruité et incomplet est décomposé en une sélection de plans. Des relations d'adjacences de polygones proches sont identifiées afin de les connecter et de raffiner le modèle. L'utilisateur guide ensuite le processus automatique en rajoutant manuellement des relations d'adjacences. Un modèle à faible nombre de polygones représentant le nuage de points 3D peut ainsi être reconstruit.

La photogrammétrie et la photo-modélisation pour les effets spéciaux

Lorsque l'on regarde les solutions existantes, on constate qu'une seule solution est pour l'instant dédiée au monde des effets spéciaux audiovisuels. Il s'agit de la solution «Enwaii» proposée par «Banzai Pipeline Ltd.». Bien que la solution soit entièrement manuelle, elle présente l'avantage de s'intégrer directement à un outil métier utilisé dans le domaine : Autodesk Maya. Cette solution propose une alternative pour prendre en compte les contraintes liées à la production de contenus multimédias pour l'industrie. Cependant les solutions manuelles présentent un défaut majeur : c'est le niveau d'implication de l'utilisateur qui conditionne la précision de saisie des correspondances et donc la qualité du résultat obtenu. Les étapes de calibration et modélisation étant manuelles, l'utilisateur doit être formé afin de produire de bons résultats. Un utilisateur avisé sera à même d'obtenir de bien meilleurs résultats qu'un novice.

Les solutions automatiques, comme Agisoft Photoscan, demandent quant à elles peu de formation, mais en contre partie ne donnent pas de contrôle à l'utilisateur sur la chaîne de traitement. Lorsqu'un jeu de photographies ne permet pas l'obtention de modèle 3D automatique, l'utilisateur ne peut obtenir aucun résultat.

L'idéal pour le marché de la postproduction audioviselle serait une solution semi-automatique fonctionnant de l'acquisition photographique sur site à la production du contenu 3D final. L'utilisateur pourrait alors guider le processus automatique sur des jeux d'images restreint ou au contraire être guidé sur des jeux de données comportant de nombreuses images.

1.3 Contexte de la thèse

Cette thèse CIFRE commencée en octobre 2010 a été effectuée au sein du groupe de recherche IMAGINE pour la tutelle laboratoire et de l'équipe recherche et développement de MIKROS IMAGE pour la tutelle entreprise.

IMAGINE. Le groupe de recherche IMAGINE est un projet collaboratif entre l'École des Ponts Paristech et Chaussée (ENPC) et du Centre Scientifique et Technique du Bâtiment (CSTB). Ce groupe de recherche appartient au Laboratoire Informatique Gaspard Monge (LIGM) de l'Université Paris-Est Marne-la-Vallée (UPEM). L'expertise d'IMAGINE se situe en vision par ordinateur, en traitement de maillage, en apprentissage statistique, en optimisation, et en programmation par contraintes. Une partie des travaux actuels concerne les thématiques suivantes :

- La reconstruction haute précision de surfaces 3D à partir de grandes quantités d'images acquises sous des conditions non contrôlées. Expertise transférée en 2011 au sein de l'entreprise Acute3D par Renaud Keriven et Jean-Philippe Pons.
- L'amélioration des méthodes de calibration de caméra par le biais de l'utilisation de méthodes statistiques avec le projet ANR Callisto. Ce projet, en collaboration avec le CNES est réalisé dans le cadre du projet MISS (Mathématiques de l'Imagerie Satellitaire Spatiale).
- L'interprétation des images et leur sémantisation pour reconstruire des façades de bâtiments riches d'informations (fenêtres, portes, ...).

Le travail de l'équipe IMAGINE a été notamment remarqué à l'échelle internationale grâce à des résultats en reconstruction de surface et d'algorithmes de stéréo-vision multiple-vues denses (Hiep et al. 2009). Les reconstructions les plus précises et les plus

complètes ont été obtenues sur le jeux de données de référence mise en place par le CVLAB de l'EPFL (STRECHA et al. 2008). L'équipe a également obtenu en 2011 le premier prix au challenge "PRoVisG Mars 3D Challenge" consistant à évaluer la précision de reconstruction de la trajectoire d'un robot terrestre et martien.

Mikros Image. Créé en 1985, Mikros Image est un prestataire de services spécialisé dans les effets numériques visuels. Mikros Image gère pour ses clients plus de 300 projets par an tous domaines confondus, depuis la supervision de tournage, jusqu'à la finalisation de films de cinéma, de spots de publicité, de programmes de télévision ou de contenus pour Internet et la téléphonie mobile. La gamme de services proposée comprend :

- Effets spéciaux et images de synthèse 2D & 3D,
- Animation,
- Montage et conformation,
- Étalonnage, transferts de support numériques/argentiques et argentiques/numériques, masterisation,
- Laboratoires vidéo, film & compression,
- Gestion d'actifs et outils de transmission numérique.

Mikros Image est une filiale à 100% de la société MTC (Multimédia Télévision Cinéma), dont le capital est majoritairement détenu depuis octobre 2006 par la société italienne Mediacontech, cotée à Milan. Son effectif compte plus de 100 salariés fixes et environ 50 intermittents free-lances.

Mikros Image possède des antennes dans quatre pays : France, Belgique, Luxembourg, Canada. Depuis 1999, Mikros Image possède une activité de recherche et développement qui mobilise environ 10% de l'effectif de la société. Ses outils «maison» permettent d'augmenter ses capacités de production, d'optimiser la qualité de ses réalisations et d'offrir de nouveaux services, notamment pour son développement stratégique et commercial. Certaines de ses applications sont développées avec des partenaires industriels ou universitaires, dans le cadre de projets collaboratifs. Le financement de ses outils est en partie assuré par des organismes de soutien public à l'innovation.

Mikros Image se trouve sur un marché en pleine expansion. En effet, les films et les publicités utilisent de plus en plus d'effets visuels. De plus, compte tenu de la variation du degré d'exigence au niveau du rendu final en raison de budgets variables, de jeunes entreprises émergent sur ce climat concurrentiel tendu. La valeur ajoutée de Mikros Image reste sa capacité à mener à bien des projets complexes, dans un temps imparti et avec un budget donné.

Chapitre 2

Introduction

Le problème de la reconstruction 3D par stéréo-vision à partir de caméras multiples calibrées capturant une scène fixe est étudié depuis plusieurs décennies. Les travaux initiés par (BEARDSLEY et al. 1996) puis étendus par (POLLEFEYS et al. 2000 ; BROWN et LOWE 2005a ; SNAVELY et al. 2006) ont démontré qu'il est possible d'estimer de manière séquentielle la *structure à partir du mouvement* d'une ou plusieurs caméras. Le principe de reconstruction, découlant de travaux de photogrammétrie, consiste à comparer les informations provenant de plusieurs images, prises de points de vue différents, pour identifier les positions et orientations de chaque cliché (le mouvement) puis la géométrie de la scène (la structure). Il a été démontré sur des jeux de données comportant une vérité terrain (STRECHA et al. 2008) que les résultats de reconstruction sont quantitativement comparables à des acquisitions lasers. Des erreurs de localisation de caméras de l'ordre du centimètre ou millimètre ont été mesurées.

Cependant l'application de ces mêmes méthodes séquentielles sur de larges jeux d'images n'est pas simple. Le passage à l'échelle n'est alors atteignable qu'en ayant recours à diverses approximations. Des implémentations massivement parallèles sont utilisées pour accélérer la recherche de correspondances entre images (AGARWAL et al. 2009). Des solutions dédiées pour l'ajustement non linéaire de paramètres sur GPU sont utilisées (WU et al. 2011a). Des localisations GPS approximatives peuvent être utilisées pour certaines images (CRANDALL et al. 2011) ou bien des informations GPS sont combinées avec des plans de cadastre issus de bases de données GIS (système d'information géographique) (STRECHA et al. 2010). Toutes ces approximations permettent effectivement de traiter des jeux de données de plus en plus grands, mais ce passage à l'échelle est réalisé au détriment de l'estimation de la position des caméras. Une précision moyenne de l'ordre du mètre est alors obtenue sur de larges jeux de données (CRANDALL et al. 2011 ; WU 2013).

Le principal défaut de ces méthodes séquentielles est l'accumulation d'erreurs due à la nature du processus. On observe des dérives lors de l'estimation des poses. Une trajectoire circulaire est ainsi souvent identifiée en spirale. Des méthodes considérant les poses de caméras de manière globale ont été développées (OLSSON et ENQVIST 2011 ; MARTINEC et PAJDLA 2007) pour supprimer ce phénomène de dérive, mais une fois de plus le passage à l'échelle n'est pas aisé.

Ces méthodes d'estimation de pose et orientation de caméras possèdent des limitations sur les points suivants :

- la robustesse,
- la précision,
- le passage à l'échelle.

Nous proposons dans cette thèse des solutions alternatives pour chacune de ces limitations. Nous démontrerons l'impact positif des solutions proposées, en termes de performances quantitatives et de temps de calcul.

2.1 Organisation et contributions du manuscrit

Cette thèse concentre son étude sur l'estimation de structure à partir du mouvement (SfM) dans le cadre d'une application pour la postproduction audiovisuelle et plus particulièrement la reconstruction de décors. Elle se focalise sur l'estimation précise de poses des caméras afin d'obtenir la meilleure représentation 3D possible de l'environnement photographié.

La thèse s'articule autour de contributions sur les axes suivants :

- le suivi de points saillants dans des images non ordonnées,
- la généralisation de l'utilisation d'un estimateur robuste statistique de modèles paramétriques,
- la vérification de l'impact, à large échelle, d'estimateurs robustes adaptatifs dans les méthodes de calibration séquentielles,
- la robustesse et le passage à l'échelle pour l'estimation globale de la position d'un réseau de caméras,
- l'harmonisation colorée d'un ensemble d'images multiple-vues.

2.1.1 Contributions théoriques

Les contributions théoriques sont axées sur :

1. Une généralisation des travaux de MOISAN et STIVAL (2004) et RABIN (2009) :

   Nous proposons de généraliser le cadre d'estimation robuste de modèle paramétrique défini par MOISAN et STIVAL (2004) afin de pouvoir utiliser cette estimation robuste adaptative à la reconstruction 3D. Nous montrerons que la formulation générique permet de réaliser des estimations de poses relatives, de matrices de projection, de tenseurs tri-focaux. Nous proposerons des travaux préliminaires pour explorer une paramétrisation *a contrario* d'erreurs angulaires appliquée à l'estimation de pose relative d'images sphériques.

2. L'utilisation d'optimisation convexe pour garantir l'obtention d'un minima global :

   Nous proposons de réaliser par minimisation convexe la fusion de translations relatives dans un repère global commun sous norme l_∞ afin de calibrer globalement en position un réseau de caméras. Nous présenterons finalement un ajustement d'histogrammes via une déformation linéaire sous norme l_∞ pour l'harmonisation colorée d'une séquence d'images.

2.1.2 Contributions appliquées

Dans le cadre d'un projet de recherche et innovation, «Mimatte3D», nous avons développé une chaîne de reconstruction 3D prenant en compte les besoins métiers liés à la postproduction audiovisuelle. Des outils permettant à l'utilisateur de guider le processus automatique de reconstruction ont été réalisés (cf. figure 2.1). Ce projet subventionné, OSEO-CNC-RIAM - 2012, a été réalisé par Benoit MAUJEAN, Bruno DUISIT,

Pierre MOULON et Christophe COURGEAU. Ce projet s'implique dans la vision industrielle liée à cette thèse CIFRE.

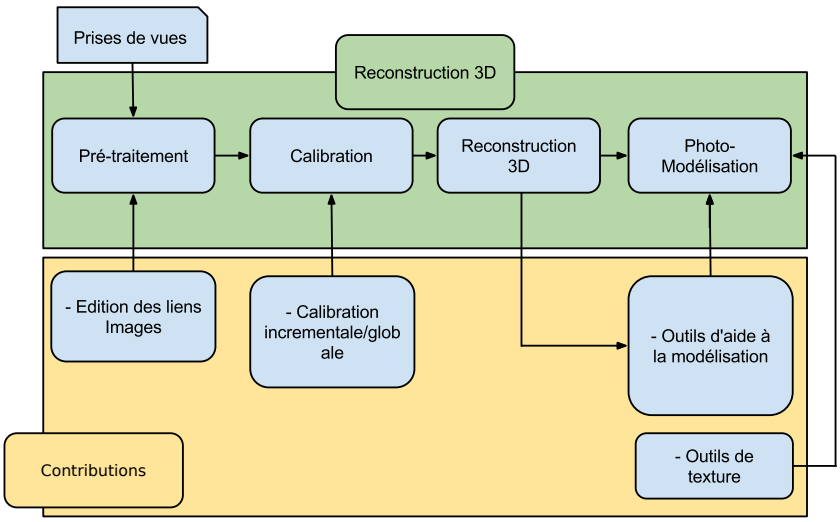


FIGURE 2.1 – Chaîne de traitement MiMatte3D pour l'aide à la construction de décors.

La chaîne «Mimatte3D» propose une suite d'outils pour l'aide à la reconstruction de décor virtuel photo-réaliste à partir de photos s'intégrant le plus facilement aux outils métiers des *mattes-painters*, les créateurs de décors. Le projet débouche sur les outils suivants :

- MILINK : un outil de visualisation et édition de réseau de connections d'images pour l'aide à la suppression de mises en correspondances d'images aberrantes,
- MICALIB : une chaîne de calibration externe séquentielle et une chaîne de calibration globale,
- MIMODE : une interface d'aide à la photo-modélisation,
- MIMATTEIMPORTER : une interface de la géométrie de calibration pour le logiciel d'édition 3D Autodesk Maya (caméras, nuages de points, plan images),
- MIPROJCAM : un outil de projection de texture sur de la géométrie pour le logiciel d'édition 3D Autodesk Maya.

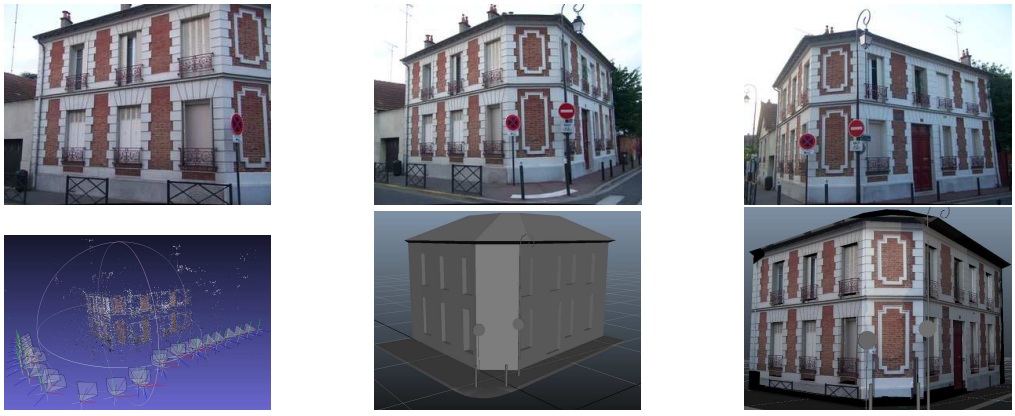


FIGURE 2.2 – Haut : une partie des images utilisées pour la reconstruction photo-réaliste. Bas : la calibration externe, la photo-modélisation et la restitution visuelle.

Cette chaîne réalisée pour le compte de Mikros Image est basée sur la librairie open source OpenMVG réalisée pendant cette thèse.

Les figures 2.3, 2.4 illustrent des exemples de reconstruction de décors réalisés à partir de collections d'images dans la phase finale du projet OSEO-CNC-RIAM - 2012 :

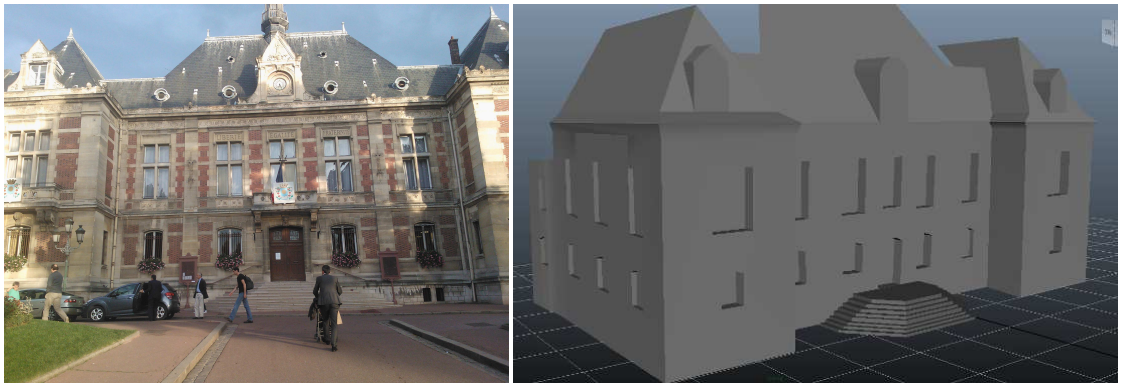


FIGURE 2.3 – Modélisation de la Mairie de Montrouge réalisée à partir de 20 images acquises à partir d'un téléphone mobile.

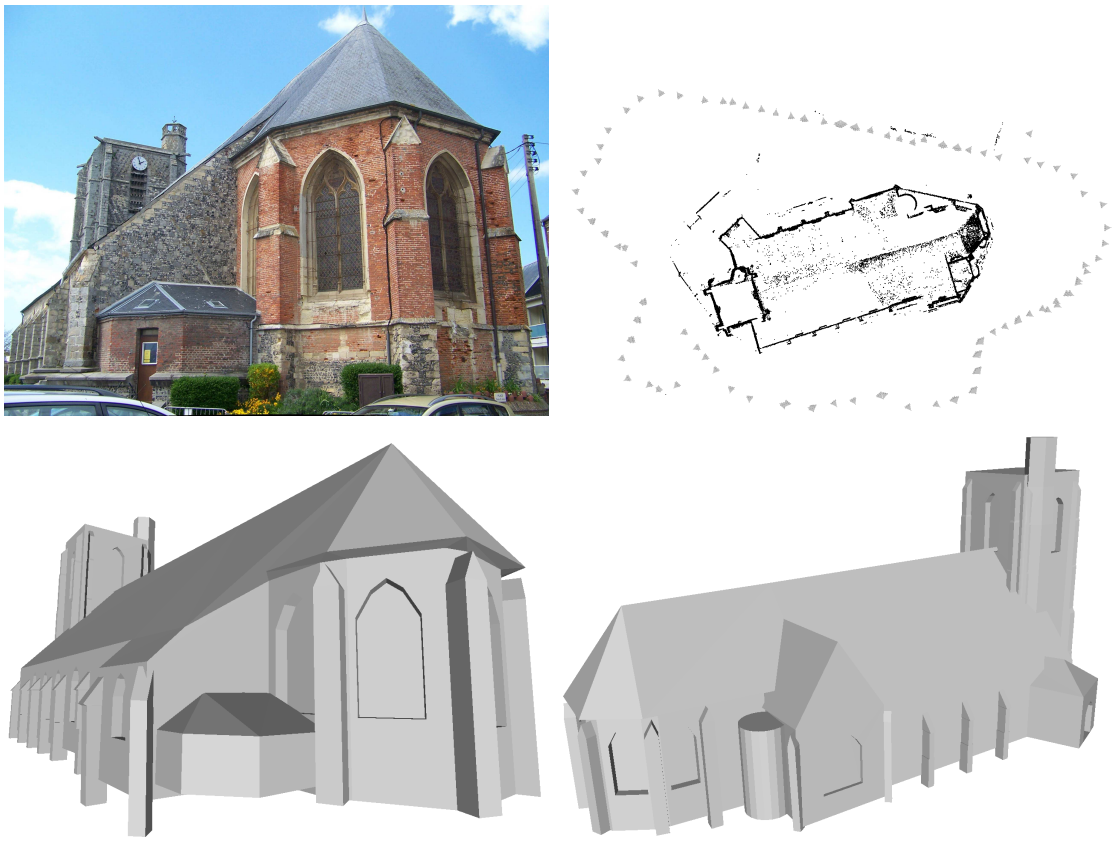


FIGURE 2.4 – Modélisation de l'église du village de Ault réalisée à partir de 109 images.

2.1.3 Contributions logicielles

– PPT-GUI (Python Photogrammetry Toolbox)
Ce projet co-développé en collaboration avec Alessandro Bezzi et Luca Bezzi d'ARC-TEAM permet un accès simplifié à des outils open-source de photogrammétrie (Bundler (SNAVELY et al. 2006), CMVS (FURUKAWA et al. 2010), PMVS (FURUKAWA et PONCE 2010)) sous les systèmes d'exploitation Linux et Windows. Ce projet est intégré à la distribution Linux ArcheOS dédiée aux archéologues et utilisé avec succès pour de nombreux projets de reconstitution faciale (cf. figure 2.5).

FIGURE 2.5 – Projets de reconstitution faciale menés par Cicero Moraes à partir de photographies de crânes reconstitués en 3D via l'utilisation de PPT-GUI et Blender.

– openMVG
openMVG (Open-source MultipleViewGeometry) est une bibliothèque C++ open-source conçue pour la recherche reproductible en vision par ordinateur. Elle fournit une implémentation de l'état de l'art et un accès facilité aux outils communs utilisés en géométrie multi-vues. La bibliothèque est multiplateforme, peut être compilée sous Windows, Linux, MacOS et est utilisable sur des architectures de type ARM (Apple Iphone iOS). Cette bibliothèque contient le code relatif aux publications CVMP2012, ACCV 2012 et IPOL 2012. La qualité de ce projet est évaluée dans le temps par une machine d'intégration continue et une série de tests unitaires garantissant la non régression des fonctionnalités délivrées.

2.1.4 Participation à la vie scientifique

Récompenses :

- Le 31/10/2011 le groupe Imagine remporte le premier prix du *PROVISG Mars 3D Challenge*. La compétition portant sur 3 thématiques :
  1. la reconstruction de cartes de disparité,
  2. la reconstruction de trajectoires de caméras à partir d'images acquises par un robot,
  3. la reconstruction 3D de la géométrie de la scène observée par le robot (cf. figure 2.6),

  nous a désignés vainqueurs parmis les 6 équipes participantes. Les évaluations ont été réalisées sur des images du CNES et du robot martien «MER Mars Exploration Rovers», fournies par le comité organisateur (le CMP (Center for Machine Perception) de l'université CTU de Prague).

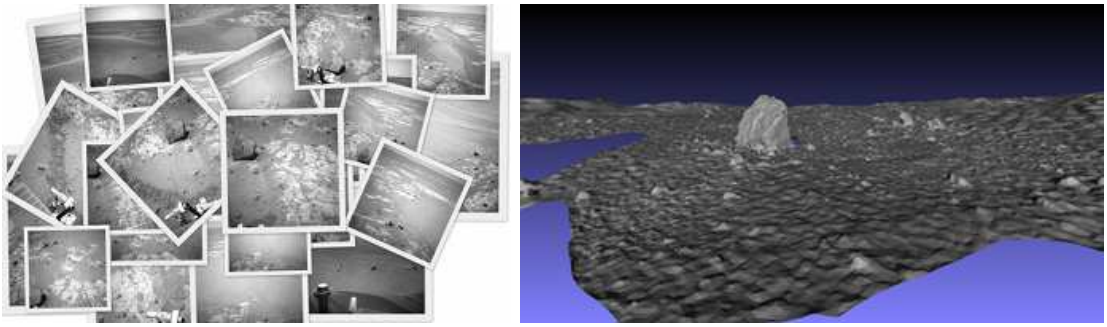


FIGURE 2.6 – Exemple de la reconstruction 3D d'un artefact de la planète Mars que nous avons pu réaliser à partir d'images du robot MER fournies par le comité organisateur.

Les résultats ont été présentés à la conférence ICCV en 2011 au groupe de travail «CVVT :E2M – Computer Vision in Vehicle Technology : From Earth to Mars» et publié dans le journal de la conférence. Je tiens à remercier David Ok, Victoria Rudakova et Pascal Monasse sans qui cette aventure n'aurait pas été couronnée de succès et aussi Gerhard Paar, Rongxing (Ron) Li et Tomas Pajdla pour leur accueil à l'université de Columbus et au JPL Nasa pour la présentation des résultats.

- Vainqueur du prix NVIDIA pour le meilleur papier court à la conférence CVMP 2013 pour le travail intitulé 'Global Multiple View Color Consistency'.

Encadrement de stages :

- Badis Djellab étudiant ENPC :

  Stage réalisé au laboratoire Imagine sur l'estimation multi-modèle en utilisant les méthodes de J-Linkage (TOLDO et FUSIELLO 2008). Étude de l'impact de la suppression *a contrario* de modèle non-significatif pour accélérer et estimer avec plus de précision le nombre de modèles à identifier.

- Bruno Duisit étudiant Polytech Paris Sud (Université Paris XI) :

  Stage réalisé au sein de l'entreprise Mikros Image sur la modélisation 3D à partir d'une image dans le logiciel Maya et la réalisation d'une interface de visualisation et édition de graphes d'images.

- Tristan Faure et Luc Girod étudiants ENSG :

  Stage réalisé au laboratoire Imagine sur la mise en place d'un protocole d'acquisition de vérité terrain pour une évaluation des méthodes de photogrammétrie.

- Emmanuel Habbets étudiant ENSG :

  Stage réalisé au laboratoire Imagine sur le calcul et la fusion de cartes de disparité. Implémentation partielle d'une chaîne de traitement similaire aux travaux de TOLA et al. (2012).

- Rafaël Marini Silva étudiant de l'école polytechnique :

  Stage réalisé sur les méthodes de recherche de plus proches voisins parmi de larges collections d'images. Implémentation d'un moteur de recherche basé sur la quantification d'espace descriptif (JEGOU et al. 2011).

- Lucas Plaetevoet étudiant ENPC :

  Stage réalisé au laboratoire Imagine sur la fusion d'acquisition de nuage de points issus de capteur de type Kinect (lumière structurée).

Encadrement salarié :

- Bruno Duisit (Mikros Image) :

  Encadrement de Bruno sur la thématique du projet MiMatte3D (cf. section 2.1.2).

Relecteur :
Relecteur pour le journal en ligne IPOL http://www.ipol.im.

2.1.5 Publications de l'auteur

- **Revue internationale :**

  Automatic Homographic Registration of a Pair of Images, with A Contrario Elimination of Outliers. L Moisan, P Moulon, P Monasse. IPOL 2012.

- **Conférence internationale avec actes :**

  Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion. P Moulon, P Monasse, R Marlet. ICCV 2013.

  Adaptive Structure from Motion with a contrario model estimation. P Moulon, P Monasse, R Marlet. ACCV 2012.

- **Démonstration en conférence internationale :**

  Adaptive model estimation, a real time demonstration. P Moulon, P Monasse, R Marlet. ACCV 2012.

- **Conférence internationale avec comité de relecture :**

  Global Multiple-View Color Consistency. P Moulon, D Bruno, P Monasse. CVMP 2013. (Vainqueur du prix NVIDIA pour le meilleur papier court).

  Unordered feature tracking made fast and easy. P Moulon, P Monasse. CVMP 2012.

- **Conférence nationale avec comité de relecture :**

  La bibliothèque openMVG : open source Multiple View Geometry. P Moulon, P Monasse, R Marlet. Orasis, Congrès des jeunes chercheurs en vision par ordinateur 2013.

  Estimation robuste de modèles a contrario, impact sur la précision en structure from motion. Présentation orale. P Moulon, P Monasse, R Marlet. ISS France 2013.

- **Groupe de travail :**

  L'utilizzo di tecniche structure from motion e imagebased modelling in ambienti estremi. P Moulon, Nicolò Dell'Unto, A Bezzi, L Bezzi, Rupert Gietl. Low Cost 3D 2012.

  Python Photogrammetry Toolbox : A free solution for Three-Dimensional Documentation. P Moulon, A Bezzi. ArchoFoss 2011.

- **Bibliothèque open source :**

  OpenMVG Open-source MultipleViewGeometry 2012. https://github.com/openMVG/openMVG.

Chapitre 3

La géométrie multiples vues et l'estimation de mouvements

Lorsqu'une scène est photographiée sous plusieurs points de vue, la connaissance du déplacement apparent des éléments de la scène à travers la série d'images permet de retrouver le déplacement de l'appareil photographique et d'obtenir une représentation 3D de la scène observée : on parle alors de *SfM Structure from Motion*.

Ce chapitre présente :

1. le modèle projectif classique de caméra,
2. une série de relations géométriques formulées à partir de correspondances de points homologues entre photographies,
3. comment détecter et suivre des éléments à travers une série d'images.

Sommaire

3.1 Notations

Dans les sections suivantes nous allons travailler avec des coordonnées définies dans un espace cartésien ou projectif. En géométrie projective, les coordonnées homogènes rendent les calculs possibles dans l'espace projectif comme les coordonnées cartésiennes le permettent dans l'espace euclidien. Les coordonnées homogènes d'un point de l'espace projectif de dimension n (x,y,z,...) sont écrites habituellement comme un vecteur de longueur $n+1$ (x,y,z, ... , w). Deux ensembles de coordonnées qui sont proportionnels dénotent le même point d'espace projectif : pour tout scalaire non-nul c, $(cx, cy, cz, ..., cw)$ est équivalent à (x, y, z, w). La coordonnée $w = 0$ permet de représenter un élément à l'infini. Le passage de coordonnées homogènes à des cordonnées cartésiennes est réalisé en divisant les n premiers éléments par le $n+1^{\text{e}}$, soit w.

Un point 3D en coordonnées homogènes $\mathbf{X}_{4\times1} = \{\mathbf{X}(1), \mathbf{X}(2), \mathbf{X}(3), \mathbf{X}(4)\}$ est représenté en coordonnées cartésiennes $X_{3\times1}$. Pour tout W différent de 0, on obtient l'équation :

$$X = (\mathbf{X}(1)/W, \mathbf{X}(2)/W, \mathbf{X}(3)/W)^T \qquad \mathbf{X} \sim (\widetilde{X}, \widetilde{Y}, \widetilde{Z}, 1)^T, \tag{3.1}$$

avec $\sim$ définissant l'égalité à une échelle $\frac{1}{W}$ près avec $W = \mathbf{X}(4)$.

De la même manière, un point 2D en coordonnées homogènes $\mathbf{x}_{3\times1} = \{\mathbf{x}(1), \mathbf{x}(2), \mathbf{x}(3)\}$, est en relation avec son équivalent en coordonnées cartésiennes $x_{2\times1}$:

$$x = (\mathbf{x}(1)/w, \mathbf{x}(2)/w)^T \qquad \mathbf{x} \sim (\widetilde{x}, \widetilde{y}, 1)^T \tag{3.2}$$

avec $\sim$ définissant l'égalité à une échelle $\frac{1}{w}$ près avec $w = \mathbf{x}(3)$.

Notations complémentaires

x	Un point en coordonnées cartésiennes
$\mathbf{x}$	Un point en coordonnées homogènes
$\{A\}$	Une liste d'éléments de type A
$[A]$	Une liste ordonnée d'éléments de type A
(A, B)	Couple d'éléments associés, ici un 2-uplet
R	Matrice de rotation
$\mathbf{t}$	Vecteur de translation
C	Position du centre de projection d'une caméra
$\mathbf{K}$	Matrice des paramètres intrinsèques d'une caméra
X_j	Point 3D d'index j
x_j^i	Projection du point 3D X_j dans l'image i
tr	Déplacements apparents des projections des points $\{X_j\}_j$ dans une série d'images
G_A	Graphe entre élements de type A
R_i^k	k^{e} ligne de la matrice de rotation de l'image i
$\mathbf{t}_i^k$	k^{e} composante du vecteur de translation de l'image i
$x_j^i(k)$	k^{e} composante du point x_j^i

3.2 La géométrie caméra

Un sténopé modélise un appareil photographique comme un système réalisant la projection centrale d'une scène en 3 dimensions en 2 dimensions. Une image est le résultat d'une intégration de rayons lumineux observés sur une surface sensible durant un court intervalle de temps à travers une série de lentilles. Lorsque ce système optique est approximé par une seule lentille, on obtient un modèle simple de projection perspective, le modèle sténopé. Dû au fait que l'ensemble des rayons lumineux observés passe par un seul et unique point (le centre de projection) ces caméras sont souvent appelées *pinhole*.

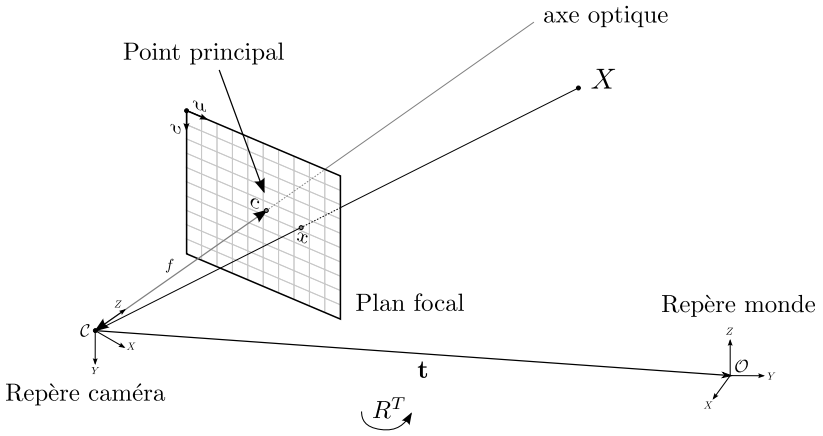


FIGURE 3.1 – Illustration d'une caméra sténopé. Un point 3D $\mathbf{X}$ est projeté en $\mathbf{x}$ sur un plan image par une projection centrale. On appelle paramètres extrinsèques, la transformation rigide $[R|\mathbf{t}]$ entre le repère monde O et la position de la caméra $\mathcal{C}$. Pour simplifier le plan image est ici montré en avant du centre optique $\mathcal{C}$. Sur une caméra réelle ce plan image est situé derrière le centre optique et tourné à 180°.

Ce système d'acquisition peut être réduit à deux composantes principales :
- **un système optique** permettant de réaliser la projection de la scène observée sur un plan focal : $\mathbb{R}^3 \rightarrow \mathbb{R}^2$.
- **une surface photo-sensible** qui capte les densités de photons par pixels pour former une image.

Une caméra sténopé transforme un point $\mathbf{X}$ de $\mathbb{R}^3$ en un point image $\mathbf{x}$ de $\mathbb{R}^2$ à travers deux opérations :

Un changement de repère. Soit $\mathbf{X}_c$ un point monde défini dans le repère de la caméra :

$$\mathbf{X}_c = \begin{bmatrix} R & \mathbf{t} \\ 0 & 1 \end{bmatrix} \mathbf{X} \tag{3.3}$$

Cette relation de passage entre le repère monde et le repère local caméra dépend de 6 degrés de liberté que l'on appelle les paramètres extrinsèques :
- 3 degrés de liberté pour l'orientation de la caméra : une matrice de rotation $R_{3\times3}$,
- 3 degrés de liberté pour la translation, décrite par le vecteur $\mathbf{t}_{3\times1}$, $\mathbf{t}$ représente la position de l'origine monde O dans le repère caméra. La position $\mathcal{C}$ du centre optique de la caméra est donc $\mathcal{C} = -R^T\mathbf{t}$.

Une projection et une mise à l'échelle. La transformation réalisée par l'optique et la géométrie du capteur est modélisée par les paramètres intrinsèques définis par le biais de 6 paramètres par une matrice $\mathbf{K}_{\times}3$. Cette matrice dite de calibration ou bien calibrage, peut s'écrire comme suit :

$$\mathbf{K} = \begin{bmatrix} fk_u & s & c_u \\ & fk_v & c_v \\ & & 1 \end{bmatrix} \tag{3.4}$$

Soit :
– f la distance focale, distance du centre optique au plan focal,
– s, k_u et k_v des facteurs d'échelles,
– c : (c_u, c_v) le point principal modélise le décalage de l'origine,
Un point 3D en repère caméra $\mathbf{X}_c$ a pour correspondant $\mathbf{x}$ image :

$$\mathbf{x} = [\mathbf{K}|0]\mathbf{X}_c = \mathbf{K}\begin{bmatrix} R & \mathbf{t} \\ 0 & 1 \end{bmatrix}\mathbf{X} \tag{3.5}$$

Ainsi un point $\mathbf{x}$ en repère image est transformé en un rayon en repère caméra $\hat{\mathbf{x}}$ comme suit :

$$\hat{\mathbf{x}} = K^{-1}\mathbf{x} \tag{3.6}$$

Pour simplifier nous utiliserons c au centre de l'image de taille $w \times h$ et des pixels carrés sur la surface photo sensible ($k_u = k_v = 1$ et $s = 0$) :

$$\mathbf{K} = \begin{bmatrix} f & & w/2 \\ & f & h/2 \\ & & 1 \end{bmatrix} \tag{3.7}$$

Finalement ces deux transformations peuvent être combinées en une seule opération matricielle. Un point $\mathbf{X}$ exprimé dans le repère monde est donc relié à sa projection image $\mathbf{x}$ par la formule suivante :

$$\mathbf{x} = \mathbf{P}\mathbf{X} \tag{3.8}$$

Avec $\mathbf{P} = \mathbf{K}[R|\mathbf{t}]$ une matrice de projection de taille 3×4.

On néglige ici la distorsion causée par l'optique. Nous invitons le lecteur à consulter les travaux de (BROWN 1966) pour plus de détails.

3.3 La géométrie à 2 vues

Homographie

Lorsqu'un objet est plan, il est possible de définir une transformation exacte entre les points homologues x et x'. La transformation la plus générale pour ce couple de points (x,x') est appelée *homographie*. L'homographie désigne une classe de transformations projectives qui conservent les alignements. Si tous les points appartiennent à un même plan, alors les projections obtenues dans les images conservent leur alignement (cf. figure 3.2). L'image d'une ligne reste donc une ligne.

La fonction de passage entre les coordonnées de l'observation dans l'image gauche et droite (x,x'), est définie par une transformation homographique $\mathbf{H}$.

$$\mathbf{x}' = \mathbf{H}\mathbf{x} \tag{3.9}$$

$\mathbf{H}$ est une matrice de transformation :
– projective 2D linéaire conservant les alignements,
– inversible entre les plans projectifs. On peut donc écrire :

$$\mathbf{x} = \mathbf{H}^{-1}\mathbf{x}'. \tag{3.10}$$

– de taille 3×3 définie à un facteur près qui lui confère 8 degrés de liberté.

La matrice $\mathbf{H}$ étant définie à un facteur d'échelle près et chaque couple de points homologues fournissant 2 équations indépendantes sur $\mathbf{H}$, quatre points homologues sont nécessaires pour définir de manière unique les huit paramètres indépendants de l'homographie. Certaines configurations de points peuvent mener à des cas dégénérés, nous invitons le lecteur à lire MOISAN et al. (2012) pour les détecter.

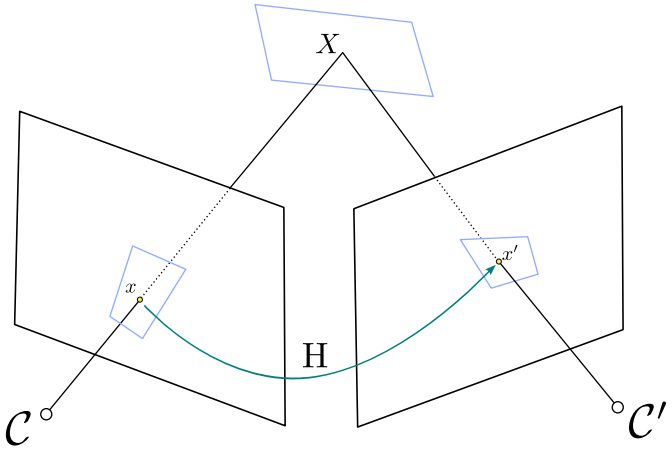


FIGURE 3.2 – La relation homographique établit une relation point à point entre les images d'une surface plane dans plans images.

Note : Lorsqu'une caméra effectue un mouvement de rotation autour de son centre optique (le point nodal), les images acquises sont reliées par des homographies. Ce mouvement particulier de caméra permet la construction d'images panoramiques.

La matrice essentielle et la géométrie épipolaire

La notion de matrice essentielle **E** a été proposée par LONGUET HIGGINS (1981). La matrice **E** modélise le changement de repère entre deux caméras : une rotation R et une translation $\mathbf{t}$.

Cette géométrie repose sur la géométrie épipolaire qui associe à un point x une droite $l'(x)$, notée l' pour simplification. Cette droite, dite ligne épipolaire, est située à l'intersection du plan image droit et du plan épipolaire défini par les points $\mathcal{C}$, $\mathcal{C}'$ et x. l' est ainsi l'observation du rayon $\overrightarrow{CX}$ par la caméra $\mathcal{C}'$. Toutes les lignes épipolaires ont un point commun, la projection du centre optique de l'autre caméra. Ces points sont appelés épipôles et sont notés e, e' respectivement pour la caméra gauche et droite.

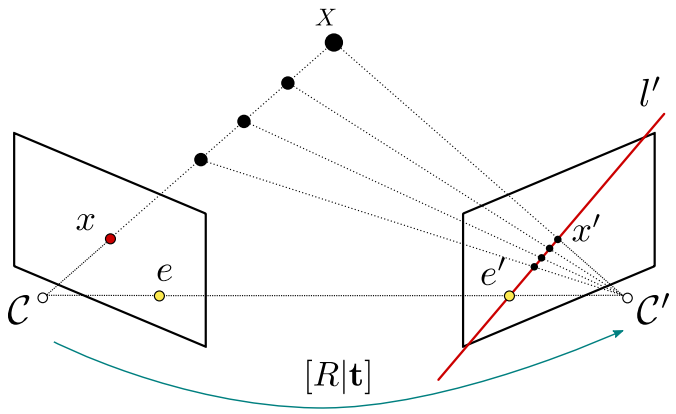


FIGURE 3.3 – Relation épipolaire entre deux images. Étant donné une projection x d'un point 3D X, sa projection x' dans l'image droite est restreinte à la ligne épipolaire correspondante l'.

La contrainte épipolaire est traduite par l'utilisation de la matrice essentielle **E**. Cette matrice permet de mettre en relation une correspondance de point entre deux images. Soit deux caméras $\mathbf{P} = [I|0]$ et $\mathbf{P}' = [R|\mathbf{t}]$. Étant donné un point X en coordonnée euclidienne dans le repère de la caméra gauche, sa position dans le repère de la caméra droite est :

$$X' = RX + \mathbf{t} \tag{3.11}$$

Une pré-multiplication par $X^T[\mathbf{t}]_\times$ permet d'obtenir :

$$X^T[\mathbf{t}]_\times RX' = X^T EX' \tag{3.12}$$

ou $\mathbf{E} \sim [\mathbf{t}]_\times R$ est une matrice 3×3 et $\mathbf{t}$ est un vecteur de taille 3. $[\mathbf{t}]_\times$ est la représentation matricielle du produit vectoriel (*cross product matrix*). Il est intéressant de noter que la relation 3.12 est aussi utilisable avec des correspondances en repère caméra :

$$\hat{\mathbf{x}}'^T E\hat{\mathbf{x}} = 0 \tag{3.13}$$

E possède 5 degrés de liberté car elle dépend seulement de R et de la direction de translation $\mathbf{t}$. Multiplier $\mathbf{t}$ par un facteur d'échelle revient à multiplier **E** par le même facteur, ce qui exprime les mêmes contraintes.

La matrice essentielle peut être identifiée à partir de 8 correspondances si l'on utilise la formulation générale $\mathbf{x}'^T\mathbf{E}\mathbf{x} = 0$ sans contraintes de structure de la matrice **E** et 5 points si l'on utilise des matrices de calibration connues cf. (NISTÉR 2004 ; LI et HARTLEY 2006).

La matrice fondamentale

La géométrie épipolaire a notamment été étudiée par LUONG (1992) et FAUGERAS (1992). Pour toute correspondance entre deux images on peut reprendre l'équation 3.13 :

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

et l'écrire en considérant des points images (en coordonnées image, pixels) :

$$\begin{aligned} (\mathbf{K}_r^{-1}\mathbf{x}')^T E (\mathbf{K}_l^{-1}\mathbf{x}) &= 0, \\ \mathbf{x}'^T (\mathbf{K}_r^{-T} E \mathbf{K}_l^{-1}) \mathbf{x} &= 0, \\ \mathbf{x}'^T \mathbf{F} \mathbf{x} &= 0, \end{aligned} \tag{3.14}$$

ou $\mathbf{F} \sim \mathbf{K}'^{-T} E \mathbf{K}^{-1}$ est la matrice fondamentale. $\mathbf{F}$ est définie de taille 3×3 à un facteur multiplicatif près, de rang 2, ce qui lui confère donc 7 degrés de liberté. Cette matrice établit une relation point-ligne tout comme la matrice essentielle.

La matrice $\mathbf{F}$ peut être estimée à partir de 8 correspondances HARTLEY (1997a) ou à partir de 7 correspondances en forçant a posteriori la contrainte de rang (TORR et MURRAY 1997). Cette dernière méthode est la solution dite minimale et identifie de 1 à 3 solutions pour un échantillon de 7 correspondances. La géométrie épipolaire fournie par la matrice $\mathbf{F}$ est particulièrement intéressante car elle établit une relation entre des points en géométrie image. Les paramètres intrinsèques ne sont pas nécessaires pour vérifier la consistance géométrique d'une paire de points.

Les épipôles étant les points d'intersection de toutes les droites épipolaires, ils définissent aussi le noyau de $\mathbf{F}$: $\mathbf{F}e = 0$ et $\mathbf{F}^T e' = 0$.

3.4 La géométrie à 3 vues

La géométrie d'un triplet d'images peut être représentée par un tenseur tri-focal $\mathbf{T}$ (HARTLEY 1997b). $\mathbf{T}$ définit les relations épipolaires entre trois vues indicées i, j, k par une matrice cube de taille $3 \times 3 \times 3$. Un point x est mis en correspondance avec ses deux lignes épipolaires correspondantes : l', l'' (cf. figure 3.4) :

$$\sum_{ijk} \mathbf{x}(i) l'_j l''_k \mathbf{T}_i^{jk} \tag{3.15}$$

Ce tenseur est une généralisation du concept de la matrice fondamentale à un ensemble de trois vues. Soit trois matrices de projections : $\mathbf{P}_1 = [Id|0]$, $\mathbf{P}_2 = [a_j^i]$ et $\mathbf{P}_3 = [b_j^i]$. Le tenseur trifocal est défini ainsi :

$$\mathbf{T}_i^{jk} = a_i^j b_4^k - a_4^j b_i^k \mid i, j, k = 1, 2, 3 \tag{3.16}$$

avec ij une entrée de la matrice tel que i désigne la ligne et j la colonne.

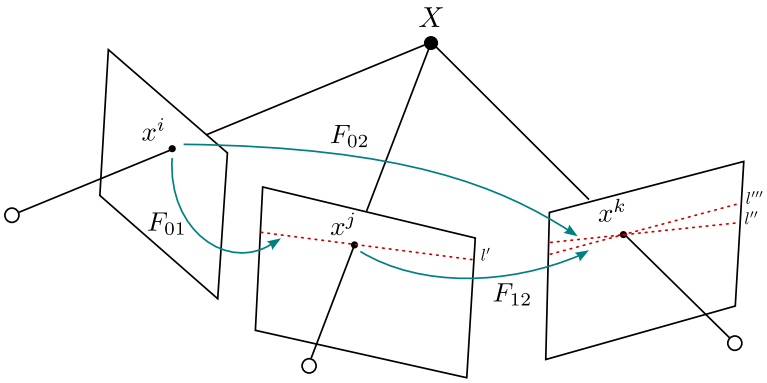


FIGURE 3.4 – Les relations épipolaires existantes au sein d'un tenseur tri-focal $\mathbf{T}$.

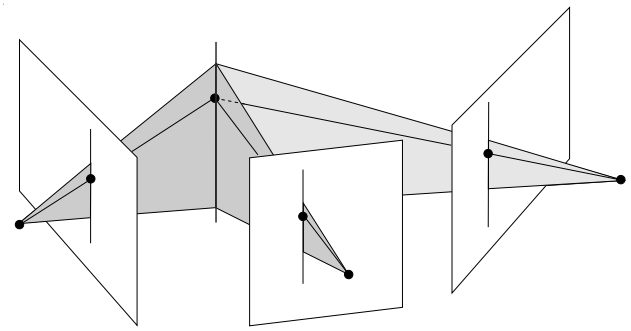
Une des propriétés les plus importantes du tenseur tri-focal est que la formulation du tenseur permet d'établir des relations linéaires entre des lignes et points en correspondances entre les trois images. Des contraintes tri-linéaires sont exprimables pour les relations suivantes :

ligne-ligne-ligne

point–ligne–ligne

point–ligne–point

point–point-point



Une autre particularité du tenseur tri-focal par rapport aux tenseurs à deux vues (bi-focaux), telle que la matrice fondamentale, est la relation de transfert. Cette relation de transfert permet d'identifier des points manquants lors de mises en correspondances.

Supposons qu'une correspondance est connue ($x^i \leftrightarrow x^j$) mais que le point correspondant dans la troisième image x^k ne l'est pas. L'utilisation de la contrainte point-ligne-ligne pour x^i et x^j identifie alors par intersection le point x^k. Il est situé à l'intersection des deux lignes épipolaires l'' et l'''.

3.5 La triangulation

La triangulation est le procédé de calcul d'un point 3D X d'après ses observations images x^i et des matrices de projections $\mathbf{P}_i$. Idéalement le point 3D X est situé à l'intersection des rayons $\overrightarrow{C_i x^i}$. Étant donné que les données sont le plus souvent bruitées, les rayons ne s'intersectent pas en pratique : cf. figure 3.5. Le candidat X est alors choisi comme le point ayant les plus faibles erreurs de re-projection entre les projetés images $\mathbf{P}_i(X)$ et les points de mesures x^i :

$$\underset{\mathbf{X}}{\text{minimiser}} \sum_{i=1}^{n} \left\| \mathbf{x}^i - \mathbf{P}_i \mathbf{X} \right\|_2 \tag{3.17}$$

avec n le nombre de vues considérées. Cependant une formulation directe menant à cette solution n'est pas évidente à mettre en place. Souvent une solution approximative est calculée, puis les erreurs résiduelles sont minimisées par itérations de l'algorithme de *Levenberg-Marquardt* (cf. (HARTLEY et ZISSERMAN 2000)).

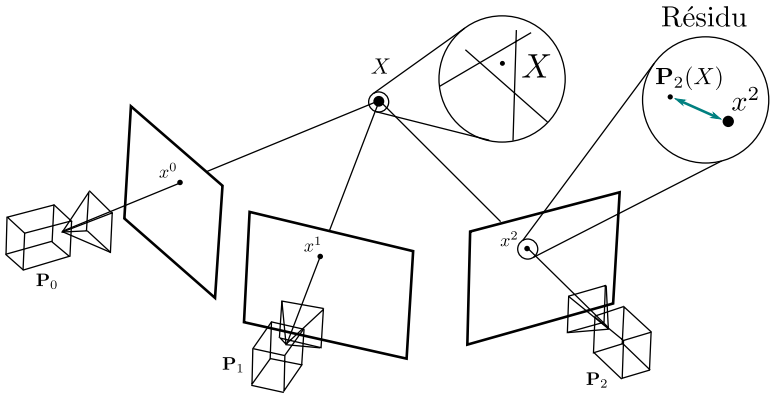


FIGURE 3.5 – La triangulation : Calcul du point X en fonction des caméras $\mathbf{P}_i$ et re-projections images. A cause du bruit de mesure, le point X n'est pas aisé à identifier.

On note dans la littérature une série de méthodes pour trouver un candidat X dans le cas à deux vues (cf. figure 3.6) :

Point milieu On recherche un point X situé sur le segment le plus court entre les deux rayons $\overrightarrow{C_i x^i}$. Ce segment est identifié grâce à la droite orthogonale aux deux rayons considérés. Le point 3D recherché est alors situé au milieu de ce segment.

Optimisation linéaire ou dite de minimisation algébrique. Avec l'utilisation des coordonnées homogènes on peut utiliser le fait que les vecteurs x^i sont colinéaires à $\mathbf{P}_i X$ pour écrire :

$$[\mathbf{x}^i]_\times \mathbf{P}_i \mathbf{X} = 0 \tag{3.18}$$

Cette équation peut être réécrite sous une forme solvable aux moindres carrés :

$$AX = 0 \tag{3.19}$$

avec A une matrice $3n \times 4$, n le nombre de vues et X un point visible (situé devant les caméras). La solution en coordonnée homogène est calculée en minimisant $||AX||$ sujet à $||X|| = 1$ pour éviter la solution triviale $X = 0$ (cf. Triangulation DLT (*Direct Linear Transform*) : HARTLEY et ZISSERMAN (2000)).

Optimisation itérative une solution initiale est identifiée puis optimisée de manière itérative (LINDSTROM 2010).

Optimale le point X est recherché en minimisant une erreur géométrique : les erreurs résiduelles. On minimise l'équation 3.17 directement (KANATANI et al. 2008). Cette méthode est dite méthode *gold-standard* (cf. (HARTLEY et ZISSERMAN 2000)).

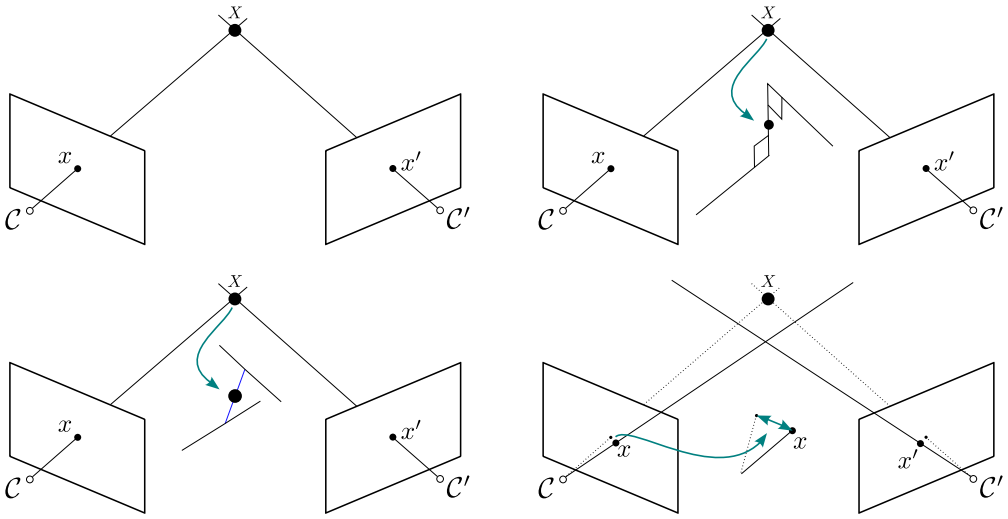


FIGURE 3.6 – De gauche à droite, de haut en bas. Situation théorique, point milieu, méthode linéaire, méthode optimale.

Limitations

Les méthodes DLT sont le plus couramment utilisées dans le cas de la triangulation à n vues, car identifier une solution optimale de manière directe implique des équations complexes qui sont non triviales à résoudre. Des solutions optimales ont été exprimées pour le cas à deux et trois vues (BYRÖD et al. 2007), mais pas au delà.

Perspectives

HARTLEY et SCHAFFALITZKY (2004a) proposent une formulation quasi-convexe du problème qui permet d'identifier X par minimisation de la norme l_∞ des erreurs résiduelles. Cette formulation minimisée par bissection permet de vérifier l'existence d'une solution et de garantir que la solution calculée est optimale par rapport au critère d'ajustement utilisé (la norme l_∞ des erreurs de re-projection).

3.6 L'estimation de pose

Étant donné des correspondances entre des points 3D X_j et les points images 2D x_j, on cherche à identifier la matrice de caméra $\mathbf{P}$ optimale (cf. figure 3.7). On recherche ainsi la pose (orientation et position) de la caméra qui fait que les rayons $\overrightarrow{CX_j}$ passent au plus près possible des m points 2D x_j projections des X_j.

$$\underset{\mathbf{P}}{minimise} \sum_{j=0}^{m} \left\| \mathbf{x}_j - \mathbf{P}\mathbf{X}_j \right\|_2 \tag{3.20}$$

Ce problème appelé *Perspective-n-Point* est traité en fonction du nombre de degrés de liberté de la pose :

1. Le cas non calibré :
   la matrice $\mathbf{P}$ de taille 3×4 est à identifier. 12 degrés de liberté sont à estimer. Une formulation linéaire de l'équation (3.20) permet de trouver aux moindres carrés une matrice $\mathbf{P}$ possible à partir de 6 correspondances 2D-3D (HARTLEY et ZISSERMAN 2000).
2. Le cas calibré (la matrice de calibration $\mathbf{K}$ est connue) :
   6 degrés de liberté sont ainsi à identifier : 3 pour l'orientation R et 3 pour la position $\mathbf{t}$ de la caméra. La connaissance a priori de la matrice de calibration permet de réduire le nombre de correspondances nécessaires. Trois correspondances sont suffisantes pour identifier un ensemble de solutions possibles (GAO et al. 2003 ; KNEIP et al. 2011). Le lecteur est invité à consulter LEPETIT et al. (2009) pour une liste plus exhaustive de différentes méthodes qui, à partir de n correspondances, estiment la pose $[R|\mathbf{t}]$ de la caméra (méthodes PnP (*Perspective-n-Point*)).

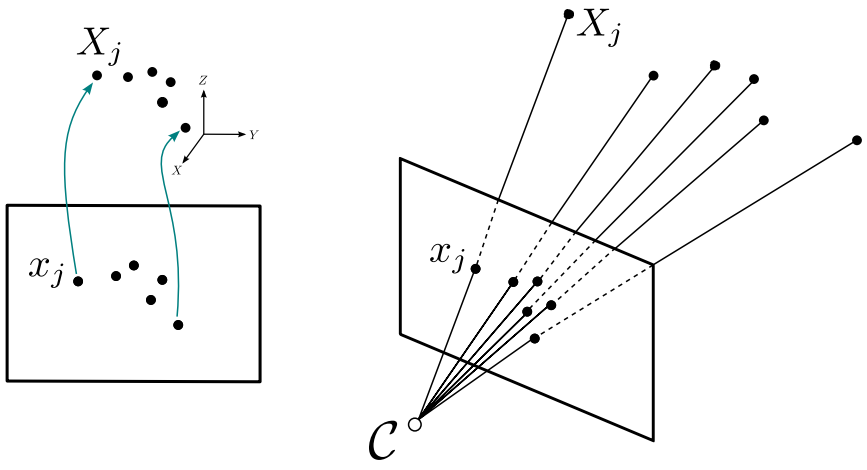


FIGURE 3.7 – L'estimation d'une pose de caméra (une orientation et position) est déterminée à partir de n correspondances 3D-2D.

3.7 L'ajustement de faisceaux

L'ajustement de faisceaux, *Bundle Adjustment*, est un processus d'optimisation non linéaire. On souhaite optimiser un vecteur de paramètres pour réduire une fonction d'objectif donné. Dans notre cas la fonction objectif vise à réduire les erreurs résiduelles de re-projection de la structure X_j aux mesures images x_j^i. x_j^i étant la projection du point 3D X_j dans l'image i. Le vecteur de paramètres est défini par une configuration initiale : les paramètres des caméras $\{\mathbf{P}_i\}_i$ et la structure de la scène $\{X_j\}_j$. Cette minimisation est réalisée par utilisation d'une procédure itérative, l'algorithme de *Levenberg-Marquardt*. Un vecteur p de départ représentant la configuration des paramètres est initialisé. A chaque itération, on remplace p par une nouvelle estimation $p+q$, q étant déterminé pour réduire la fonction objectif à minimiser. Lorsque la fonction objectif ne varie plus ou que le vecteur de paramètres est stable l'algorithme est arrêté. Une convergence vers la solution optimale est observée si le vecteur de départ n'est pas trop éloigné de la solution. Par contre, si la solution initiale est éloignée, une solution locale peut être identifiée.

Le problème d'ajustement de faisceaux est donc posé pour réduire la fonction coût suivante :

$$\underset{\{\mathbf{P}_i\}_i,\{\mathbf{X}_j\}_j}{minimise} \left\| \sum_{j=0}^{m} \sum_{i=0}^{n} \mathbf{x}_j^i - \mathbf{P}_i \mathbf{X}_j \right\|_2 \tag{3.21}$$

avec n le nombre de matrices de projection $\mathbf{P}$ et m le nombre de points 3D considérés dans la scène.

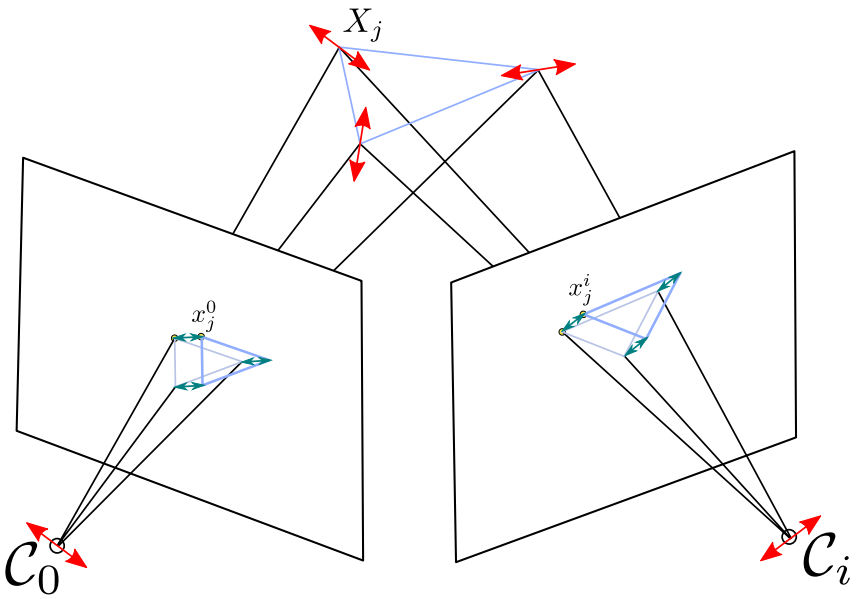


FIGURE 3.8 – L'ajustement de faisceaux : Une minimisation non linéaire des paramètres de projection des caméras et de la structure 3D de la scène est réalisée pour réduire les erreurs de re-projection résiduelles observé en domaine image.

Une synthèse complète sur l'ajustement de faisceaux est proposée par TRIGGS et al. (2000). Ce problème d'optimisation peut être résolu avec l'usage de matrices parcimonieuses (LOURAKIS et ARGYROS 2004). WU et al. (2011a) proposent une implémentation parallèle (GPU ou CPU) et AGARWAL et MIERLE (2012) proposent une implémentation parallèle générique pouvant utiliser une norme robuste pour éviter l'influence de mesures aberrantes : Ceres-solver. Ceres présente l'avantage d'être générique, la spécifica-

tion des variables et la fonction objectif intervenant dans un problème non-linéaire sont très facilement paramétrables, ce qui facilite l'implémentation de l'équation (3.21).

3.8 La géométrie multiples-vues et l'estimation de mouvements

Les techniques de structure à partir du mouvement, *Structure-from-Motion*, estiment le déplacement d'une caméra ou d'un appareil photographique et reconstruisent la structure de la scène à partir d'une séquence d'images. Soit la séquence d'images I_j, $j \in \{0,n\}$. Le procédé est le suivant : des éléments (ou primitives, par exemple des points saillants 2D) sont détectés puis suivis à travers l'ensemble des images : x_j^i. La visibilité des points 3D X_j image est ainsi connue. Les techniques de SfM cherchent alors à identifier des caméras $\mathbf{P}_i$ ainsi qu'une structure X_j représentant au mieux les données de visibilité x_j^i. C'est un problème d'optimisation ou l'on cherche à minimiser la somme des erreurs résiduelles en domaine image, l'équation (3.21), ou intervient n images et m points 3D. Ces erreurs résiduelles mesurent la précision de la reconstruction : l'adéquation entre la structure, les positions de caméra et les mesures images. Ces erreurs de re-projections, appelées erreurs résiduelles, représentent la distance entre les observations x_j^i images et la re-projection des points 3D $\mathbf{P}_i(X_j)$ reconstruit.

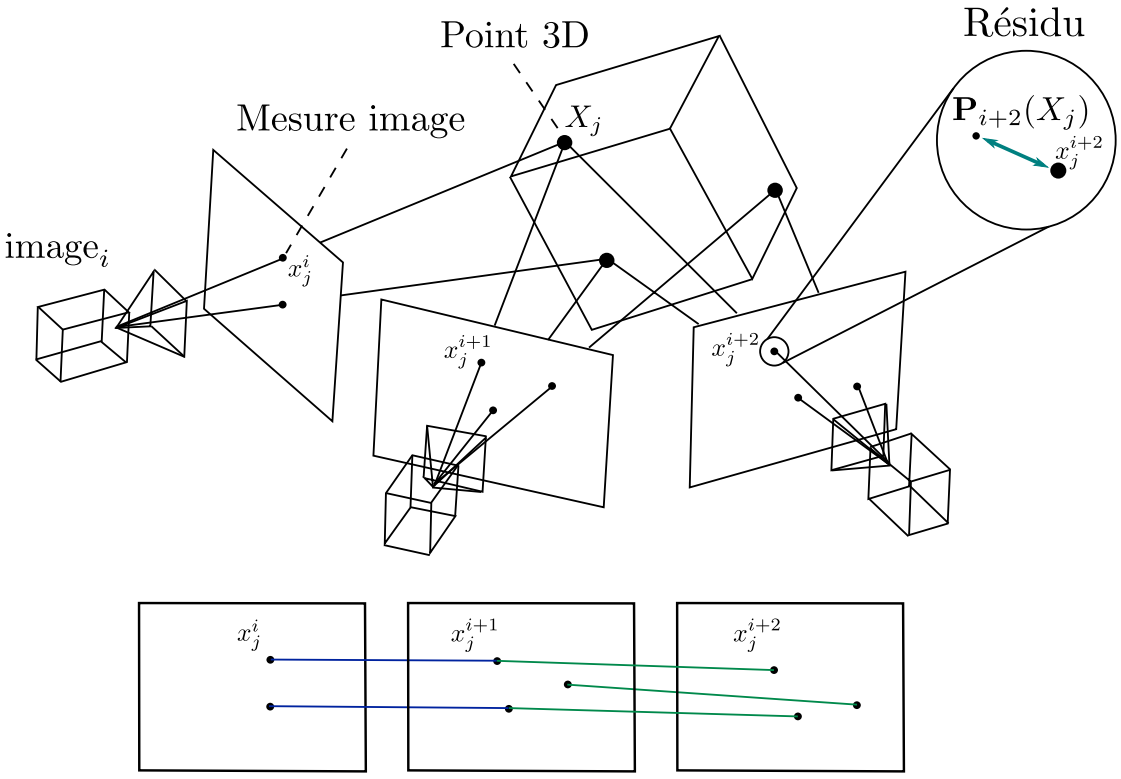


FIGURE 3.9 – Illustration de la problématique de *SfM*. A partir de correspondances entre images relatant des projections de points 3D commun, il s'agit d'identifier la structure et les positions et orientations des clichés dans l'espace.

On note trois grandes catégories d'algorithmes de *Structure-from-Motion* (cf. figure 3.10) :

1. **Les méthodes séquentielles, *Sequential SfM* :**

   La méthode de reconstruction fait naître une première graine 3D, une reconstruction initiale créée à partir de deux vues, puis la méthode fait croître cette reconstruction en agrégeant les images restantes par estimation de pose. Des itérations répétées d'ajustement de faisceaux sont utilisées pour limiter les effets de dérive et d'accumulation d'erreurs. On distingue deux sous cas en fonction de la manière d'établir les correspondances visuelles considérées entre images :

   Dans le cas de séquences d'images ordonnées :
   Les images sont traitées les unes après les autres dans leur ordre d'arrivée : odométrie visuelle ou SLAM (communauté robotique). Les traces sont construites de proche en proche à chaque arrivée d'image.

   Dans le cas de séquences d'images non ordonnées :
   Les images sont traitées dans leur ensemble pour construire les correspondances visuelles et identifier les traces.

2. **Les méthodes hiérarchiques, *Hierarchic SfM* :**

   Les images sont traitées par sous ensembles et sont fusionnées de manière hiérarchique afin de reconstruire l'intégralité de la scène.

3. **Les méthodes globales, *Global SfM, Batch SfM* :**

   Les images sont traitées dans leur ensemble. On distingue ici deux méthodes :
   – Factorisation, *Batch SfM* : On recherche les matrices de projection et la structure de la scène simultanément,
   – *Global SfM* : Supposant les paramètres intrinsèques connus (une caméra calibrée) le problème de *SfM* est découpé en deux sous tâches. D'abord les rotations globales sont identifiées pour toutes les images puis dans un second temps la structure et les translations des caméras sont identifiées.

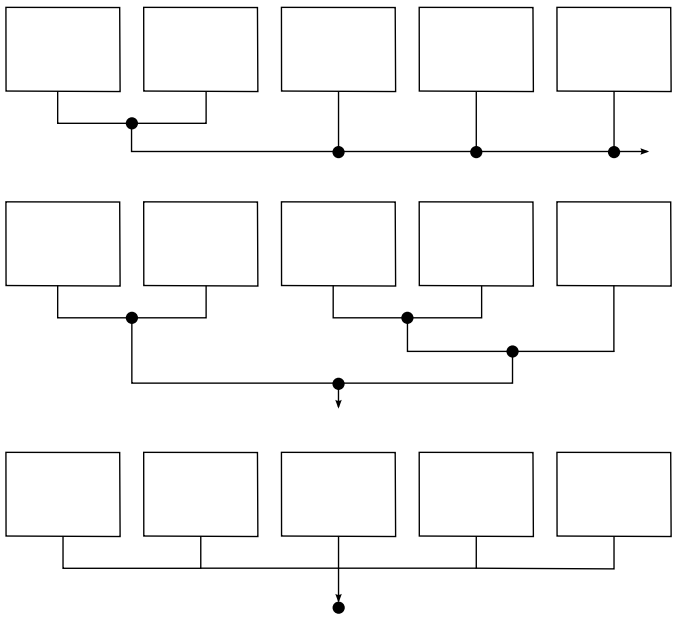


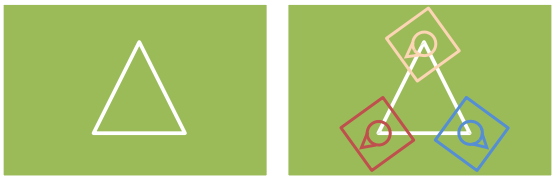
FIGURE 3.10 – Illustration du processus des méthodes de *SfM*, de haut en bas, séquentielle, hiérarchique et globale. Chaque point noir implique une reconstruction 3D ou un assemblage de reconstructions 3D.

3.9 La mise en correspondances de points saillants

L'extraction de caractéristiques visuelles, *visual features extraction*, consiste en des transformations mathématiques calculées sur les pixels d'une image numérique. Ces transformations permettent de mettre en évidence des éléments saillants possédant certaines propriétés visuelles de l'image et de rechercher si des images possèdent du contenu en commun localement similaire. Cette mise en correspondances photométrique est réalisée en trois étapes principales :

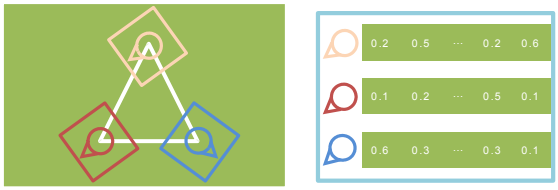
1. **La détection :**

   Une prise de décision locale en chaque point de l'image détermine si la zone de l'image présente une caractéristique intéressante. Les zones mises en évidence représentent des sous-ensembles du domaine de l'image, souvent sous la forme de points isolés, de segments, de courbes continues ou de régions.



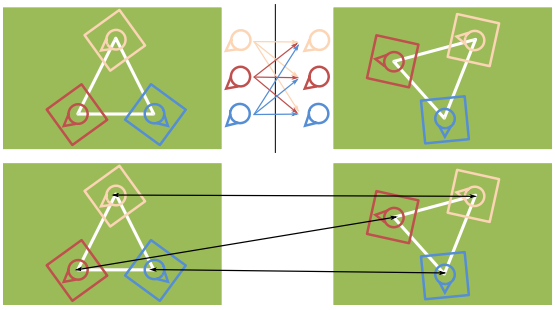
2. **La description :**

   Une zone locale est utilisée autour de chaque zone d'intérêt pour construire une signature venant décrire la région autour du point d'intérêt.



3. **L'appariement :**

   Une comparaison des signatures entre deux images permet d'identifier les zones similaires et ainsi d'identifier des points saillants images d'un même point 3D de la scène observée.



Le résultat de la phase d'appariement est utilisé par de nombreuses applications en vision par ordinateur :

- la reconnaissance et le suivi d'objets,
- l'assemblage d'images panoramiques,
- la stabilisation vidéo,
- la reconstruction 3D et l'odométrie visuelle.

3.9.1 La détection de points saillants

Il est primordial que la détection de points saillants dans une image soit le plus robuste possible. L'invariance à certaines transformations comme la translation, la rotation et l'échelle permettront ainsi d'établir avec plus de succès les futures étapes d'appariements de points, et donc la robustesse des correspondances établies avec le déplacement de la caméra. Plus un détecteur sera à même d'exhiber des points localisés précisément d'une image à l'autre plus il sera pertinent. Plusieurs catégories de détecteurs sont identifiables :

- les bords, *edges, curves* (Canny, LSD, ...),
- les coins, *corners* (Harris, Fast, ...),
- les régions, *blobs* (Sift, Surf, Kaze, Mser, ...).

De nombreuses approches ont été proposées pour améliorer la robustesse et la répétabilité de la détection de points saillants. L'une des premières approches a avoir été largement utilisée est le détecteur de coins de HARRIS et STEPHENS (1988), invariant à l'orientation de la structure détectée. LINDEBERG (1998) a ensuite proposé une représentation en espace échelle linéaire des images qui permet de définir une famille de détecteurs de structures invariantes par changement d'échelle. L'utilisation de cet espace échelle a été généralisé à la détection de coins par Harris-Laplace (MIKOLAJCZYK et SCHMID 2001) et à la détection de blobs avec l'utilisation de différences de gaussiennes par LOWE (1999). En définissant un point d'intérêt comme un extremum local de la représentation en espace-échelle, ces approches permettent d'attribuer à ce point une échelle caractéristique. Ces approches sont souvent coûteuses en calcul à cause de la construction de l'espace échelle. Certaines approximations sont alors réalisées, au détriment de la précision de localisation des zones saillantes, mais au profit de gains de calculs non négligeables. L'utilisation d'images intégrales par la méthode SURF (BAY et al. 2006) permet de réduire la consommation mémoire et de réaliser les opérations de filtrage de manière très efficace. L'exploration de nouveau espace échelle mené par ALCANTARILLA et al. (2012) démontre de meilleures stabilités sur les détections réalisées et laisse envisager un détecteur encore plus performant (proche du temps réel : ALCANTARILLA et al. (2013)).

FIGURE 3.11 – Une image i est décrite par un ensemble j de points saillants : $\{P_j^i\}_{i,j}$. Dans le cas de SIFT une position, une orientation et une échelle caractéristique sont extraits par points.

3.9.2 La description de point saillants

Afin de retrouver les observations image x_j^i d'un même point 3D X_j parmi une séquence d'images, il est nécessaire d'identifier chaque observation image de manière unique. Pour cela, chaque point est décrit par une signature. Cette signature définit un ensemble de caractéristiques, un descripteur local qui est une représentation compacte du voisinage du point d'intérêt (cf. figure 3.12).

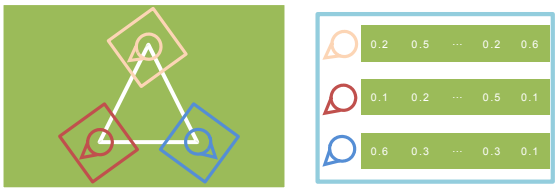


FIGURE 3.12 – Chaque points saillants : $\{P_j^i\}_{i,j}$ est décrit par une transformation de la zone image locale l'entourant : $\{\text{desc}(P_j^i)\}_{i,j}$.

Une solution simple, mais peu robuste, consiste à extraire un patch centré sur le point d'intérêt. Mais le support d'invariance de ce type de descripteur est faible et limite donc son usage. La robustesse de la description locale des points saillants peut être améliorée en supportant l'invariance à des transformations géométriques et aux changements d'éclairement. LOWE (1999) propose une représentation locale appelée SIFT. Ce descripteur SIFT est composé d'histogrammes d'orientation du gradient. Ces histogrammes sont estimés à partir de régions distinctes du voisinage normalisé et centré de chaque point d'intérêt considéré. Il a été montré par MIKOLAJCZYK et SCHMID (2005) que ce type de descripteur est très robuste à différents phénomènes, tels que : bruit, compression JPEG, changement d'éclairement, rotation et changement d'échelle. Ce type de descripteur est très utilisé pour les applications de recherche par le contenu et la photogrammétrie car il possède une répétabilité élevée.

Un inconvénient majeur des descripteurs de type SIFT est l'occupation mémoire. On construit en effet un descripteur de 128 valeurs flottantes par point d'intérêt. Cet espace de relativement haute dimension n'est pas idéal pour les calculs sur des collections d'images à large échelle. Des alternatives permettent de réduire la taille des signatures en utilisant des signatures binaires. Ces signatures ont l'avantage d'être plus compactes en mémoire et de proposer un espace de faible dimension. Leur dimension étant plus faible, la phase d'appariement sera réalisée de manière plus rapide. STRECHA et al. (2012) projettent les descripteurs SIFT en une représentation compacte via une matrice de projection apprise par *machine-learning*. CALONDER et al. (2012) calculent nativement un descripteur binaire par l'utilisation du signe de la différence de couple de points sur une grille autour du point d'intérêt.

Idéalement on souhaite disposer de détecteurs et descripteurs de points d'intérêt ayant les qualités suivantes :

- invariance de détection en translation, rotation et échelle,
- invariance aux variations d'éclairement (luminosité, contraste),
- un critère suffisamment local pour gagner en robustesse aux occultations mais suffisamment large pour décrire suffisamment de contenu.

L'ajout de nouveaux degrés d'invariance tend à créer des faux positifs lors de l'établissement de correspondances, car plus on devient invariant plus les points ont de chance de se ressembler. C'est pourquoi lors de conditions d'acquisition contrôlée (comme c'est le cas avec des robots mobiles), l'invariance en rotation n'est pas considérée. En considérant une orientation verticale, des appariements plus stables seront ainsi identifiés.

3.9.3 L'appariement de point saillants

Soit deux images, A et B, représentées par un ensemble de points saillants et descripteurs. La phase d'appariement consiste à identifier les points ayant une forte similarité entre les deux images. On note N_A le nombre de points d'intérêts de l'image A et N_B ceux de l'image B. Pour chaque point saillant de A on recherche les points les plus similaires de B. Cette mesure de ressemblance est réalisée par l'utilisation d'une métrique entre les descripteurs liés aux points. Les N_B distances sont évaluées pour chaque point de A dans B. Cet ensemble présentant de nombreuses fausses hypothèses, il convient d'utiliser un critère de rejet basé sur l'analyse des distances calculées pour retenir uniquement les distances les plus vraisemblables (cf. figure 3.13).

L'étape d'appariement nécessite trois éléments :

- la **recherche de plus proches voisins** pour obtenir les correspondances hypothèses,
- l'utilisation d'une **métrique** pour mesurer la similarité d'une correspondance,
- l'utilisation d'une **politique de rejet** pour valider une correspondance.

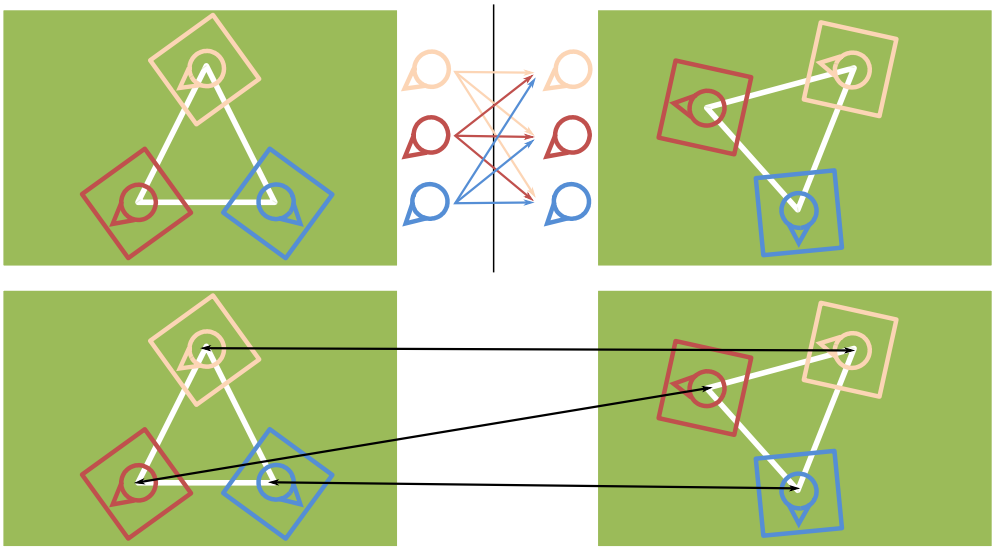


FIGURE 3.13 – Appariements : les points les plus similaires de l'image A (gauche) sont recherchés dans l'image B droite. Un filtre est utilisé pour déterminer de tous les candidats possibles si une correspondance est dominante ou non.

La recherche de plus proches voisins pour chaque point d'intérêt est un problème coûteux. La méthode naïve teste de manière exhaustive toutes les possibilités et les ordonne par distance. Cette méthode est dite de **force brute**, *Brute force*. Des méthodes de calculs approchées identifient les k plus proches voisins d'un descripteur plus rapidement. Ces **méthodes approchées**, *ANN, Approximate Nearest Neighbour*, répartissent les descripteurs suivant leur ressemblance dans un arbre binaire (KD-TREE) (MUJA et LOWE 2009). Cet arbre binaire permet un parcours rapide pour évaluer quel sous-ensemble de l'arbre est le plus similaire à un élément donné en requête. Ce partitionnement permet de limiter le nombre de candidats sur lesquels la métrique est évaluée lors d'une opération de recherche et d'identifier rapidement k voisins. La complexité de recherche sur de larges ensembles de descripteurs est réalisable au prix d'une légère dégradation des performances en précision de calcul, qui est fonction de la taille du groupe considéré et des paramètres de l'arbre de partition.

La métrique est choisie en fonction du type de descripteurs utilisés :

Euclidienne $d(x,y) := \|x-y\|_2$ pour des descripteurs composés de valeurs réelles,

Hamming $d(x,y) = \sum(x \oplus y)$ pour des descripteurs composés de valeurs binaires,

EMD Earth Mover Distance pour des descripteurs circulaires (cf. RABIN (2009)).

Le critère de rejet permet de réduire avec une heuristique les fausses hypothèses parmi les correspondances établies précédemment. Parmi les correspondances établies, seulement quelques-unes sont valides et on souhaite les conserver. Couramment, un filtre réalise le rejet de candidats, parmi les N_B distances évaluées, pour chaque point de A. On note plusieurs politiques de rejet dans la littérature (RABIN 2009) :

Critère FNN ***First Nearest Neighbour*** Pour chaque requête, on garde le plus proche voisin ; le descripteur présentant le plus de similarité :

$$\{(P_A^i, P_B^j) : j = \underset{k}{\operatorname{argmin}} \quad d(\operatorname{desc}(P_A^i), \operatorname{desc}(P_B^k))\}$$

Critère DT ***Distance Threshold*** Un seuil de validation global est utilisé sur la distance : Pour chaque requête, l'ensemble des descripteurs candidats ayant une distance plus petite que le seuil global δ sont validés :

$$\{(P_A^i, P_B^j) : d(\operatorname{desc}(P_A^i), \operatorname{desc}(P_B^j)) < \delta\}$$

Critère DR ***Distance Ratio*** Le pourcentage de ressemblance entre les 2 plus proches voisins dans la seconde image est utilisé : pour chaque requête $(P_i)_A$, les 2 plus proches voisins $(P_j)_B, (P_k)_B$ sont identifiés. Le plus proche voisin est conservé comme point homologue si le ratio des distances $d((P_i)_A, (P_j)_B)/d((P_i)_A, (P_k)_B)$ est inférieur à un seuil δ. L'idée utilisée est que plus des candidats sont similaires, plus la chance de confusion est forte. On évite ainsi de mettre en relation des correspondances ambiguës. δ est souvent choisi entre 0.6 et 0.8 (cf. les expérimentations de LOWE (1999) pour identifier les meilleures valeurs possibles de ce paramètre).

$$\{(P_A^i, P_B^j) : j = \underset{k}{\operatorname{argmin}}\, d(\operatorname{desc}(P_A^i), \operatorname{desc}(P_B^k)) < \delta \min_{k \neq j} d(\operatorname{desc}(P_A^i), \operatorname{desc}(P_B^k))\}$$

Critère SD ***Symmetric distance*** Une correspondance n'est conservée que si les correspondances sont réciproques : les indices mis en correspondance doivent être les mêmes quel que soit le sens de calcul $A \rightarrow B$ et $A \leftarrow B$:

$$\{(P_A^i, P_B^j) : j = \underset{k}{\operatorname{argmin}}\, d(\operatorname{desc}(P_A^i), \operatorname{desc}(P_B^k)), i = \min_{k} d(\operatorname{desc}(P_A^k), \operatorname{desc}(P_B^j))\}$$

Les correspondances établies étant photométriques de faux positifs peuvent toujours être présents. Il convient par la suite de vérifier si les correspondances établies sont géométriquement cohérentes (cf. chapitre 4).

Dans le cadre de cette thèse nous utiliserons les détecteurs et descripteurs SIFT (LOWE 1999). Les appariements ayant passé la politique de rejet DR sont retenus. La méthode approchée ANN accompagnée de la norme euclidienne l_2 est utilisée pour rechercher les candidats.

3.10 Méthode de fusion rapide de paires de correspondances de points saillants entre images

Un des pré-requis souvent utilisé en vision par ordinateur est l'information de visibilité, c'est-à-dire la connaissance qu'un point 3D donné se re-projette dans une série d'image. Se pose alors le problème suivant : étant donné des détections image nous voulons suivre le déplacement de ces détections dans une série d'images. Ce problème est appelé suivi de points, ou *point/feature tracking*. Nous appellerons le déplacement apparent d'un point de l'espace dans une série d'images une trace, (*track*).

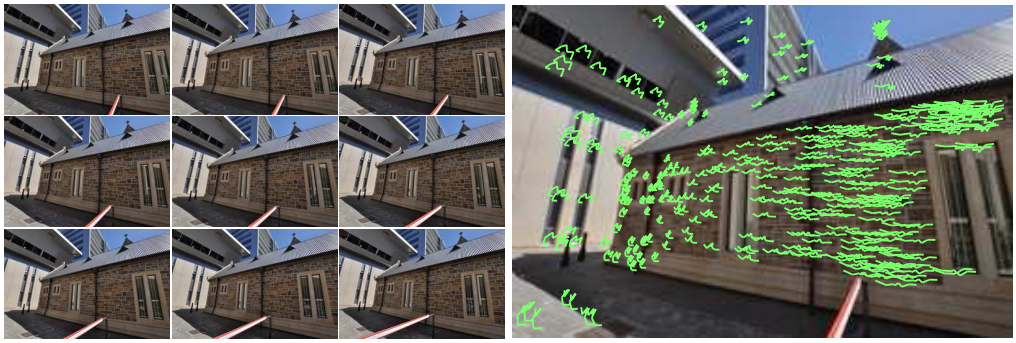


FIGURE 3.14 – A gauche, une séquence de 9 images fournies avec le logiciel VideoTrace. A droite, une série de traces identifiant le mouvement des points saillants qui ont pu être suivis à travers toute la série d'images en utilisant nos algorithmes.

Le problème de suivi de points à travers une série d'images peut être abordé de différentes manières en fonction de la nature de la collection d'images à traiter :

Une séquence d'images ordonnée. *Narrow-baseline matching*.

Ce type de séquence, comme des vidéos, induit par nature une amplitude de mouvement faible. Ce faible mouvement des points à suivre permet de construire les trajectoires de points de proche en proche par des zones de recherches locales. Les méthodes sont basées sur des approches de corrélation (LUCAS et KANADE 1981 ; TOMASI et KANADE 1991) (cf. figure 3.15), ou de flux optique (HORN et SCHUNCK 1981). Les méthodes les plus récentes présentent soit des boucles d'apprentissage et de détection : TLD (KALAL et al. 2012) ou des analyses très rapides des champs de déplacement : Zero Shift points (DUPAČ et al. 2012).

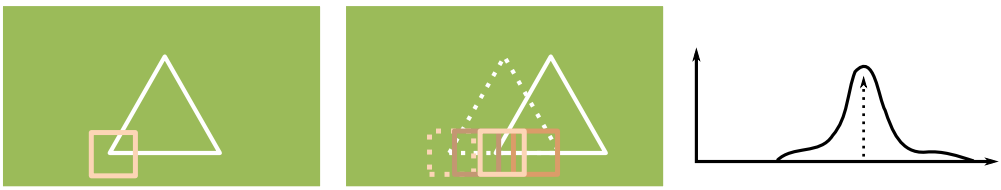


FIGURE 3.15 – Suivi de trajectoire d'un point saillant par maximum de corrélation.

Une série d'images non ordonnée. *Wide-baseline matching*.

Dans ce cas la cohérence de mouvement ne peut être supposée. En effet les points n'ont pas forcément de cohérence d'une image à l'autre car une partie différente de la scène peut être vue. Le suivi de points est plus difficile, on ne sait pas où rechercher d'une image à l'autre. Les points similaires entre des paires d'images sont alors identifiés par des méthodes d'*image-matching* (cf. section 3.9)

puis filtrés pour vérifier leur cohérence géométrique (cf. chapitre 4). Ces correspondances par paires d'images sont ensuite assemblées en trajectoires lorsqu'elles partagent des points communs (cf. figure 3.16).

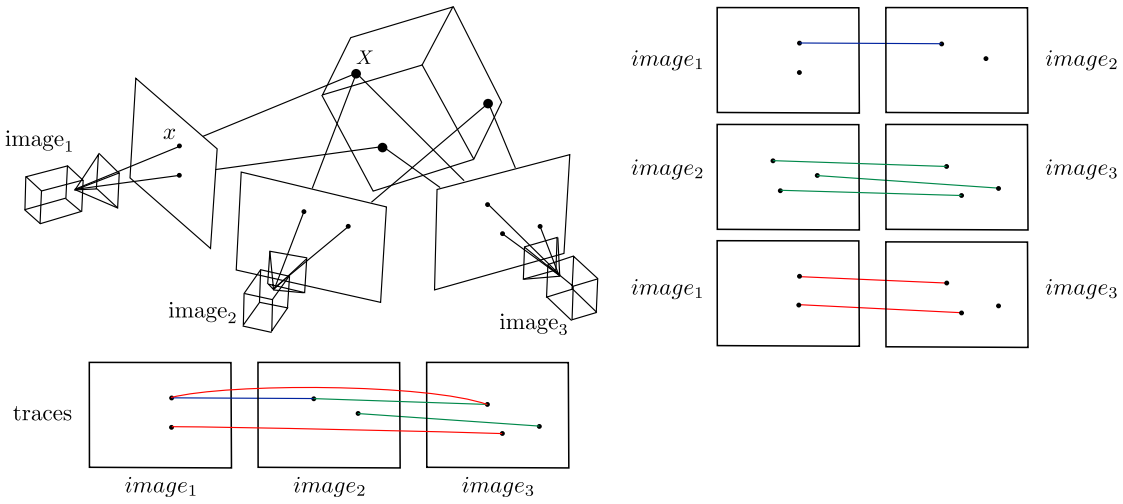


FIGURE 3.16 – Dans une série d'images non ordonnées ($image_1$, $image_2$, $image_3$), des correspondances sont identifiées par paire d'images (droite). Relier ces correspondances si elles partagent des points en commun permet d'identifier les traces (bas gauche). Note : L'analyse d'une série de paires $1 \to 2, 2 \to 3, 1 \to 3$ permet de créer une correspondance entre l'image 1 et 3 qui ne serait pas identifiée par une analyse en séquence : $1 \to 2 \to 3$.

Nous nous intéressons ici au cas générique, le cas des images non ordonnées. On se retrouve donc avec la tâche suivante : combiner les correspondances géométriques identifiées par paires en traces cohérentes. L'état de l'art traite le plus souvent ce problème par la construction d'un graphe suivi d'une phase d'analyse (SNAVELY et al. 2006).

Soit i^k le $k^{\text{ième}}$ point de l'image i et (i^m, j^n) une mise en correspondance établie entre l'image i et j avec les points indicés m et n. Étant donné une liste de correspondances, $\{(i^m, j^n)\}$, on cherche à construire un graphe $\mathcal{G}$ puis on l'analyse pour identifier les traces. Soit $\mathcal{G} = \{\mathcal{S}, \mathcal{E}\}$ avec $\mathcal{S}$, un ensemble de sommets, et $\mathcal{E}$ un ensemble d'arêtes.

Les sommets $\mathcal{S} : \{i^k\}$ représentent les points saillants détectés dans les images.

Les arêtes $\mathcal{E} : \{(i^m, j^n)\}$ représentent les correspondances établies par paires d'images.

Les deux étapes requises pour identifier les traces grâce au graphe $\mathcal{G}$ sont les suivantes :

1. $\mathcal{G}$ est construit en utilisant l'ensemble des correspondances par paires : $\{(i^m, j^n)\}$:
   Des liens entre points saillants $\mathcal{S}$ sont ainsi créés pour chaque correspondance (i^m, j^n) établie. Les correspondances deux à deux sont ainsi reliées entre elles.
2. Une analyse en composante connexe permet d'identifier les traces :
   Chaque composante connexe est une trace identifiant le déplacement apparent d'un point saillant dans une série d'images.

Cette approche est fonctionnelle mais non optimale, nous démontrerons dans la section suivante que l'utilisation d'une structure de données plus adaptée permet de gagner en efficacité.

3.10.1 Une solution ensembliste pour la construction de traces de points saillants

Au lieu de voir le problème de construction de traces comme la construction d'un graphe et son analyse, nous considérons le problème de manière ensembliste. Nous montrons que ce problème est soluble par l'utilisation de la théorie des ensembles et que son utilisation est plus efficace que les méthodes de l'état de l'art sur le plan de la complexité algorithmique. De plus, son utilisation possède plusieurs avantages aux vues des solutions concurrentes disponibles.

La théorie

Proposition 1. *En théorie des ensembles, la notion de relation d'équivalence sur un ensemble permet de mettre en relation des éléments qui sont similaires par une certaine propriété.*

Soit E un ensemble et $\mathcal{R}$ une relation d'équivalence. L'utilisation de la relation d'équivalence $\mathcal{R}$ sur E permet la construction du groupe quotient $E|\mathcal{R}$ composé de classes Q. Chaque Q représente ainsi la fusion des éléments similaires de E : les éléments de E suivant la relation d'équivalence $\mathcal{R}$. Les ensembles quotients $\{Q\}$ obtenus représentent des classes disjointes.

Proposition 2. *L'utilisation de la relation d'équivalence sur une ensemble permet de créer les classes par complétion par transitivité. Étant donné un ensemble d'éléments (les points saillants) nous partitionnons en un certain nombre de classes disjointes les relations établies par les correspondances. L'utilisation des correspondances de points homologues comme relation d'équivalence permet d'établir les ensembles disjoints désirés : les traces.*

La proposition 2 démontre que la théorie des ensembles est applicable à notre problème. Voyons désormais comment réaliser ces opérations de manière concrète.

La solution logicielle

GALLER et FISHER (1964) propose de mener efficacement la construction et la manipulation de classes d'équivalence à travers une structure de données, les *disjoint-set*, et des algorithmes : *union-find*. La structure de données *disjoint-set* permet de maintenir une forêt d'arbres, chaque arbre représentant un ensemble disjoint. Les algorithmes *union-find* permettent de maintenir et créer des partitions entre les ensembles disjoints grâce à deux opérations :

Trouver, *Find* détermine la classe d'équivalence d'un élément. Elle sert aussi à déterminer si deux éléments appartiennent à la même classe d'équivalence.

Unir, *Union* réunit deux classes d'équivalence en une seule.

Une condition préalablement nécessaire à leur utilisation est la construction des singletons : des ensembles d'équivalences contenant un seul élément par l'instruction *MakeSet*.

La mise en pratique

L'utilisation de la théorie des ensembles pour notre problème de construction de traces nécessite la définition de l'ensemble et de la relation d'équivalence :

Soit $\{E\}$ une collection d'ensembles : chaque point saillant considéré par les correspondances est un ensemble disjoint,

Chaque point saillant est considéré comme une trace de taille 1.

Soit une relation d'équivalence $\mathcal{R}$: $E(find(i^m)) = E(find(j^n))$,

L'utilisation de la relation d'équivalence permet de fusionner deux classes, les classes contenant les points homologues désignés par une correspondance (i^m, j^n) par utilisation de la fonction union : union$(find(i^m), find(j^n))$. On réalise la complétion par transitivité.

La procédure 1 de calcul de traces est alors réalisée, elle est composée de trois étapes qui nécessitent de parcourir deux fois l'ensemble de correspondances :

1. **Pour créer les ensembles de bases :**

   Un ensemble est créé par point saillant utilisé,

2. **Pour appliquer la relation d'équivalence :**

   Les correspondances relatives sont fusionnées en traces par l'utilisation répétée des fonctions *find* et *union* sur l'ensemble des correspondances. *Find* identifie les ensembles d'appartenance des points saillants de la correspondance considérée et les fusionne en utilisant l'opérateur *union*.

3. Les classes calculées sont parcourues pour lister et identifier les points appartenant à chaque trace.

La fusion des correspondances est ainsi réalisée par fusion itérative des correspondances à deux vues. Des traces de points de taille 1 sont unies et évoluent pour former les traces désirées au cours du processus. L'algorithme obtenu est très simple à lire et implémenter. Il est constitué de deux boucles sur les correspondances relatives.

Procédure 1 Calcul des traces de points saillants pour une série de paires de correspondances

Entrée: une liste de correspondances entre différentes paires d'images : $\mathcal{L} : \{(i^m, j^n)\}$
Sortie: les traces

(1) **Construction des ensembles initiaux, les singletons :**
pour $(i^m, j^n) \in \mathcal{L}$ **faire**
  si find$(i^m) \neq \emptyset$ **alors**
    MakeSet(i^m)
  fin si
  si find$(j^n) \neq \emptyset$ **alors**
    MakeSet(j^n)
  fin si
fin pour
(2) **Complétion par transitivité :**
pour $(i^m, j^n) \in \mathcal{L}$ **faire**
  union(find(i^m), find(j^n))
fin pour
(3) **Récupération des traces :**
Retourne chaque arbre de la forêt comme une trace

La complexité associée à une utilisation naïve de l'algorithme *Union-Find* et des *disjoint-sets* est $O(n\log(n))$. TARJAN (1975) a ensuite montré que l'utilisation de deux optimisations, *union by rank* et *path compression*, permettent de rendre la complexité quasi-linéaire en pratique : $O(n\alpha(n))$ (α étant l'inverse de la fonction de Ackermann). Il n'est

pas possible d'obtenir un meilleur résultat : FREDMAN et SAKS (1989) ont montré que $\Omega(\alpha(n))$ mots en moyenne doivent être lus par opération sur toute structure de données pour le problème des classes disjointes.

Le comportement de l'algorithme est illustré en figure 3.17. Des correspondances par paires sont identifiées par différentes couleurs (bleu, vert, rose, violet). Ces correspondances impliquent 12 points, 12 ensembles sont alors créées. Les correspondances sont alors parcourues par paires (marquées par les différentes couleurs) et les ensembles contenant les points homologues sont assemblés (*union*). Une forêt d'arbres est alors construite et évolue pour créer les 4 trajectoires associées aux correspondances initiales.

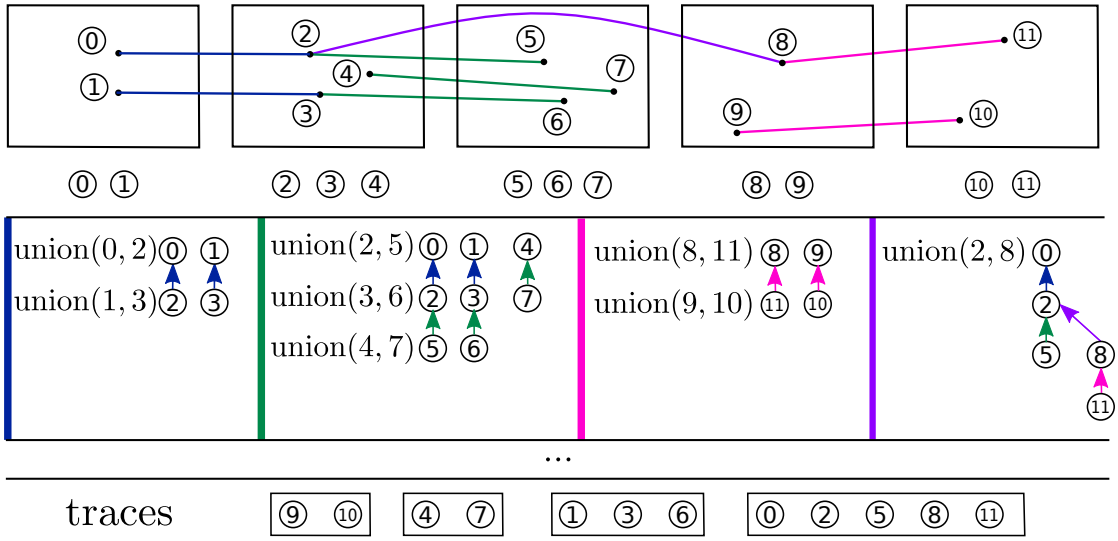


FIGURE 3.17 – Illustration des étapes de l'algorithme de fusion de traces. De haut en bas. Les correspondances initiales par paires identifiées par différentes couleurs. La création des 12 singletons, désignant les 12 points considérés par les correspondances. L'évolution de la forêt de traces (de gauche à droite avec le parcours des appariements par couleur). Et enfin chaque arbre de la forêt est parcouru pour identifier les points appartenant à une trace commune.

Expérimentations

Notre solution , désignée sous l'acronyme UF, a été évaluée face à deux autres solutions :

1. SNAVELY et al. (2006) : solution contenue dans le code source du logiciel associé Bundler.
2. ZACH (2010-2011) : solution contenue dans le code source de la bibliothèque associée ETH-V3D.

Le protocole de test est le suivant : pour des données initiales identiques, une liste de correspondances $\{(i^m, j^n)\}$ géométriquement valide par géométrie fondamentale (cf. section 3.3) est fournie aux trois solutions d'identification de traces. Les temps de calcul et nombre de traces identifiées sont mesurés et comparés. Afin de pouvoir tester différents scénarios le test est réalisé sur différents jeux de données images présentant de 8 à 314 images. Ces tests permettent de faire intervenir de 10000 à 1 million de correspondances initiales et ainsi d'évaluer le comportement des solutions à faible et large échelle. Les résultats de l'expérience sont disponibles dans le tableau 3.1.

	Jeux d'images	nb. Images	$\#\{i^m, j^n\}$	Méthodes					
				UF		Zach		Bundler	
				temps s.	# traces	temps s.	# traces	temps s.	# traces
1	HerzJesus	8	13726	**10**	2454	20	2383	**10**	2314
2	Castle	19	17853	**12**	2673	18	2543	30	2229
3	Entry	10	18914	**13**	2697	27	2524	20	2566
4	Fountain	11	29338	**21**	4279	52	4139	30	3513
5	Jean-fontana	66	53901	**49**	8551	72	7773	90	7672
6	Castle	30	56477	**40**	5639	73	5272	60	5033
7	DeteniceFountain	59	63437	**56**	7949	86	7445	250	7447
8	HerzJesus	25	68284	**50**	6603	130	5781	90	6160
9	SceauxCastle	282	364751	**411**	39639	613	36508	680	35845
10	StMartin	124	651990	531	51473	1505	49109	**410**	41380
11	Temple	314	1012804	**640**	20623	1295	15580	1280	19730

TABLE 3.1 – Statistiques sur l'évaluation la fusion de correspondances pour le calcul de traces. Les résultats sont triés par ordre croissant du nombre de correspondances relatives. Le temps le plus court est affiché en gras.

Pour faciliter l'interprétation des résultats les temps des différentes méthodes ont été représentés sous forme graphique sur la figure 3.18.

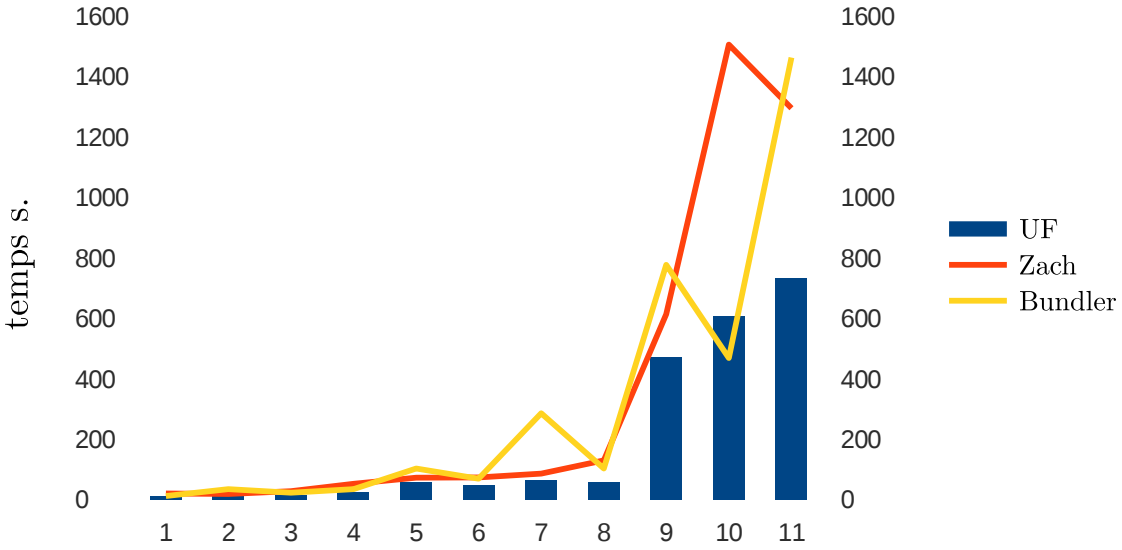


FIGURE 3.18 – Représentation graphique des temps nécessaires pour identifier les traces du tableau 3.1. Les jeux de données sont triés de manière croissante en fonction du nombre de correspondances initiales.

L'expérience permet de faire des remarques sur les points suivants :

la vitesse d'exécution On remarque que la solution **UF** est dans 90% des cas plus rapide que les deux autres solutions (cf. courbes de la figure 3.18).

la complexité à large échelle Pour de larges jeux de données on constate que les méthodes Zach et Bundler présentent des résultats en dents de scies avec la taille du jeu de correspondances relatives fournies. Notre solution par contre réagit de manière beaucoup plus linéaire.

la complétude de la solution : le nombre de traces identifiées Il est important de noter que le nombre de traces varie d'une implémentation à l'autre. Notre approche étant ensembliste nous avons des garanties que pour notre critère d'équivalence la solution identifiée soit optimale. Le fait que nous détectons tout le temps plus de traces que les deux autres méthodes démontrent que leur implémentation n'est pas parfaite et que des traces ne sont pas identifiées, ou rejetées à tort.

Concernant la complétude des solutions identifiées les arguments suivants sont avancés : les deux implémentations disponibles évitent l'utilisation de graphes et utilisent des tables d'indices. Ce qui rend les algorithmes sensibles à plusieurs facteurs :

Bundler : SNAVELY et al. (2006)
Résultat dépendant d'un index de départ,
Requiert plusieurs opérations de tri.

ETH-V3D : ZACH (2010-2011)
Résultat dépendant de l'ordre des paires d'images,
Grosse consommation mémoire.

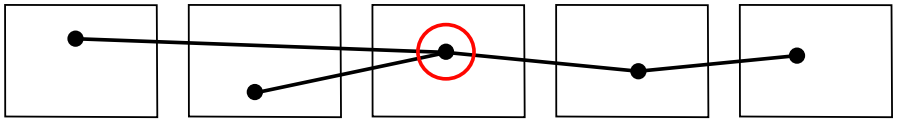
La complexité de tels algorithmes est au minimum de $O(n\log(n))$ (due aux opérations de tri requises). Nos tests ont confirmé que notre solution avec une complexité plus faible, $O(n\alpha(n))$ minimale théorique, présente une meilleure réactivité aux données d'entrées.

Limitations et perspectives

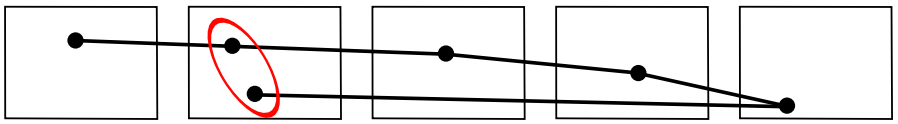
Les temps d'exécution de notre approche pourraient être encore réduits par l'utilisation d'une version non bloquante de l'algorithme *union-find* (ANDERSON et WOLL 1991). Une version parallélisable de notre algorithme est alors envisageable, laissant imaginer de meilleurs temps de réponse sur de larges jeux de données.

Cependant l'assemblage naïf des correspondances donne lieu à plusieurs problèmes (AGARWAL et al. 2009) :

- plusieurs traces peuvent se croiser et donc contenir plusieurs fois le même point,



- plusieurs traces peuvent contenir des points dans la même image.



Notre approche actuelle se limite juste à détecter et ne pas exporter les traces qui portent ces cas de conflits. Étant capable de détecter les arbres présentant des conflits nous pouvons envisager un post-traitement afin de couper ses traces corrompues en traces cohérentes en utilisant la méthode proposée par SVARM et al. (2012). Leur approche propose une solution pour garder les traces les plus probables parmi les traces corrompues qui se croisent. Leur solution est basée sur une analyse des arbres dits de Gomory-Hu. Étant données des traces corrompues, l'algorithme utilise des coupes récursives afin de conserver les n sous-traces portant les plus grandes pondérations. Chaque arête porte pour poids le nombre de points géométriquement validés pour la paire d'images considérée par cette arête.

3.11 Contributions de ce chapitre

Nous avons présenté les fondamentaux de la géométrie multi-vues et comment des correspondances de points pouvaient être identifiées de manière automatique et assemblées en trajectoire au sein de collections d'images non ordonnées.

Une nouvelle méthode permettant de calculer la fusion de correspondances dans le but d'identifier les traces de points saillants à travers une série d'images a été exposée. L'utilisation d'une structure de données et des algorithmes adaptés nous permet de résoudre le problème de fusion avec une complexité optimale, quasi-linéaire en pratique. Notre solution, utilisant «la théorie des ensembles», permet de mettre en œuvre une solution élégante qui ne réalise aucune approximation et aucun biais dans les résultats.

Nos expériences ont confirmé les résultats théoriques sur le fait que notre méthode a une complexité moindre et donc un temps d'exécution plus faible que les solutions concurrentes. L'utilisation de notre algorithme a toujours démontré qu'il était capable d'identifier plus de traces que les solutions concurrentes pour l'ensemble des jeux de données. Cela démontre que les deux implémentations concurrentes sont biaisées et ne garantissent pas un résultat complet au problème traité.

Ce travail a été présenté à la conférence CVMP (MOULON et MONASSE 2012) et largement utilisé pour les autres travaux de ce manuscrit. Une implémentation libre est disponible avec la librairie open-source openMVG (MOULON et al. 2013d).

Chapitre 4

L'estimation robuste de modèles paramétriques

Lorsque des données sont légèrement bruitées il est courant d'utiliser une méthode d'ajustement aux moindres carrés afin d'identifier les paramètres d'un modèle. Cependant lorsque les données sont bruitées et polluées ces méthodes ne permettent plus de trouver un modèle adéquat. Il est alors courant d'utiliser des méthodes d'estimation robuste qui recherchent le sous ensemble de données s'ajustant le mieux au modèle paramétrique choisi. Ces méthodes reposent sur des tests d'hypothèses pour identifier un modèle et classifier les données suivant leur nature en :

- **mesures fiables** appelées *inliers*, auxquels le modèle s'ajuste,
- **fausses mesures** appelées *outliers*, les fausses mesures que le modèle réfute.

Nous allons dans ce chapitre :

1. Étudier les méthodes couramment utilisées : MAX-CONSENSUS et RANSAC et discuter leurs limitations,
2. Expliquer et discuter un estimateur robuste incorporant un critère statistique permettant de s'adapter de manière dynamique au bruit de mesure,
3. Montrer comment généraliser l'utilisation de cet estimateur robuste adaptatif à différents modèles d'erreur.

Sommaire

4.1 MAX-CONSENSUS

Le but de l'estimation robuste est d'identifier, parmi un ensemble $\mathcal{D}$, le sous-ensemble de points auxquels s'ajuste le mieux un modèle paramétrique $\mathcal{H}$ recherché. L'ensemble $\mathcal{D}$ est en sortie classifié en deux sous-ensembles : les *inliers* et les *outliers*.

Soit $\mathcal{D}_i$ le i^{e} échantillon de $\mathcal{D}$, $\mathcal{M}$ une métrique calculant l'erreur de re-projection d'un échantillon au modèle $\mathcal{H}$ et δ un seuil d'acceptation. MAX-CONSENSUS est une procédure itérative qui repose sur quatre étapes :

1. **La génération d'hypothèses :**
   un échantillonnage stochastique aléatoire de **s**-uplets est réalisée afin de générer des hypothèses $\mathcal{H}$, **s** étant suffisant pour estimer les paramètres d'une hypothèse $\mathcal{H}$.
2. **Une mesure de consensus :**
   l'ensemble des erreurs de re-projection au modèle en cours d'hypothèse $\mathcal{H}$ est évalué pour chaque échantillon.
3. **Un critère de validation :**
   Si l'erreur est inférieure à un seuil δ alors l'échantillon est ajouté au consensus.
4. **Un critère d'arrêt :**
   un nombre d'itérations N.

MAX-CONSENSUS est une méthode qui teste successivement des hypothèses et mesure la taille du consensus généré. La taille du consensus acceptant $\mathcal{H}$ sous une précision δ est maximisée et ainsi la consensus d'échantillon identifié est retenu comme *inliers* :

$$\underset{\mathcal{H}}{\operatorname{argmax}} \sum_{i=1}^{\#\mathcal{D}} \mathbb{1}\left(\mathcal{M}(\mathcal{H}, \mathcal{D}_i) < \delta\right) \tag{4.1}$$

Une recherche exhaustive des **s**-uplets est nécessaire afin de rechercher le modèle idéal. L'évaluation de toutes les combinaisons de **s**-uplets n'est pas réalisable en pratique car $N = \binom{\#\mathcal{D}}{\mathbf{s}}$ combinaisons seraient à évaluer. Notant que $N \to \infty$ en fonction de **s** et la taille de l'ensemble échantillon $\mathcal{D}$ un problème combinatoire se pose. Puisqu'il n'est pas pensable de réaliser tous ces tirages, une façon naïve mais efficace pour limiter la complexité est de choisir un N fixe et d'utiliser un échantillonnage stochastique. MAX-CONSENSUS est une méthode itérative qui tire au sort, N fois, un **s**-uplet pour générer une hypothèse et la vérifier. Note : plus **s** est petit, plus l'exploration stochastique de l'espace des **s**-uplets sera large et rapidement réalisée.

MAX-CONSENSUS est une méthode à deux paramètres :

- δ : précision/erreur maximale tolérée pouvant être acceptée pour une appartenance à l'ensemble de consensus,
- N : le nombre d'itérations à réaliser.

4.2 RANSAC

La méthode RANSAC (RANdom SAmpling Consensus) (FISCHLER et BOLLES 1981) est une évolution de la méthode MAX-CONSENSUS. L'idée est de réduire le nombre de tirages N à réaliser. Si l'on a une idée du nombre d'*inliers* a-priori dans les données ; on peut calculer le nombre de tirages nécessaires pour être sûr d'avoir statistiquement parcouru l'espace des solutions de manière suffisante. A première vue cette solution permet de réduire la complexité algorithmique de la recherche robuste de modèle, mais elle a comme inconvénient d'ajouter un nouveau paramètre p sur la proportion estimée de contamination des données.

RANSAC évalue le nombre suffisant de tirages N à réaliser pour assurer qu'avec une probabilité p, au moins 1 échantillon de taille $\mathbf{s}$ n'est pas pollué. Si le taux d'*inliers* w est connu, alors la probabilité de choisir tous les échantillons pollués est $(1-w^{\mathbf{s}})^N$, c'est à dire Soit la probabilité de tirer N fois un s-uplet contenant au moins 1 *outlier*. On a alors $(1-w^{\mathbf{s}})^N \leq 1-p$, soit encore :

$$N \geq \frac{log(1-p)}{log(1-w^{\mathbf{s}})} \tag{4.2}$$

RANSAC peut ainsi au fur et à mesure de son évaluation estimer le nombre d'itérations N lui restant à effectuer en fonction de la taille du consensus le plus grand rencontré jusqu'alors. A chaque fois qu'un meilleur consensus est identifié, N est remis à jour grâce à l'équation 4.2 (cf. procédure 2).

Procédure 2 RANSAC : Recherche du plus large consensus pour un modèle paramétrique $\mathcal{H}$

Entrée: $\mathcal{D} = \{\mathcal{D}_0, \ldots, \mathcal{D}_i\}$: un ensemble d'échantillons
Entrée: δ : un seuil de précision, borne haute pour l'acceptation des erreurs
Entrée: p : une probabilité sur la contamination de l'ensemble échantillon,
Entrée: N : un nombre maximal de tirages.
Sortie: le plus large ensemble consensus S_{opt} et les paramètres du modèle $\mathcal{H}_{opt}$ retenu.
compteur i = 0, $\#S_{opt} = \emptyset$
(1) **Échantillonnage aléatoire :**
Tirage d'un $\mathbf{s}$-uplet
Estimation d'un modèle $\mathcal{H}$ // Génération d'une hypothèse
(2) **Sélection des inliers :**
$S = \{\mathcal{D}_i\} \mid \mathcal{M}(\mathcal{H}, \mathcal{D}_i) < \delta$ // Évaluation de l'hypothèse
(3) **Consensus optimal :**
si $\#S > \#S_{opt}$ **alors**
$S_{opt} = S$
$\mathcal{H}_{opt} = \mathcal{H}$
N est mis à jour via l'équation 4.2 // Évaluation du #tirages restant à effectuer
fin si
(4) **Critère d'arrêt :**
tant que $i < N$, $i = i + 1$. Retour à l'étape 1.

La seule différence entre MAX-CONSENSUS et RANSAC consiste en la mise à jour du nombre de tirages restant à réaliser. L'algorithme RANSAC peut donc terminer plus rapidement son estimation robuste, mais un paramètre supplémentaire p est rajouté. Cependant, dans la majorité des cas le niveau de bruit des données est inconnu et variable d'un jeu de données à l'autre. Le choix des paramètres δ et p est donc loin d'être

évident.

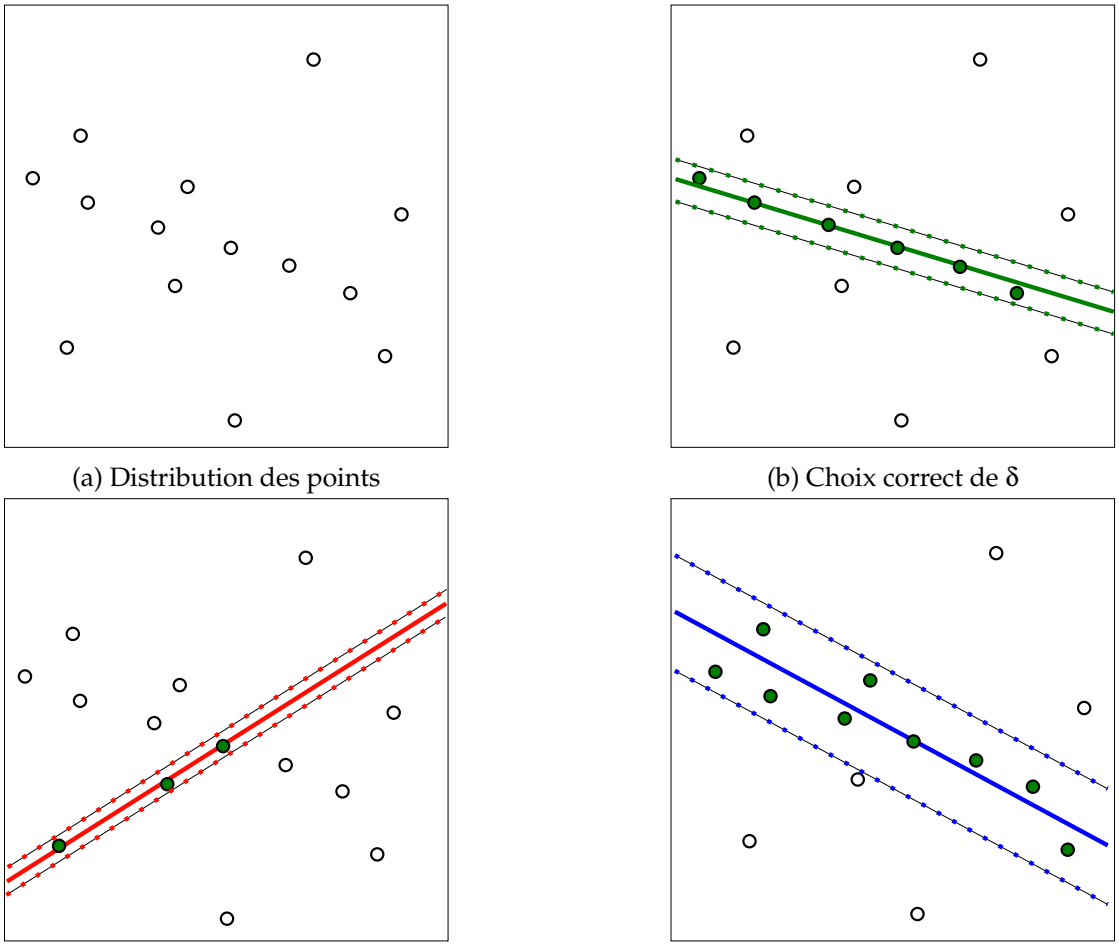
RANSAC est une méthode à trois paramètres :

- δ : précision maximale acceptée pour construire un ensemble de consensus,
- N : le nombre maximal d'itérations pouvant être réalisées,
- p : une probabilité de succès fixé apriori.

4.2.1 Limitations et variantes

Les méthodes MAX-CONSENSUS et RANSAC ont une limitation majeure commune par rapport à ce que l'on appelle l'efficacité relative.

L'efficacité relative est dépendante du choix arbitraire du seuil de précision δ. C'est l'introduction de ce seuil qui permet une robustesse d'estimation tolérant jusqu'à plus de 50 % d'*outliers*, mais en contrepartie le choix de ce paramètre est très critique. Comme le montre la figure 4.1 lorsque ce seuil est trop élevé, quelques *outliers* sont sélectionnés à tort. Lorsque ce seuil est trop faible, une transformation fiable n'est pas estimable, car trop peu d'échantillons sont utilisés. On parle alors de situation de sur-évaluation, *over-fitting*, ou de sous-évaluation, *under-fitting*.



(a) Distribution des points

(b) Choix correct de δ

(c) δ trop petit, sous-évaluation

(d) δ trop large, sur-évaluation

FIGURE 4.1 – Ambiguité du choix du seuil de sélection pour l'estimation d'un modèle de droite.

Diverses variantes de MAX-CONSENSUS et RANSAC sont proposées dans la littérature pour chacune des 4 étapes mises en jeu : la génération d'hypothèses, la mesure de consensus, le critère de validation et enfin le critère d'arrêt. Plutôt que de réaliser une liste exhaustive (cf. CHOI et al. (2009)) nous allons lister quelques méthodes marquantes par domaine d'étude.

Amélioration de la rapidité :

Échantillonnage guidé. L'échantillonnage de groupes de **s**-uplets peut être vu comme un processus de génération d'hypothèses. En l'absence de connaissance apriori sur le modèle suivi par les données, un échantillonnage uniforme est utilisé, chaque hypothèse est ainsi générée indépendamment des précédentes. Plusieurs heuristiques ont été proposées pour faire converger l'algorithme plus rapidement.

MOISAN et STIVAL (2004) propose avec ORSA *(Optimized Random Sampling Algorithm)* de tirer les échantillons parmi le meilleur ensemble S_{opt} lorsqu'une hypothèse est jugée valide.

CHUM et MATAS (2005) propose avec PROSAC (PROgressive SAmple Consensus) de tirer les échantillons en fonction d'un indice de confiance qui leur est associé. Le tirage n'est plus réalisé de manière uniforme mais en tenant compte de cette mesure de qualité. L'intuition est qu'il vaut mieux tester en premier lieu les configurations sur lesquelles on a le plus confiance. Dans le cas de la mise en correspondances d'images, la similarité entre les descripteurs est associée aux échantillons.

NI et al. (2009) propose avec GROUPSAC de partitionner en sous-groupes les échantillons. Dans le cas de correspondances images, un critère géométrique est utilisé. La ressemblance des vecteurs directeurs entre les points images en correspondance est ainsi utilisée dans un algorithme de regroupement hiérarchique. La probabilité de tirer un échantillon est alors déterminée par la taille d'un groupe et non plus uniforme.

Amélioration de la robustesse :

Adaptivité au bruit. Contrairement aux méthodes de type MAX-CONSENSUS qui utilisent un seuil fixe δ sur les résidus observés, les méthodes citées ici visent à mesurer la qualité et la validation d'un groupe associé à une transformation. Il s'agit de déterminer de manière automatique le groupe de consensus qui s'ajuste le mieux au modèle en cours d'hypothèse et de ne plus dépendre d'un seuil δ fixé de manière heuristique. Pour chaque modèle, le bruit des données est estimé, l'ensemble consensus satisfaisant le modèle paramétrique est déterminé statistiquement. Ce problème n'est pas trivial, mais apporter une solution permet de devenir adaptatif aux données.

Une façon de ne plus dépendre d'un seuil δ peut être réalisé en changeant la métrique. ROUSSEEUW (1984) recherche l'ensemble consensus qui minimise la médiane des résidus observés par la méthode LMedS, *Least-Median-of-Squares*. On observe que la méthode permet de rejeter efficacement les données aberrantes mais en contrepartie elle est très sensible à un bruit de type gaussien. L'utilisation de la médiane limite l'identification d'un consensus à des données polluées à moins de 50%.

Hypothèse : distributions normales des inliers et distribution uniforme des outliers : L'algorithme MLESAC *Maximum Likelihood SAC* (TORR et ZISSERMAN 2000) introduit une mesure de qualité basée sur la probabilité de

distribution des *inliers* et *outliers*. La distribution des *inliers* est modélisée comme une distribution gaussienne et les *outliers* comme une distribution uniforme.

Hypothèse : distributions uniformes des outliers : MINPRAN *MINimize the Probability of RANdomness* (STEWART 1995) recherche un ensemble consensus, qui associé à un modèle paramétrique, n'est pas expliqué par la chance (le modèle de fond). La taille de cet ensemble consensus est évaluée par la minimisation d'une probabilité $\mathcal{P}(\mathcal{S}|\mathcal{H})$. Considérant les résidus des *outliers* uniformément distribués, une mesure de consistance est définie par l'utilisation de la probabilité. MINPRAN modélise par des probabilités le fait d'observer un groupe de k résidus plus petit qu'une erreur r parmi N résidus selon une loi uniforme. Le groupe de k points présentant la plus faible probabilité est retenu. Le calcul des probabilités permet d'obtenir une méthode adaptative mais rajoute une complexité importante pour les calculs.

Hypothèses : points d'intérêt indépendants et uniformément distribués dans les images : MOISAN et STIVAL (2004) proposent de mesurer la qualité d'un groupe de correspondances dans le cadre de la théorie de la détection *a contrario*. Cette approche présente de nombreuses similitudes avec l'algorithme MINPRAN mais les hypothèses pour le modèle de fond sont différentes. Cette méthode que nous référons par l'acronyme AC-RANSAC (A Contrario RANdom SAmple Consensus) est expliquée plus en détail dans la section 4.3.

Hypothèses : distributions uniformes des *alpha-consistance* de modèles : StaRSaC (CHOI et MEDIONI 2009) propose de tester de manière exhaustive différentes valeurs de seuil δ. Le consensus conservé est estimé en fonction de la variance des paramètres du modèle $\mathcal{H}$ en estimation. RAGURAM et FRAHM (2011) propose avec la méthode RECON *REsidual CONsensus* de rechercher K hypothèses qui sont consistantes. La mesure de consistance repose sur un test dit d'α-consistance permettant d'identifier la variance du bruit d'un modèle en cours d'évaluation. RECON itère parmi différentes valeurs de seuil α et garde le plus petit α donnant un ensemble de modèles partageant des distributions similaires d'erreur résiduelle. L'inconvénient de ces méthodes est que les seuils sont contraints apriori dans un intervalle fixe et discrétisé en K sous seuils à évaluer.

Amélioration de la précision :

Optimisation locale. CHUM et al. (2003) propose avec LO-RANSAC pour chaque hypothèse en cours d'acceptation de l'optimiser localement. C'est à dire de lancer des estimations d'hypothèses parmi les données sélectionnées en *inlier*. L'hypothèse donnant la plus petite erreur moyenne est retenue.
La méthode de MOISAN et STIVAL (2004) réalise à la fois un échantillonnage guidé et une optimisation locale du modèle.

Nous venons de voir qu'il existe toute une famille de méthodes RANSAC, chaque méthode apporte des optimisations de certaines parties de l'algorithme de base. Chaque méthode a ses avantages et inconvénients : fiabilité et paramètres plus ou moins visibles. Nous allons nous intéresser par la suite à la méthode nommée *AC-RANSAC*, *A Contrario RANSAC*, car elle repose sur la définition et l'usage de critères statistiques d'aide à la décision bien fondés. Les points abordés par *AC-RANSAC* sont particulièrement intéressants et permettent :

- la modélisation statistique du nombre de fausses alarmes,
- l'adaptabilité au bruit des données et donc une meilleure précision pour les modèles identifiés (cf. figure 4.2),
- l'absence de paramètres autres qu'un nombre d'itération maximal,
- un point de rupture plus large que les autres méthodes (pouvant aller jusqu'à 90% d'*outliers* si suffisamment d'hypothèses sont testées).

Des expériences, sur *A Contrario RANSAC*, réalisées par MOISAN et STIVAL (2004) et NOURY (2011) ont démontré sur images synthétiques et réelles :

- l'amélioration du taux de réussite général et ce jusqu'a 90% d'*outliers*,
- l'amélioration générale de la précision (comparé à RANSAC et MSAC),
- l'amélioration systématique de la solution identifiée lorsque le taux d'*outliers* dépasse les 50%.

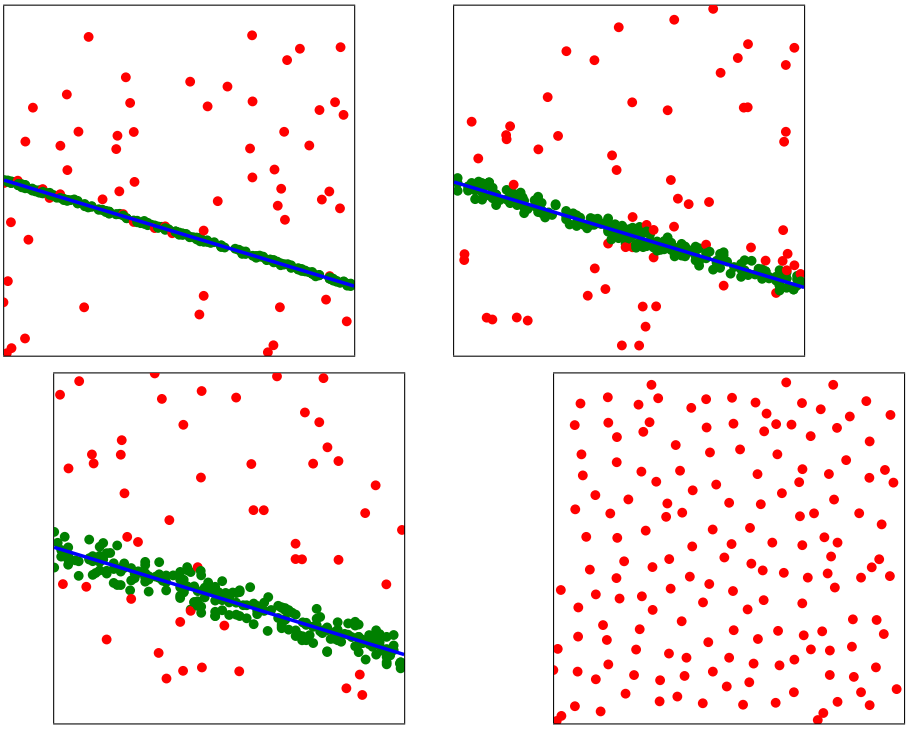


FIGURE 4.2 – Comportement d'*AC-RANSAC* pour la recherche d'un modèle paramétrique de ligne ou un bruit gaussien de plus en plus grand est ajouté. En vert les données validées *a contrario* en rouge les données rejetées et en bleu le modèle identifié. On note, en bas à droite, qu'en présence de bruit pur *AC-RANSAC* n'identifie aucun modèle, RANSAC aurait lui retourné une fausse hypothèse.

4.3 *A Contrario-RANSAC*

La théorie de la détection *a contrario* a été proposée initialement par DESOLNEUX et al. (2000) pour la détection de segments puis généralisée à d'autres propos par la suite : DESOLNEUX et al. (2007). Elle s'inspire des tests d'hypothèses pour détecter des groupes significatifs d'objets partageant des caractéristiques similaires. Les «méthodes a contrario» reposent sur la définition d'**un modèle de fond** et **une mesure de significativité**.

4.3.1 Le principe de la détection *a contrario*

La méthodologie *a contrario (AC)* repose sur le postulat qu'une structure n'est perçue que lorsqu'elle n'a que très peu de chance d'être due au hasard. Ce principe est défini par le «principe de Helmholtz» (cf. figure 4.3). Lionel MOISAN (2003) définit l'idée à exploiter comme suit :

Proposition 3. *"Il est beaucoup plus simple de définir un modèle que l'on souhaite réfuter (typiquement un modèle uniforme) qu'un modèle précis des objets que l'on souhaite détecter".*

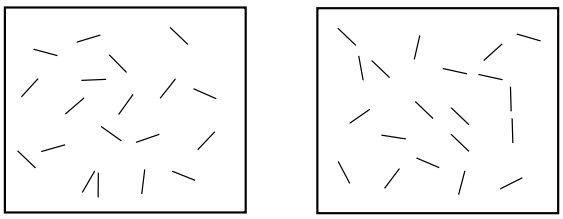


FIGURE 4.3 – Illustration du principe de Helmholtz (groupement perceptuel). A gauche, aucune structure ne se détache de l'image, où les segments ont été tirés aléatoirement de manière indépendante. A droite, on ne peut s'empêcher de regrouper certain segments car les alignements visibles ont peu de chances d'arriver par hasard.

L'application de la méthodologie *a contrario* demande la définition de quatre critères :

1. **Un modèle de fond :** La définition de l'hypothèse à réfuter.
2. **Une mesure de similarité :** Une mesure de l'adéquation d'un échantillon à une hypothèse en cours d'évaluation.
3. **Une mesure de significativité :** Une mesure de l'adéquation d'un groupe d'échantillons à une hypothèse en cours (la détection d'une structure significative).
4. **Un critère d'optimisation :** Optimisation permettant de retenir la meilleure hypothèse rencontrée : celle qui réfute le plus le modèle de fond.

Appliquée à la recherche de modèle, la méthode a contrario répond à la question : "Est-ce que le modèle considéré s'ajuste aux données par chance ?" Le cadre statistique repose sur deux notions : la définition d'un modèle de fond, qui décrit le processus génératif, pour lequel aucune structure significative n'est perçue, et une mesure de similarité de caractéristiques composant un groupe. Cette similarité permet d'évaluer la qualité des groupes testés afin de détecter automatiquement quel sous-groupe est cohérent, rigide.

4.3.2 Mise en correspondance *a contrario* pour l'estimation de la géométrie épipolaire

Dans le but de s'affranchir des limitations de RANSAC, MOISAN et STIVAL (2004) utilisent la méthodologie *a contrario* pour réaliser les tâches de sélection et validation de groupe dans le but d'estimer la géométrie épipolaire à partir de correspondances. Ils apportent les éléments suivants :

Rappel. On dispose d'un ensemble de correspondances $C : \{(m,m')\}, \#C = n$ entre deux images I et I'. On considère qu'un sous-groupe de 7 points est nécessaire pour calculer de 1 à 3 matrices fondamentales.

Le modèle de fond à réfuter, l'hypothèse nulle. On souhaite identifier un sous-groupe de ces correspondances qui peut être expliqué par une unique transformation. Pour estimer cette transformation dans la méthodologie *a contrario*, on définit une hypothèse nulle $\mathcal{H}_0$ qui décrit la distribution des correspondances aléatoire C pour lesquelles aucun groupement ne doit être validé. Un groupe de correspondances est considéré comme significatif s'il réfute l'hypothèse nulle, en d'autres termes si l'observation d'un tel groupe sous $\mathcal{H}_0$ est peu probable.

Proposition 4. *Un ensemble C de n correspondances aléatoires $\{(m,m')\}$ suit l'hypothèse nulle $\mathcal{H}_0$ lorsque :*

- *les correspondances (m,m') sont des variables aléatoires mutuellement indépendantes,*
- *les points m et m' sont uniformément distribués dans leur image respective I, I'.*

La mesure de similarité. La mesure permettant de vérifier la qualité d'un échantillon de correspondances dans le cas de la géométrie épipolaire utilise une erreur de type point-droite. Cette erreur résiduelle implique pour un modèle $\mathbf{F}$ et un couple de point (m,m') une distance des points m et m' aux lignes épipolaires $\mathbf{F}^T m'$ dans I et $\mathbf{F}m$ dans I' respectivement.

La mesure de significativité. On souhaite ici mesurer l'adéquation d'un modèle en cours d'hypothèse aux données de manière statistique. Soit S' un sous-groupe de C, tel que $\#S' = \mathbf{s}$ et $\mathbf{F}_{S'}$ la matrice fondamentale évaluée à partir du $\mathbf{s}$-uplet. Si l'on considère que C suit le modèle de fond et que l'on a estimé la matrice $\mathbf{F}_{S'}$ à partir d'un sous-groupe $S' \subset C$. Pour n'importe quelle correspondance aléatoire (m,m') de C, la probabilité que la distance entre m' et la ligne épipolaire $\mathbf{F}_{S'}m$ soit plus petite que α peut être majorée (cf. figure 4.4). Cette borne supérieure est le rapport entre l'aire maximale d'une bande de largeur 2α et l'aire A de l'image I'.

En notant $\mathcal{M}(\mathbf{F}_{S'}m,m')$ la distance euclidienne entre le point m' et la ligne épipolaire $\mathbf{F}_{S'}m$ on note :

$$\forall \alpha > 0, \mathbb{P}_{\mathcal{H}_0}[\mathcal{M}(\mathbf{F}_{S'}m,m') \leq \alpha] \leq \frac{2\alpha D_{I'}}{A_{I'}} \tag{4.3}$$

où $D_{I'}$ et $A_{I'}$ désignent respectivement la longueur de la diagonale et l'aire de l'image I'. On définit l'erreur symétrique de transfert pour la géométrie épipolaire :

$$\max\left\{\frac{2D_{I'}}{A_{I'}}\mathcal{M}(\mathbf{F}_{S'}m,m'), \frac{2D_I}{A_I}\mathcal{M}(m,\mathbf{F}_{S'}^T m')\right\} \in [0,1]$$

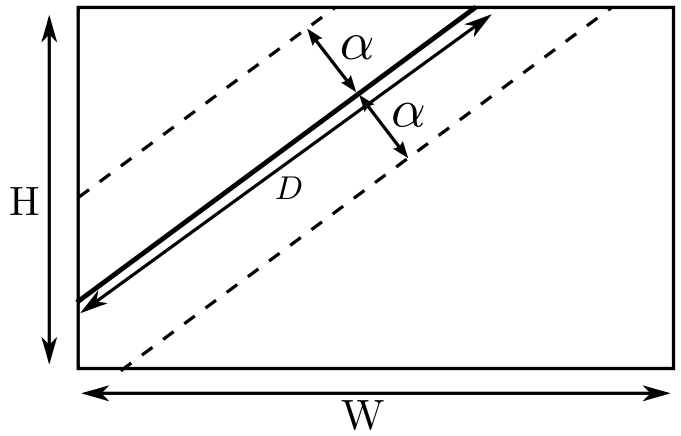


FIGURE 4.4 – Définition de la mesure de significativité de l'erreur α pour la géométrie épipolaire.

Considérant un couple de point aléatoire $(m,m') \subset C \mid S'$ on peut écrire

$$\mathbb{P}_{\mathcal{H}_0}\left[\max\left\{\frac{2D_{I'}}{A_{I'}}\mathcal{M}(\mathbf{F}_{S'}m,m'),\frac{2D_I}{A_I}\mathcal{M}(m,\mathbf{F}_{S'}^T m')\right\} \leq \alpha\right] \leq \alpha^2 \leq \alpha \tag{4.4}$$

Pour tout sous-ensemble S de C tel que $S \cap S' = \emptyset$ on peut ainsi mesurer l'adéquation de la matrice fondamentale $\mathbf{F}_{S'}$ pour les correspondances de S comme l'erreur de transfert symétrique normalisée maximale sur tous les points de S :

$$\alpha(S,\mathbf{F}_{S'}) := \max_{(m,m')\in S}\max\left(\frac{2D_{I'}}{A_{I'}}\mathcal{M}(\mathbf{F}_{S'}m,m'),\frac{2D_I}{A_I}\mathcal{M}(m,\mathbf{F}_{S'}^T m')\right) \tag{4.5}$$

On appelle cette mesure l'α-rigidité. Plus α est petit, moins il est probable que l'ensemble de correspondances soit dû au hasard. Les correspondances étant supposées indépendantes, on obtient une borne $\alpha^{\#S}$ pour la probabilité d'observer une rigidité $\alpha(S,\mathbf{F}_{S'})$:

$$\forall \alpha > 0, \quad \mathbb{P}_{\mathcal{H}_0}\left[\alpha(S,\mathbf{F}_{S'}) \leq \alpha\right] \leq \alpha^{\#S} \tag{4.6}$$

On peut ainsi mesurer la cohérence d'un sous-groupe S de correspondances réelles selon une transformation $\mathbf{F}$ en considérant que la probabilité de la rigidité aléatoire de $\alpha(S,\mathbf{F}_{S'})$ soit plus petite que la rigidité observée $\alpha(S,\mathbf{F})$ sous l'hypothèse nulle $\mathcal{H}_0$. La quantité $\alpha(S,\mathbf{F})^{\#S}$ mesure à quel point on s'étonne d'observer un groupe de taille $\#S$ et de rigidité $\alpha(S,\mathbf{F})$ en supposant que le groupe est généré aléatoirement. Comme on recherche des groupes qui ne sont pas composés de bruit, seuls les groupes pour lesquels la probabilité est faible seront validés.

Une optimisation. Afin de connaître de manière automatique quel sous-groupe S est α-rigide, un critère de validation automatique est utilisé. Ce critère de validation s'appuie sur l'espérance du nombre de fausses alarmes, le *NFA* : une probabilité α pondérée par un nombre de tests. Cette mesure de qualité associe une borne supérieure de l'espérance du nombre de fausses alarmes au nombre de groupes de taille k de S qui suivent le modèle de fond :

Proposition 5. *Soit $C = \{(m_i,m_i')|i=1,\ldots,n\}$ un ensemble de n appariements entre les images I et I'. Soit S un sous-ensemble de C, constitué de $\#S = k$ correspondances, avec $k \leq n-7$. L'ensemble S est dit ε-significatif s'il existe un sous-ensemble S' de C, tel que $\#S' = 7, S' \cap S = \emptyset$ et*

$$NFA(S,\mathbf{F}_{S'},k) = 3(n-7)\binom{n}{k}\binom{k}{7}\alpha^{k-7} \leq \varepsilon. \tag{4.7}$$

Le *NFA* permet d'estimer quel sous-groupe de taille k réfute l'hypothèse de fond $\mathcal{H}_0$. On mesure ainsi l'α-rigidité d'un sous-groupe de S de taille k pour la matrice $\mathbf{F}_{S'}$. Cette mesure est d'autant plus significative que la quantité $NFA(S,\mathbf{F}_{S'},k)$ est faible. Le nombre de tests utilisé est composé de :

1. **le nombre de tirage aléatoire** : Le terme $3\binom{k}{7}$ correspond au nombre de transformations $\mathbf{F}$ qu'il est possible d'estimer parmi les k correspondances restantes. Le nombre de 7-uplets multiplié par le nombre de modèles hypothèses maximum pouvant être calculé.
2. **le nombre de groupe de résidus** $(n-7)$. Les appariements restants dont les erreurs résiduelles sont ordonnées par ordre croissant (les différentes bornes supérieures α à évaluer),
3. **le nombre de groupe de taille** $k \leq n-7$: le terme $\binom{n}{k}$.

Identifier le groupe optimal pour la matrice $\mathbf{F}_{S'}$ consiste à trouver le nombre de valeurs étant le plus α-consistant : le groupe S_k ayant le plus petit NFA. Soit, rechercher le groupe de taille k minimisant l'équation $NFA(S,\mathbf{F}_S,k)$:

$$NFA(S_k) = \min_{k=8\ldots n} NFA(S,\mathbf{F}_{S'},k) \leq \varepsilon\ , \tag{4.8}$$

avec $\varepsilon = 1$ comme borne naturelle pour indiquer que l'on autorise au plus une fausse alarme par détection. Les sous-groupes sont explorés en faisant varier $k \in [8,n]$.

Tester tous les sous-ensembles de 7-uplets n'étant pas envisageable, il convient d'utiliser les mêmes idées que RANSAC pour créer l'algorithme *AC-RANSAC* (cf. procédure 3). Á chaque itération un 7-uplets S est tiré parmi les n correspondances. De une à trois matrices fondamentales sont alors estimées. Pour chacune on recherche le sous-groupe le plus α-consistant : Les erreurs pour les $n-7$ appariements restant $(m_i,m_i') \in C \mid S$ sont évaluées et ordonnées par ordre croissant puis le groupe de taille k optimal est identifié. On itère jusqu'à ce qu'un nombre maximal d'itérations ait été atteint ou que l'on a identifié une hypothèse donnant un $NFA < 1$, phase où l'on va pouvoir optimiser localement le modèle pour continuer à identifier de nouvelle matrice $\mathbf{F}$ ayant sous-groupe avec un *NFA* plus petit.

On obtient la procédure 3 :

Procédure 3 *AC-RANSAC*

Entrée: $\mathcal{D} = \{(m, m')\}$: un ensemble de correspondances
Entrée: N : un nombre maximal de tirage
Sortie: l'ensemble consensus S_{opt}, le modèle $\mathbf{F}_{opt}$ validé *a contrario* et son *NFA*.

iter = 0, $S_{opt} = \emptyset$, optim = 0
$NFA_{opt} = 1$
$\mathcal{D}_{\text{copie}} = \mathcal{D}$

(1) **Échantillonnage aléatoire :**
Tirage d'un 7-uplet S parmi $\mathcal{D}_{\text{copie}}$
Estimation de(s) matrices **F** (au plus 3)
pour chaque matrice **F faire**

(2) **Sélection des *inliers* :**
Tri des correspondances (m, m') selon leur erreur résiduelle α_i
Sélection du groupe S'de taille k minimisant le $NFA(S, \mathbf{F}, \alpha_i, k)$

(3) **Validation :**
si $\#S' > \#S_{opt}$ et $NFA(S') < NFA_{opt}$ **alors**
$S_{opt} = S'$
$\mathbf{F}_{opt} = \mathbf{F}$

(3.1) **Optimisation du modèle et réduction du nombre d'itération :**
si $NFA(S') < 1$ et optim = 0 **alors**
$\mathcal{D}_{\text{copie}} = S'$
$N = \text{iter} + N/10$;
optim = 1
fin si
fin si
fin pour

(4) **Critère d'arrêt :**
Tant que $iter < N$, $iter = iter + 1$. Retour à l'étape 1.

(5) **Optimisation du modèle final :**
Estimation aux moindres carrés de F_{opt} en utilisant S_{opt}.

4.4 Généralisation de la mise en correspondance *a contrario* pour l'estimation de modèles paramétriques

Le modèle de fond proposé pour la géométrie épipolaire par MOISAN et STIVAL (2004) est très générique (indépendance mutuelle et distribution uniforme des points homologues), il peut donc être utilisé pour l'estimation de modèles paramétriques autres que la matrice fondamentale. Dans un premier temps nous nous intéressons au cas des transformations géométriques du plan. Puis nous proposons d'explorer une formulation générique du calcul *NFA* pour appliquer l'estimation robuste *a contrario* *AC-RANSAC* à des modèles inexplorés jusqu'alors.

Étendre la formulation *a contrario* initiale pour le cas des transformations géométriques du plan (similitudes, transformations affines et homographie) requiert de redéfinir les points suivants :

1. **La mesure de similarité** : la mesure de l'erreur résiduelle,

   Les transformations géométriques du plan impliquent non plus une distance à une ligne épipolaire, mais une correspondance point à point. Soit $\mathbf{M}_p$ un modèle paramétrique réalisant une transformation géométrique du plan. L'erreur résiduelle de transfert dans l'image droite s'exprime par la distance euclidienne entre le point x' de l'image droite et le point $\mathbf{M_p}x$, transfert du point x de l'image gauche à l'image droite :

   $$\mathcal{M}(\mathbf{M}_p x, x') = \|\mathbf{M}_p x - x'\|_2. \tag{4.9}$$

   Pour tout correspondance (x,x') la probabilité conditionnellement à $\mathcal{H}_0$ que la distance $\mathcal{M}(\mathbf{M}_p x, x')$ soit plus petite que α est bornée supérieurement par le rapport de l'aire du disque de rayon α divisé par l'aire A' de l'image I' :

   $$\forall \alpha > 0, \quad \mathbb{P}_{\mathcal{H}_0}[\mathcal{M}(\mathbf{M}_p x, x') \leq \alpha] \leq \pi \frac{\alpha^2}{A'}. \tag{4.10}$$

   Autrement dit,

   $$\forall \alpha > 0, \quad \mathbb{P}_{\mathcal{H}_0}[\frac{\pi}{A'} \mathcal{M}(\mathbf{M_p} x, x')^2 \leq \alpha] \leq \alpha. \tag{4.11}$$

2. **La mesure de significativité** : la mesure de l'α-rigidité,

   Une nouvelle définition de la rigidité (en considérant les erreurs de transfert dans les deux images) est exprimée :

   $$\alpha(S, \mathbf{M_p}_{S'}) := \max_{(x,x') \in S} \max \left(\frac{\pi}{A'} \mathcal{M}(\mathbf{M}_p x, x')^2, \frac{\pi}{A} \mathcal{M}(x, \mathbf{M}_p^{-1} x')^2 \right) \tag{4.12}$$

   La cohérence d'un sous-groupe S' de correspondances réelles selon une transformation $\mathbf{M_p}$, en considérant la probabilité que la rigidité aléatoire de $\alpha(S, \mathbf{M_p}_{S'})$ soit plus petite que la rigidité observée $\alpha(S, \mathbf{M_p})$ sous l'hypothèse nulle $\mathcal{H}_0$, est mesurée comme précédemment :

   $$\forall \alpha > 0, \quad \mathbb{P}_{\mathcal{H}_0} \left[\alpha(S, \mathbf{M}_{pS'}) \leq \alpha \right] \leq \alpha^{\#S} \tag{4.13}$$

3. **La phase d'optimisation** : le calcul du *NFA*.

   En considérant maintenant un ensemble S de n appariements entre deux images I et I' on exprime de manière analogue le critère de validation du *NFA* proposé par MOISAN et STIVAL (2004) :

   $$NFA(S, \mathbf{M}_{pS'}, k) = (n - N_s) \binom{n}{k} \binom{k}{N_s} \alpha^{k - N_s} \leq \varepsilon, \tag{4.14}$$

N_s étant le nombre d'appariements nécessaires pour l'estimation d'un modèle $\mathbf{M_p}$. La taille de ce s-uplet est définie à $s = 2$ pour une similitude, $s = 3$ pour une affinité et $s = 4$ pour une homographie $\mathbf{H}$. On optimise alors :

$$NFA(S_k) = \min_{k=N_{s+1}\ldots n} NFA(S, \mathbf{M}_{pS'}, k) \leq \varepsilon \ ,$$

On observe deux changements par rapport à la formulation originale de MOISAN et STIVAL (2004) :

1. L'introduction de N_s qui correspond à la taille du s-uplet nécessaire pour l'estimation d'un modèle $\mathbf{M_p}$.
2. La probabilité qui dépend du carré de la distance entre deux points (4.9).

Ces changements ont été démontrés et évalués par RABIN (2009) ; NOURY (2011) mais il n'y a pas eu de généralisation à d'autres modèles paramétriques.

Afin de généraliser *AC-RANSAC* au sein d'une procédure unique, nous proposons une formulation générique du calcul du *NFA* pour une interopérabilité plus aisée (MOISAN et al. 2012).

4.4.1 Généralisation du calcul du *NFA* et utilisations

Si nous résumons les changements constatés entre la modélisation initiale et l'extension réalisée pour les contraintes de correspondances planaires on observe :

1. Que la modélisation de la probabilité est différente. On note l'apparition de facteurs de pondération différents de la métrique $\mathcal{M}$;
2. L'apparition d'un s-uplet de tailles variable en fonction du modèle paramétrique estimé ;
3. L'évaluation d'un nombre de modèles candidats différe en fonction du modèle paramétrique. De 1 à 3 modèles hypothèses sont générés pour un 7-uplet dans le cas de la matrice fondamentale alors qu'un seul modèle hypothèse est estimé à partir d'un 4-uplet dans le cas de la matrice homographique.

Ces observations nous permettent de formuler une généralisation du *NFA* de Moisan et Stival. La mesure du nombre de fausses alarmes pour un modèle M à k *inliers* sous l'hypothèse nulle $\mathcal{H}_0$ devient :

$$NFA(\mathbf{M}, k) = N_o(n - N_s)\binom{n}{k}\binom{k}{N_s}(\alpha_0 e_k(\mathbf{M})^d)^{k-N_s} \tag{4.15}$$

Avec
- N_o : nombre maximal de modèles pouvant être estimés à partir d'un s-uplet,
- k : taille du sous-groupe en cours d'évaluation (l'ensemble supposé *inlier*),
- n : taille du groupe de correspondances,
- $N_s = s$: taille du s-uplet nécessaire pour l'estimation d'un modèle $\mathbf{M}$,
- $e_k(\mathbf{M})$: la k^e plus petite erreur résiduelle pour le modèle $\mathbf{M}$ parmi les n correspondances,
- α_0 : probabilité d'une correspondance aléatoire d'avoir une erreur unité ou majorant de cette probabilité.
- d : dimension de l'erreur ($d = 2$ pour une distance point à point).

On optimise alors pour chaque hypothèse modèle $\mathbf{M}_{S'}$:

$$NFA(S_k) = \min_{k=N_s+1\ldots n} NFA(S, \mathbf{M}_{S'}, k) \leq \varepsilon. \tag{4.16}$$

En fonction de la métrique associée à un modèle paramétrique **M** on obtient alors la classification récapitulée dans le tableau 4.1.

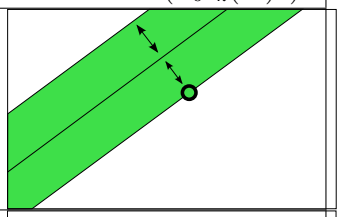
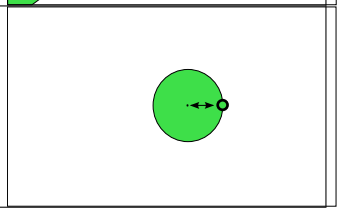
Type d'erreur	α_0	d	Probabilité $(\alpha_0 e_k(M)^d)$
point-ligne	$2D/A$	1	
point-point	π/A	2	

TABLE 4.1 – Définition des paramètres α_0 et d en fonction de l'erreur utilisée par la métrique de calcul d'erreur résiduelle et représentation graphique de la probabilité des erreurs résiduelles.

L'utilisation de cette classification nous permet ainsi de définir la procédure 4 d'*AC-RANSAC* générique, où le *NFA* est évalué de manière dynamique en fonction du type de modèle paramétrique considéré. Une extension directe à de nouveaux modèles paramétriques non explorés jusqu'à alors est ainsi obtenue pour l'estimation robuste sans précision a priori de matrices essentielles **E**, de matrices de projection **P** et de tenseurs tri-focaux **T** (cf. tableau 4.2).

Type de modèle	Métrique $\mathcal{M}$	#Images utilisées	Publications
F	point-ligne	2	MOISAN et STIVAL (2004)
H	point-point	2	RABIN (2009) ; NOURY (2011) ; MOISAN et al. (2012)
$*$**E**	point-ligne point-point	2 2	MOULON et al. (2013a)
$*$**P**	point-point	1	MOULON et al. (2013a)
$*$**T** $*$**T** depuis 2 **F** $*$**T** réduit depuis 2 rotations	point-point point-point point-point	3 3 3	MOULON et al. (2013b)

TABLE 4.2 – Récapitulatif des modèles supportés par notre généralisation de la méthodologie d'estimation robuste *a contrario*. Les modèles paramétriques marqués d'une $*$ sont utilisés pour la première fois dans un contexte *a contrario*.

Procédure 4 *AC-RANSAC* générique

Entrée: $\mathcal{D} = \{(m, m')\}$: un ensemble de correspondances
Entrée: N : un nombre maximal de tirage
Sortie: l'ensemble consensus S_{opt}, le modèle $\mathbf{M}_{opt}$ validé *a contrario* et son *NFA*.
iter = 0, $S_{opt} = \emptyset$, optim = 0
$NFA_{opt} = 1$
$\mathcal{D}_{\text{copie}} = \mathcal{D}$
(1) **Échantillonnage aléatoire :**
tirage d'un s-uplets S parmi $\mathcal{D}_{\text{copie}}$
Estimation des N_o modèles paramétriques $\mathbf{M}$ hypothèse
(2) **Sélection des *inliers* :**
pour $j \in 0, ..., N_o$ **faire**
$e_k(\mathbf{M}_j)$:= tri des appariements (m, m') selon leur erreur résiduelle $\alpha(m, m', \mathbf{M}_j)$
sélection du groupe S'de taille k minimisant le $NFA(\mathbf{M}_j, k)$
(3) **Validation :**
si $\#S' > \#S_{opt}$ et $NFA(S') < NFA_{opt}$ **alors**
$S_{opt} = S'$
$\mathbf{M}_{opt} = \mathbf{M}_j$
fin si
(3.1) **Optimisation du modèle et réduction du nombre d'itération :**
si $NFA(S') < 1$ et optim = 0 **alors**
$\mathcal{D}_{\text{copie}} = S'$
$N = \text{iter} + N/10$;
optim = 1
fin si
fin pour
(4) **Critère d'arrêt :**
tant que $iter < N$, $iter = iter + 1$. Retour à l'étape 1.
(5) **Optimisation du modèle final :**
Estimation aux moindres carrés de $\mathbf{M}_{opt}$ en utilisant S_{opt}.

4.4.2 Application pour l'estimation de la géométrie relative entre deux images sphériques

Les images panoramiques sphériques permettent de représenter l'environnement entourant un point de l'espace sous la forme d'une seule image. Au départ utilisé afin de représenter un environnement et générer des éclairages réalistes en synthèse d'images, de nouvelles applications utilisant plusieurs panoramas sont apparues. SHUM et al. (1998) utilisent des saisies utilisateur afin de reconstruire un environnement 3D simple à partir d'un ou plusieurs panoramas. TORII et al. (2005) présentent un modèle de caméra sphérique permettant de modéliser des contraintes géométriques entre deux et trois panoramas. En utilisant cette paramétrisation de caméra, PAGANI et STRICKER (2011) proposent une approche automatique pour la reconstruction 3D d'un environnement à partir d'une séquence de panoramas ainsi qu'une comparaison de différentes métriques pour mesurer l'adéquation des re-projections images. Récemment COLBERT et al. (2012) proposent un système permettant de construire des visites virtuelles à partir d'une séquence de panoramas. Des panoramas verticalement alignés sont acquis, puis une variante de RANSAC aligne en relatif les panoramas entre eux dans le plan. L'utilisateur peut ensuite éditer les mouvements relatifs pour corriger les positionnements relatifs des différents panoramas estimés.

Dans le but de pouvoir calculer le positionnement relatif de deux panoramas sans précision apriori nous proposons d'appliquer la méthodologie *a contrario* pour l'estimation d'une matrice essentielle entre deux images sphériques.

Estimation *a contrario* de la géométrie essentielle entre deux images sphériques.

Nous considérons une caméra sphérique identique à la modélisation proposée par (TORII et al. 2005). Une caméra sphérique est une caméra centrale où une collection de rayons passe par un seul point de l'espace : le centre de la caméra. L'image panoramique sphérique est obtenue par projection de l'espace sur la surface de la sphère d'unité 1 centrée à la même position que le centre de caméra.

La pose de la caméra (position et orientation) est définie par une rotation R et une translation $\mathbf{t}$ dans un système de coordonnées global. Un point monde X est défini dans le repère local loc de la caméra comme suit :

$$X_{loc} = RX + \mathbf{t} \tag{4.17}$$

Ce point X_{loc} se projette sur la sphère en $\mathbf{x}$ tel que :

$$\mathbf{x} = \frac{1}{|X_{loc}|} X_{loc}. \tag{4.18}$$

Un point image x est ainsi projeté sur la sphère en un point m exprimé en coordonnées sphériques :

$$\mathbf{m} = \begin{pmatrix} cos\theta sin\phi \\ sin\theta sin\phi \\ sin\phi \end{pmatrix} \tag{4.19}$$

L'image sphérique est définie de manière paramétrique par deux angles : $I(\theta, \phi)$ avec $0 \leq \theta \leq 2\pi$ et $0 \leq \phi \leq \pi$. Chaque point de l'image est ainsi associé à un rayon dans l'espace (cf. figure 4.5).

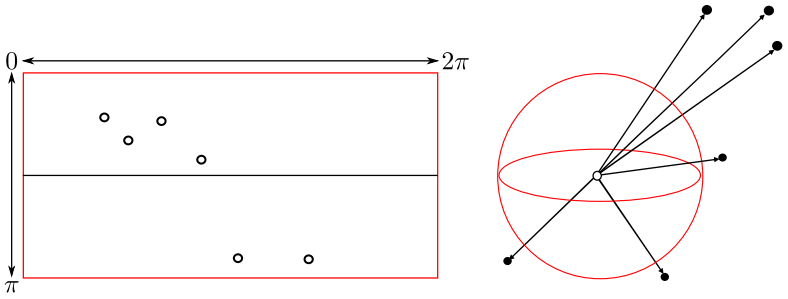


FIGURE 4.5 – Le modèle de caméra sphérique : une sphère d'unité 1. Chaque point de l'image panoramique sphérique de taille $(w,h) = (w,w/2)$ représente un rayon de l'espace. La projection image x d'un point de l'espace X est ainsi l'intersection de la sphère avec le rayon $\overrightarrow{CX}$.

La géométrie relative entre les deux images sphériques est représentée par la matrice essentielle $\mathbf{E}$ (cf. figure 4.6) et peut être estimée à partir de 8 correspondances de points :

$$\mathbf{m}'^T\mathbf{E}\mathbf{m} = 0, \quad \mathbf{E} = [\mathbf{t}]_\times R. \tag{4.20}$$

$\mathbf{m}$ et $\mathbf{m}'$ étant les vecteurs normalisés des coordonnées sphériques représentant les points images x et x'.

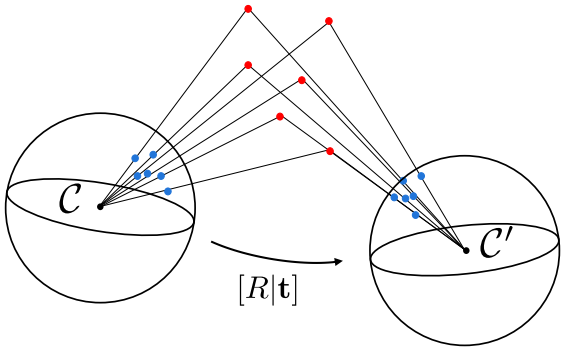


FIGURE 4.6 – Représentation de la géométrie relative entre deux images sphériques.

Pour estimer cette géométrie relative sous la méthologie *a contrario* nous devons définir :

1. **La mesure de similarité** : la mesure de l'erreur résiduelle,
2. **La mesure de significativité** : la mesure de l'α-rigidité,
3. **La phase d'optimisation** : le calcul du *NFA*.

L'erreur épipolaire peut s'exprimer sous une forme angulaire :

$$\alpha = \mathcal{M}(\mathbf{E}\mathbf{m}, \mathbf{m}') = \cos^{-1}\frac{\mathbf{m}'^T\mathbf{E}\mathbf{m}}{\|\mathbf{m}'\|\|\mathbf{E}\mathbf{m}\|} \tag{4.21}$$

qui exprime la distance géodésique entre le plan épipolaire et le point m' sur la surface de la sphère. Pour tout correspondance (x,x') la probabilité conditionnellement à $\mathcal{H}_0$ que la distance $\mathcal{M}(\mathbf{E}\mathbf{m},\mathbf{m}')$ soit plus petite que α est bornée supérieurement par le rapport de l'aire de la calotte d'angle α divisé par l'aire d'une sphère unité :

$$\forall \alpha > 0, \quad \mathbb{P}_{\mathcal{H}_0}[\mathcal{M}(\mathbf{E}\mathbf{m},\mathbf{m}') \leq \alpha] \leq \frac{2\pi r^2}{4\pi r^2}(1-\cos\alpha) = \frac{1}{2}(1-cos\alpha). \tag{4.22}$$

On note que le NFA générique (4.15) est formulé pour des erreurs dans le plan, mais ici nous avons des erreurs sur une sphère. Le calcul de la probabilité $(\alpha_0 e_k(M)^d)$ ne peut donc pas être généralisé dans ce cas-ci. Cependant en considérant que seuls les petits angles α sont intéressants on peut approximer $(1-\cos\alpha)$ par son développement de Taylor au deuxième ordre : $\frac{\alpha^2}{2}$, ce qui aboutit à l'approximation de notre terme α_0 à la valeur $\frac{1}{4}$. Autrement dit,

$$\forall \alpha > 0, \mathbb{P}_{\mathcal{H}_0}[\frac{1}{4} \leq \alpha] \leq \alpha \tag{4.23}$$

Cependant en toute rigueur on devrait modéliser la probabilité par $\frac{1}{2}(1-\cos\alpha)$. En reliant ses notations à la formule générique du *NFA* (4.15) on obtient la configuration des paramètres N_o, α_0, d du tableau 4.3 pour calculer la géométrie essentielle entre deux images sphériques *a contrario*.

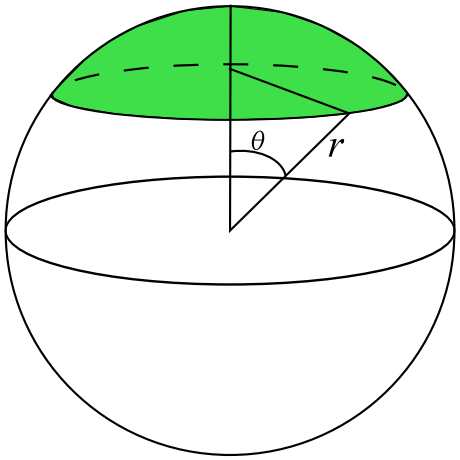
Type d'erreur	N_o	α_0	d	Probabilité $(\alpha_0 e_k(M)^d)$
angle-angle	1	$\frac{1}{4}$	2	

TABLE 4.3 – Définition des paramètres pour la formulation générique du *NFA* (eq. 4.15) pour des erreurs résiduelles angulaires.

Remarque : Une fois la meilleure matrice essentielle identifiée, la décomposition de la matrice $\mathbf{E}$ forme 4 couples de rotation et translation $[R|\mathbf{t}]$ (cf. (HARTLEY et ZISSERMAN 2000)). Le meilleur couple $[R|\mathbf{t}]$ solution est celui qui présente le plus de point 3D devant les caméras. Les caméras étant ici sphériques, le test de visibilité vérifie que les rayons sont dirigés vers le point 3D en vérifiant la positivité des produits scalaires $m^T X_{loc}$ et $m'^T X_{loc}$.

4.4.3 Évaluation expérimentale

Cette section est consacrée à la validation expérimentale du critère *a contrario* générique qui vient d'être présenté. Cette validation repose sur une analyse comparative de ses performances sur plusieurs jeux d'images pour tester différents scénarios.

1. Réaction à la taille du contexte image,
2. Vérification de fonctionnement de l'estimation *a contrario* pour la géométrie trifocale,
3. Vérification de fonctionnement de l'estimation *a contrario* pour la géométrie essentielle de caméra sphérique avec des images réelles et des images synthétiques.

Expérience 1 : Réaction à la taille du contexte image. L'utilisation d'une méthode adaptative doit faire réagir le seuil de validation en fonction de la taille d'image. Nous proposons de vérifier la réaction du seuil identifié *a contrario* pour l'estimation robuste d'une homographie sur une scène identique mais à différentes résolutions. Le contexte de l'expérience est illustré sur la figure 4.7 pour une résolution d'image 800*600.

FIGURE 4.7 – De gauche à droite : les deux images considérées, les images désignant les correspondances validées et rejetées *a contrario* et enfin la mosaïque combinant les deux images d'entrée avec l'homographie identifiée.

L'expérience d'ajustement d'une homographie *a contrario* est réalisée pour une réduction de résolution d'un facteur 2 à chaque itération. Sont comparés : le seuil identifié *a contrario*, la valeur moyenne des erreurs du consensus identifié et l'erreur moyenne une fois que le modèle est été ajusté par moindres carrés à l'ensemble consensus *inlier*. On constate dans le résultat de l'expérience (cf. table 4.4) que le seuil identifié *a contrario* diminue en relation avec la taille d'image. Les rapports des seuils étant compris entre 1.7 et 2.2, on observe qu'ils sont bien en cohérence avec le facteur 2 de réduction. L'estimation *a contrario* réagit bien de manière adaptative tout en découpant correctement les correspondances en ensembles *inliers* et *outliers* valides.

Taille d'image / Erreurs résiduelles	1600*1280	800*640	400*320	200*160
δ_{ac} : seuil identifié *a contrario*	15	6.8	3.9	2
erreur moyenne du consensus validé	6.7	3.7	2.2	1.2
erreur moyenne raffinée	5.8	3.1	1.8	0.9

TABLE 4.4 – Évaluation de la variation des seuils de détection *a contrario* de la géométrie homographique pour une scène identique à taille d'image variable. Erreurs en pixels.

Expérience 2 : Vérification de fonctionnement de l'estimation *a contrario* pour la géométrie tri-focale. Nous proposons de vérifier la réaction du seuil identifié *a contrario* pour l'estimation robuste d'une géométrie tri-focale sur une scène identique mais à différentes résolutions. Le contexte de l'expérience est le même que l'expérience précédente (illustrée en figure 4.8 pour la taille 708*532).

FIGURE 4.8 – De haut en bas : les trois images considérées, les images désignant les correspondances validées et rejetées *a contrario*, respectivement en vert puis rouge. Sur l'image du bas, les erreurs résiduelles sont affichées en jaune pour montrer où devrait se trouver le point pour satisfaire le modèle identifié *a contrario*.

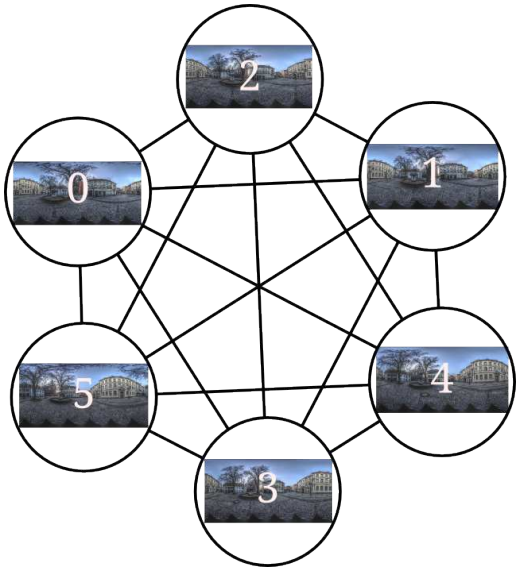
On remarque dans le tableau 4.8 que les seuils identifiés varient avec des rapports proche de 2 : $\delta_{ac} \in [1.4, 3.1]$ et que lorsque le modèle est raffiné les erreurs moyennes sont encore plus proches de ce facteur de réduction $f \in [1.7, 2.1]$. L'estimation *a contrario* réagit bien de manière adaptative tout en découpant correctement les correspondances en ensembles *inliers* et *outliers* valides. Les seuils identifiés *a contrario* varient avec une plus grande amplitude que l'expérience des homographies, mais comme cette expérience implique trois domaines image, elle est donc sujette à plus de bruit de détection sur les points saillants utilisés. Ce comportement est donc tout à fait normal.

Taille d'image / Erreurs résiduelles	2832*2128	1416*1064	708*532	354*266
δ_{ac} : seuil identifié *a contrario*	8.1	4	2.8	0.9
erreur moyenne du consensus validé	0.7	0.3	0.2	0.1
erreur moyenne raffinée	0.4	0.19	0.11	0.06

TABLE 4.5 – Évaluation de la variation des seuils de détection *a contrario* de la géométrie tri-focale pour une scène identique à taille d'image variable. Erreurs en pixels.

Expérience 3 : Vérification de fonctionnement de l'estimation *a contrario* pour la géométrie essentielle de caméra sphérique.

Cette expérience propose de vérifier que l'estimation de la géométrie essentielle robuste *a contrario* est fonctionnelle. La géométrie essentielle robuste est estimée pour une série d'images acquises sur une place de village [1]. On propose de vérifier que les seuils sont différents à chaque estimation pour démontrer l'indépendance des expériences (cf. tableau 4.6). Afin d'avoir un retour visuel de la qualité de la matrice essentielle identifiée *a contrario* nous affichons les reconstructions associées aux scènes testées en figures 4.9,4.10.



Scènes	Seuils *a contrario*(°)	Seuils pixels
0-1	0.071	2.81
0-2	0.123	4.86
0-3	0.071	2.80
0-4	0.111	4.37
0-5	0.120	4.71
1-2	0.085	3.36
1-3	0.118	4.66
1-4	0.093	3.65
1-5	0.101	4.00
2-3	0.100	3.95
2-4	0.097	3.84
2-5	0.095	3.74
3-4	0.100	3.95
3-5	0.133	5.25
4-5	0.075	2.95

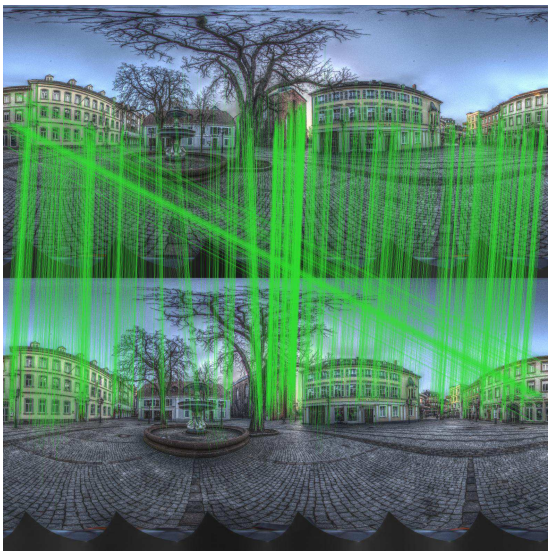
TABLE 4.6 – Variation des seuils identifiés *a contrario* pour la mise en correspondance de différentes images panoramiques sphériques de résolution 14142*7071.

On note que :

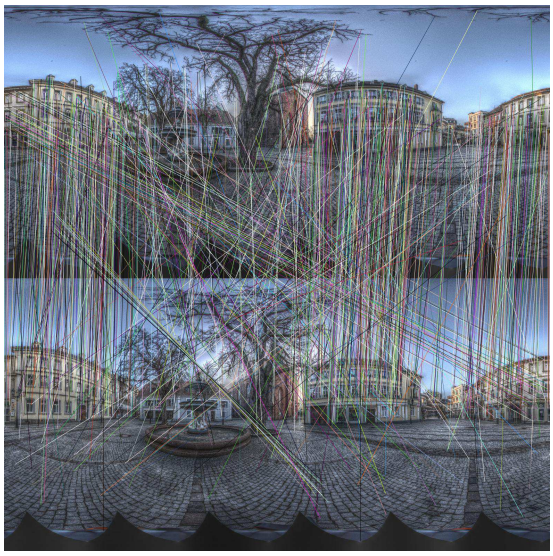
- des seuils différents sont identifiés à chaque expérience,
- les reconstructions associées aux expériences des figures 4.9,4.10 sont cohérentes.

L'approximation utilisée pour le calcul de la probabilité est donc tout à fait cohérente pour notre cas d'utilisation et démontre un bon comportement sur l'ensemble des expériences.

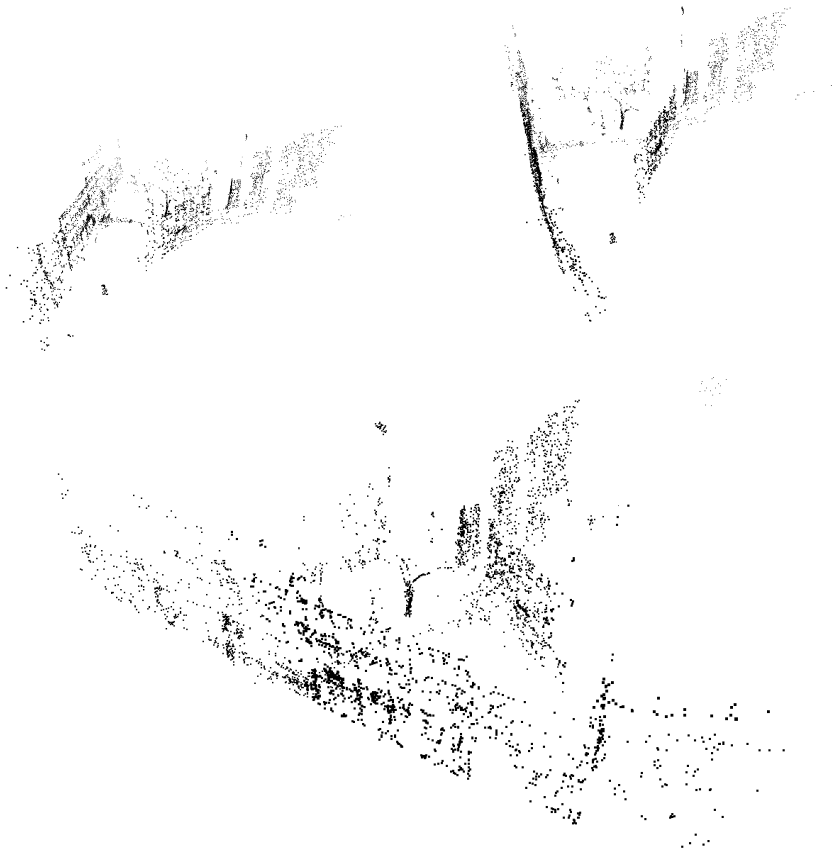
1. Remerciements à Bernd Krolla ainsi qu'aux auteurs PAGANI et STRICKER (2011) du laboratoire *DFKI - German Research Center for Artificial Intelligence* pour avoir fourni ces images.



(a) 4467 *inliers* identifiés automatiquement parmi 5307 correspondances sous une précision de 0.07°, soit 2.81 pixels.

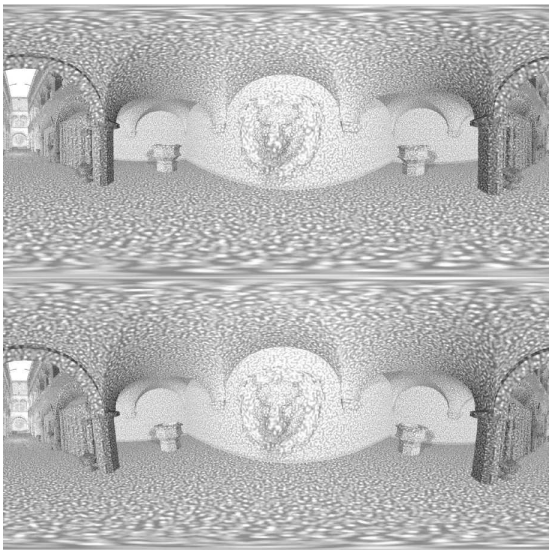


(b) Correspondances rejetées : 840.

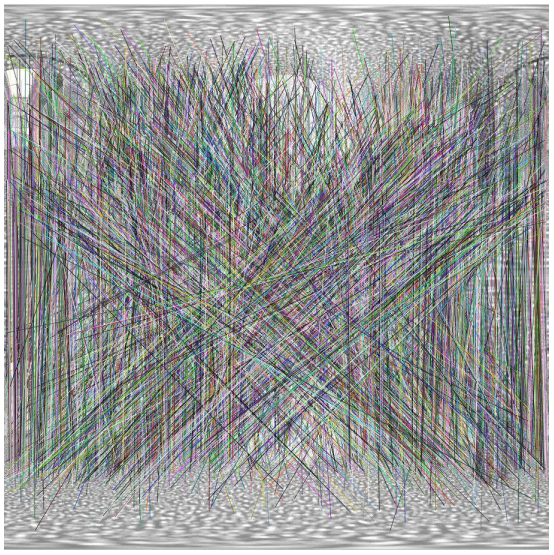


(c) 3 vues de la triangulation du consensus *inliers* avec la matrice essentielle identifiée *a contrario*. La structure fait apparaitre l'arbre et les façades, et montre la structure triangulaire de la place.

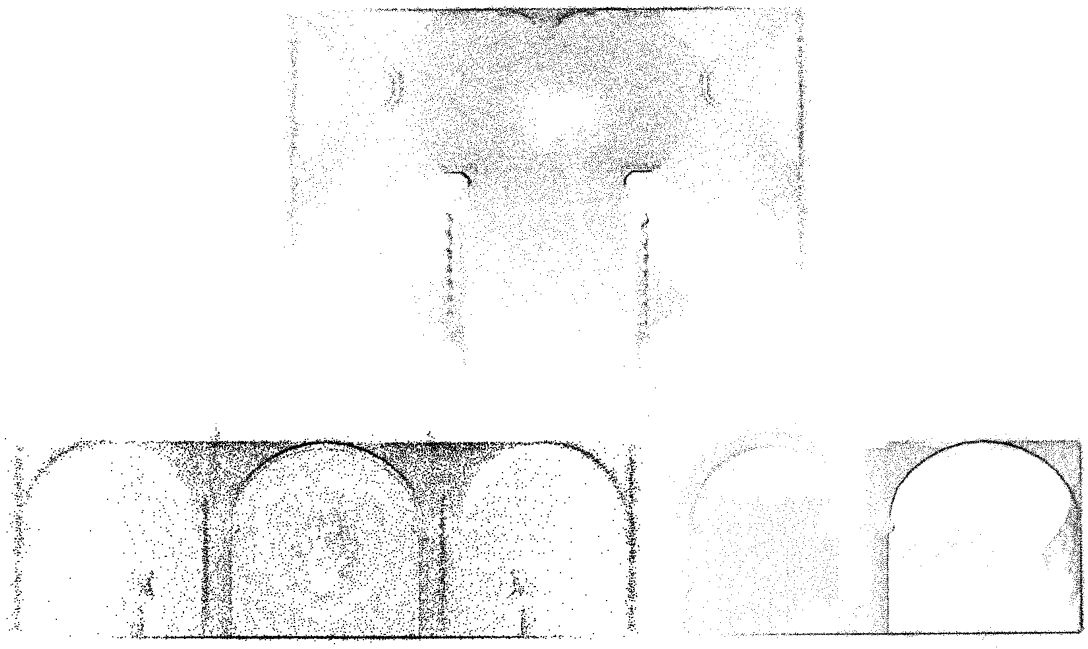
FIGURE 4.9 – Détail de l'expérience 0-1 du tableau 4.6 (cf. note 1 pour la source des images).



(a) Images utilisées pour l'estimation. 11209 *inliers* identifiés automatiquement parmi 14849 correspondances. Le seuil identifié *a contrario* est de 0.13°, soit 0.74 pixels.



(b) Correspondances rejetées : 3640.



(c) Triangulation du consensus *inliers* avec la matrice essentielle identifiée *a contrario*. Vue de haut, face et de droite, la structure fait apparaitre les voûtes, les colonnes et la tête de lion.

FIGURE 4.10 – Utilisation de la scène Sponza après un rendu panoramique de résolution 2048*1024 dans un logiciel de synthèse d'images. L'utilisation d'images synthétiques génère des appariements SIFT de meilleure précision et donc un seuil identifié *a contrario* plus faible que les expériences précédentes.

Les expériences ont prouvé que chaque situation donnait bien lieu à un seuil différent validé *a contrario*. Le fait de considérer chaque situation comme indépendante permet ainsi d'éviter des situations de sous/sur-estimation que l'on aurait rencontré avec l'usage de RANSAC et son seuil fixe. L'usage d'*AC-RANSAC* et son adaptabilité à la taille d'image et aux données est un réel avantage pour réaliser une estimation robuste stable.

4.5 Contributions de ce chapitre

Dans ce chapitre nous avons présenté et comparé sur le plan théorique les différentes méthodes d'estimations robustes de type MAX-CONSENSUS et RANSAC. Nous nous sommes attardés sur l'ambiguïté que pose le choix d'un seuil fixe : l'efficacité relative. Puis nous avons présenté et généralisé le formalisme *a contrario* et la méthode *AC-RANSAC* à l'estimation de modèles paramétriques. Enfin une série d'expériences a démontré l'adaptabilité au bruit des données et le bon fonctionnement de la méthode.

La description d'une généralisation concise de la définition du nombre de fausses alarmes a permis l'application d'*AC-RANSAC* :

- A l'estimation de nouveaux modèles paramétriques :
  Estimation de poses relatives (matrice essentielle),
  Estimation de poses monde (matrice de projection),
  Estimation de tenseurs tri-focaux.
- Sur des erreurs angulaires.
  Son application à l'estimation *a contrario* de la matrice essentielle entre images panoramiques sphérique a été démontrée expérimentalement.

Les travaux pour la généralisation de l'utilisation de l'*a contrario* RANSAC à des modèles d'erreur point-point et point-ligne ont été publiés dans une revue avec code source et démonstration en ligne (MOISAN et al. 2012). Le papier, une interface Internet d'expérimentation de l'algorithme ainsi qu'une implémentation libre d'*AC-RANSAC* sont disponibles. En octobre 2013, 2940 expérimentations en ligne de l'algorithme ont été réalisées par des utilisateurs anonymes.

Chapitre 5

Une chaîne de calibration séquentielle

Dans ce chapitre nous allons analyser le modèle de chaîne de calibration externe séquentielle le plus couramment utilisé pour estimer la structure à partir du mouvement d'un ensemble d'images. Cette chaîne de traitement repose sur l'utilisation répétée d'estimations robustes de modèles paramétriques afin d'étendre une reconstruction initiale. Ces estimations répétées nécessitent la définition de seuils qui décident de manière irrévocable les données validées ou rejetées au fur et à mesure de l'ajout d'images. Une question se pose alors : existe-t-il des seuils permettant l'obtention de la solution optimale ?

Pour répondre à cette question nous allons :

1. Étudier la chaîne classique de calibration séquentielle et identifier où l'utilisation de seuils intervient,
2. Adapter la chaîne pour utiliser l'estimateur robuste adaptatif *a contrario* du chapitre 4.
3. Évaluer la précision d'une chaîne de reconstruction 3D qui s'adapte aux données et où aucun seuil a priori n'est à fournir.

Sommaire

5.1 État de l'art

A partir de la connaissance de points homologues à travers une ensemble d'images il est possible de déterminer la structure de la scène à partir du mouvement de la caméra. Le mouvement de la caméra et la structure peuvent être calculés de manière séquentielle : *incremental SfM*. Nous supposons que les paramètres internes (la matrice de calibration **K**) de l'appareil photographique utilisé sont connus de manière approximative. Étant donné qu'aucune méthode ne permettait de calculer la visibilité d'un point 3D à travers une séquence d'image complète de manière fiable, les travaux se sont d'abord intéressés à des séquences d'images ordonnées. Puis avec les progrès réalisés sur la mise en correspondance de points saillants des solutions pour des jeux d'images non ordonnées ont été proposées.

Utilisation de séquence d'images ordonnées

La première chaîne générale a été proposée par (BEARDSLEY et al. 1996) (cf. figure 5.1). Les auteurs proposent d'établir une reconstruction initiale à partir de deux ou trois images puis d'étendre cette structure en ajoutant les images une par une par estimation de pose. A chaque ajout d'image la reconstruction est étendue en ajoutant de nouveaux point 3D à la structure par triangulation. Ces travaux généralisent la chaîne proposée par HARRIS et PIKE (1988).

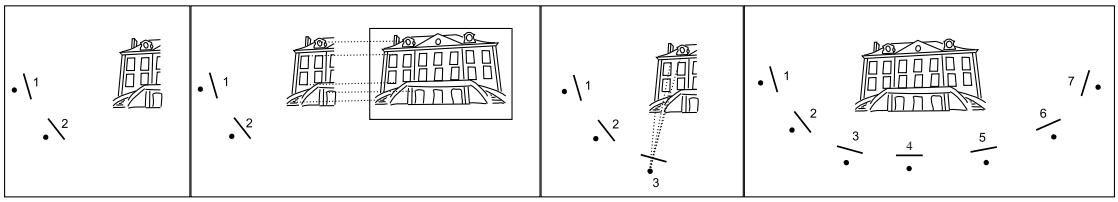


FIGURE 5.1 – Illustration de la reconstruction séquentielle. De gauche à droite : Les images 1 et 2 sont utilisées pour créer une reconstruction initiale. Les correspondances entre l'image 3 et la scène reconstruite sont listées. Ces correspondances 3D-2D permettent l'estimation de la pose et orientation de la caméra 3. La scène est complétée par ajout de nouveaux points 3D à la structure par triangulation et par itération du processus pour estimer la pose des images 4 à 7.

La reconstruction étant sensible aux bruits de mesure sur les positions des points saillants détectés et à l'accumulation d'erreurs inhérentes au processus séquentiel, le processus de reconstruction doit être stabilisé. HARRIS et PIKE (1988) utilisent un filtre de Kalman pour raffiner la structure 3D à chaque ajout d'image. BEARDSLEY et al. (1996) ajoutent des contraintes a priori sur les paramètres intrinsèques pour obtenir une reconstruction quasi-euclidienne (à un facteur d'échelle près de la réalité) et utilisent une étape de minimisation non linéaire pour optimiser les matrices de projection des caméras et réduire les erreurs résiduelles observées. BEARDSLEY et al. (1996) montrent que l'utilisation de tenseur tri-focaux permet d'obtenir une reconstruction initiale plus stable et un meilleur filtrage des mises en correspondance aberrante, comparé à l'utilisation de la géométrie essentielle seule. La procédure 5 résume l'algorithme.

Afin de faciliter le processus on suppose une cohérence temporelle. Cela présente un avantage pour la mise en correspondance de points saillants dans la séquence d'images ordonnées mais cela présente aussi un inconvénient majeur : si un nombre insuffisant de correspondances est détecté tout le processus de reconstruction s'arrête. On

Procédure 5 Structure from Motion séquentiel pour des images ordonnées

Entrée: Une collection d'images $\{I_i\}_i : i \in (1,...,n)$
Sortie: Une pose de caméra pour chaque image i et une structure 3D pour la scène observée
Calcul des points d'intérêts pour chaque image
Calcul des correspondances de points entre images consécutives
Reconstruction 3D initiale à partir des 2 ou 3 premières images
tant que il reste des images **faire**
Estimation de pose de l'image courante avec la reconstruction à partir des correspondances 2D-3D
Ajustement non linéaire de la géométrie et des poses de caméras
fin tant que

peut bien sûr étendre à deux, trois images la zone considérée comme contiguës entre les images pour la recherche de correspondances mais les coupures de reconstruction peuvent toujours apparaitre, c'est pourquoi il est intéressant de considérer le cas général : une collection d'images non ordonnées.

Utilisation d'un ensemble d'images non ordonné

L'utilisation de collection d'images non ordonnées pose avant tout un problème pour l'établissement des points saillants homologues entre images, la cohérence temporelle n'est plus supposée d'image à image. Des correspondances photométriques sont détectées par paires d'images (cf. section 3.9) puis filtrées pour ne conserver que les correspondances valides sous géométrie épipolaire (cf. chapitre 4). Partant d'un ensemble d'images non ordonnées on peut construire un graphe ayant pour sommets les images et pour arêtes les géométries épipolaires validées par estimation robuste. Un graphe de géométrie épipolaire $G_{\mathcal{E}}$ est ainsi construit, il décrit les relations de points saillants à travers la collection d'images (cf. figure 5.2). Ce graphe $G_{\mathcal{E}}$ est construit avec la procédure 6.

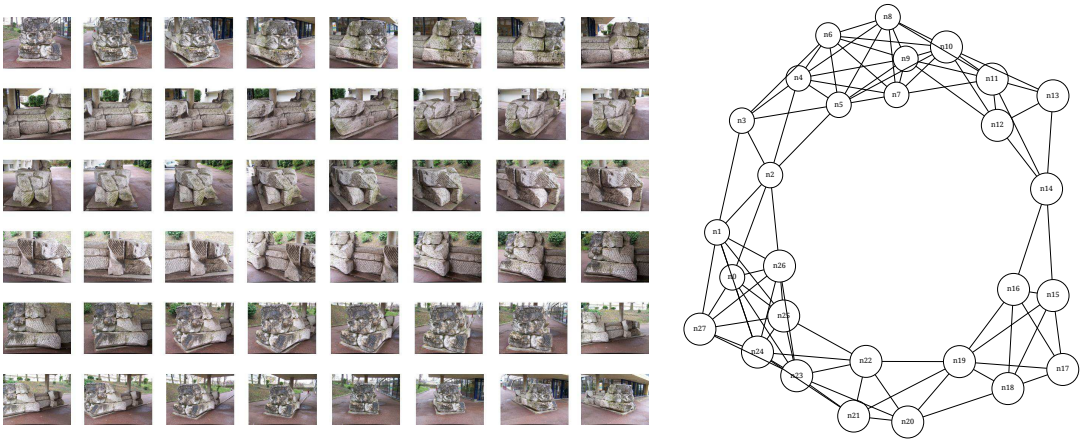


FIGURE 5.2 – Exemple de graphe épipolaire obtenu avec un ensemble d'images acquises autour d'un objet.

Procédure 6 Calcul des correspondances épipolaires géométriquement valides par paires d'images

Entrée: Une collection d'images $\{I_i\}_i : i \in (1,...,n)$
Sortie: Des correspondances par paires géométriquement valides : $G_{\mathcal{E}}$
$G_{\mathcal{E}} = \emptyset$
(1) **Calcul de la description locale des images :**
pour Toute image $i \in (1,...,n)$ **faire**
Détection et description des points saillants de l'image I_i
fin pour
(2) **Calcul des correspondances photométriques puis géométriques :**
pour Toutes paires d'images $(i,j) \in (1,...,n)$ **faire**
Calcul des correspondances photométriques entre I_i et I_j
$*$ Estimation robuste de la matrice fondamentale **F**
si Estimation réussie **alors**
Ajout de la paire (i,j) au graphe $G_{\mathcal{E}}$
fin si
fin pour

Une fois les points homologues identifiés, on recherche la structure de la scène et le mouvement des caméras. La construction et l'utilisation de ce graphe a été proposée par SCHAFFALITZKY et ZISSERMAN (2002), puis simplifié par BROWN et LOWE (2005b) en généralisant l'usage des points saillants de type SIFT (LOWE 1999) à la reconstruction 3D. Ces travaux ont permis de déterminer des méthodes de reconstruction séquentielle adaptées à des collections d'images non ordonnées. La reconstruction est réalisée en parcourant le graphe épipolaire $G_{\mathcal{E}}$ et non plus en considérant les images dans un ordre pré-déterminé. Les mêmes éléments de bases que ceux exposés par (BEARDSLEY et al. 1996) sont utilisés :

1. Une estimation robuste de la matrice essentielle pour établir une reconstruction initiale,
2. Des estimations de poses répétées et triangulations ajoutent du contenu à la scène en reconstruction. A chaque itération, l'image considérée est celle présentant le plus de correspondances avec la structure en cours de reconstruction (des correspondances 2D-3D sont identifiées grâce aux contraintes de visibilité du graphe épipolaire : les traces).
3. Des étapes d'ajustements de faisceaux pour réduire les erreurs d'accumulations : (*drift errors*). Ces étapes d'ajustements non linéaires sont globales : elles considèrent l'ensemble des caméras et les points mis en correspondance à la structure 3D reconstruite jusque là.

On note les évolutions suivantes :

Une généralisation de l'utilisation de l'ajustement de faisceaux.

Une minimisation de l'erreur résiduelle entre la projection de la structure et les points saillants est réalisée par BROWN et LOWE (2005b). LOURAKIS et ARGYROS (2004) proposent une implémentation parcimonieuse utilisant beaucoup moins de mémoire que la version dense.

La généralisation de l'utilisation des traces.

Les images étant non ordonnées, la visibilité d'un point 3D est estimée en reliant les correspondances établies par le graphe épipolaire, on obtient des traces.

Chaque trace représente le mouvement apparent d'un point 3D pour une série d'images, une information de visibilité (cf. section 3.10). Des images non contigües partagent ainsi des informations et facilitent la reconstruction. BROWN et LOWE (2005b) complètent les traces en parcourant le graphe au fur et à mesure des estimations de poses de caméras. SNAVELY et al. (2006) utilisent les traces en entrées de la chaîne, la visibilité des points 3D est ainsi directement disponible. L'utilisation des traces en entrée de la chaîne permet d'avoir des informations de visibilité plus denses que des informations considérées entre arêtes contiguës. Les informations de visibilité étant plus denses, l'étape d'ajustement de faisceaux est plus précise car plus de points sont considérés. Les résultats de l'estimation robuste de poses sont plus précis et robustes car plus d'*inliers* sont déterminés.

L'utilisation de données initiales approximatives pour les paramètres intrinsèques. SNAVELY et al. (2006) utilisent une matrice de calibration $\mathbf{K}$ initiale par image. La focale en pixels est calculée à partir d'une focale en *mm* extraite des données EXIF des images ainsi que de la taille de capteur de l'appareil photographique utilisé, supposée connue. Le point principal est supposé au centre de l'image. L'utilisation de ces données initiales permet l'estimation d'une structure euclidienne (à une échelle près de la scène réel) et rendent les estimations de poses et résultats des étapes d'ajustement de faisceaux plus stables.

La procédure de reconstruction séquentielle pour une collection d'images non ordonnées est résumé par la procédure 7 :

Procédure 7 *Structure from Motion* séquentiel pour un ensemble d'images non ordonné

Entrée: Paramètres intrinsèques $\mathbf{K}$, $G_{\mathcal{E}}$: correspondances épipolaires par paires d'images
Sortie: Structure 3D $\{X_j\}_j$, poses de caméra $\{\mathbf{P}_i\}_i$
Calcul des traces de points saillant tr
(1) **Reconstruction 3D initiale :**
Choix d'une arête e dans $G_{\mathcal{E}}$
∗ Estimation robuste de la matrice essentielle $\mathbf{E}$ pour e
$\{X_j\}_j$ = Triangulation des traces appartenant à e : $tr \cap e$
$\{\mathbf{P}_i\}_i = (P_0, P_1)$
Suppression de l'arête e
(2) **Ajout séquentiel d'images :**
tant que $G_{\mathcal{E}} \neq \emptyset$ **faire**
Choix d'une arête e dans $G_{\mathcal{E}}$ qui maximise la cardinalité de points 3D visibles entre la reconstruction et une nouvelle image : $e : \underset{e}{\operatorname{argmax}}(\#tr(e) \cap \{X_j\}_j)$
∗ Estimation robuste de la pose $\mathbf{P}_i$ de la caméra i
si Estimation robuste réussie **alors**
Ajout à $\{\mathbf{P}_i\}_i$ de P_i
Ajout à $\{X_j\}_j$ de nouveaux points 3D : triangulation des points 3D non reconstruit correspondant aux traces entre l'image I_i et les images déja positionnées en 3D
fin si
Suppression de l'arête e
(3) **Réduction des accumulations d'erreurs :**
Ajustement de faisceaux
fin tant que

Maintenant que la chaîne générique de reconstruction séquentielle est présentée nous allons analyser les points cruciaux qui garantissent le succès et la précision d'une reconstruction malgré la présence de données aberrantes.

5.1.1 Analyse du point clef des méthodes de reconstructions séquentielles

En représentant de manière graphique les procédures 6 et 7 (cf. figure 5.3) on observe que l'estimation robuste est un élément clef de la chaîne de reconstruction :

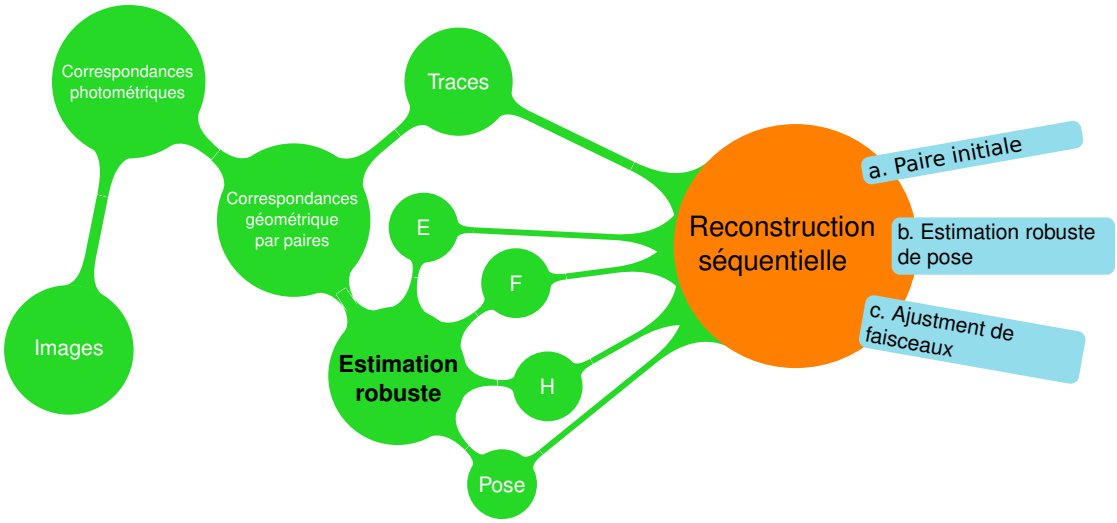


FIGURE 5.3 – Une représentation graphique de la reconstruction séquentielle. En bleu, la partie algorithmique puis en vert les outils et données d'entrées.

On observe que la brique d'estimation robuste est la colonne vertébrale du processus. C'est elle qui garantit la qualité des modèles paramétriques déterminés ainsi que les données validées et rejetées par le processus. On note que les estimations robustes de la chaîne portent sur les modèles paramétriques suivants :

- **La géométrie fondamentale F** pour la mise en correspondance de points saillants,
- **La géométrie homographique H** pour le choix d'une paire initiale ayant de la parallaxe,
- **La géométrie essentielle E** pour établir la reconstruction 3D initiale,
- **La matrice de projection P** pour ajouter une image à la scène.

Ces estimations de modèles paramétriques robustes sont nécessaires car les correspondances de points homologues sont perturbées par du bruit de mesure et corrompues par de fausses correspondances. Les méthodes de l'état de l'art utilisant le plus souvent RANSAC pour déterminer ses modèles paramétriques, nous proposons d'évaluer l'impact du choix d'un seuil δ sur la qualité d'estimation des poses de caméras. Rappel : ce seuil δ est une borne haute de la précision acceptée par RANSAC, c'est ce qui décide si une mesure est conservée ou rejetée (les correspondances *inliers*).

5.2 Impact de l'estimation robuste contrainte sur une chaîne de calibration séquentielle

Nous proposons ici d'évaluer l'efficacité relative d'une chaîne de calibration séquentielle en fonction du seuil δ choisi pour les phases d'estimation robustes. La solution open-source «Bundler» (SNAVELY et al. 2006) est utilisée comme boite noire de reconstruction séquentielle. L'implémentation utilise RANSAC et différentes valeurs de seuil δ fixées par défaut en fonction du modèle à estimer :

- 9 pixels pour l'estimation des matrices fondamentales,
- 6 pixels pour l'estimation des matrices homographiques (choix de la paire initiale),
- 4 pixels pour l'estimation des matrices de projections.

Ces choix heuristiques peuvent mener à de bons résultats mais ne peuvent s'adapter à toutes les situations, à cause de l'efficacité relative en fonction du seuil δ choisi. Nous proposons à travers une expérience de montrer qu'utiliser différentes valeurs de seuil δ produit des résultats complètement différents et que le choix d'un unique seuil global n'est pas optimal.

Le protocole de l'expérience est le suivant. Nous comparons la précision de la reconstruction obtenue en fonction de la valeur d'un seuil fixé à différentes valeurs δ, pour l'estimation des matrices fondamentales. Cette précision de reconstruction est mesurée sur des jeux de données d'images accompagnés d'une vérité terrain (cf. sous-section 5.4). La distance résiduelle entre les caméras vérité terrain et les caméras calculées est évaluée à la suite de l'estimation d'une transformation rigide de degré 7 (échelle, rotation, translation) entre les deux repères (cf. HARALICK et SHAPIRO (1992)). Les caméras calculées sont ainsi projetées dans le repère vérité terrain métrique et une erreur moyenne résiduelle de positionnement peut être évaluée. Cette transformation rigide préserve les angles et ratios de distance, donc elle ne déforme pas la structure. Les résultats suivants sont observés en faisant varier le seuil d'estimation des matrices fondamentales $\delta \in 1, 3, 6, 9, 12$ pixels :

Scène	δ pour **F** (pixels)				
	1	3	6	9	12
EntryP10	44.7	60.7	54.8	55.1	**30.2**
FountainP11	4.11	**4.02**	4.50	7.03	8.68
HerzJesusP8	**9.21**	9.62	10.0	16.4	9.97
HerzJesusP25	26.9	31.9	**16.3**	21.5	29.4
CastleP19	6817	**312**	398	344	4475
CastleP30	280	225	**125**	300	150

TABLE 5.1 – Erreurs moyennes en millimètres observées par rapport à la vérité terrain-pour différentes valeurs de seuil δ pour l'estimation des matrices fondamentales par Bundler. Les erreurs les plus faibles sont en gras.

On observe qu'il n'existe pas un seuil fixe idéal permettant d'obtenir à chaque fois la meilleure solution. Le meilleur seuil dépend du jeu de données considéré. Plus généralement fixer un seuil uniforme qui tranche entre les données validées ou rejetées

est sous-optimal au sein même d'un ensemble d'images. Note : nous avons ici fait varier uniquement le seuil d'estimation des matrices fondamentales, mais des résultats similaires sont obtenus si l'on fait varier les autres seuils.

5.3 Une chaîne de calibration séquentielle *a contrario*

Plutôt que d'utiliser des seuils fixes pour réaliser des estimations robustes avec des méthodes de type RANSAC ou MAX-CONSENSUS ordinaires, nous proposons d'utiliser la méthodologie *a contrario*.

Nous avons vu que les méthodes séquentielles requièrent des estimations robustes répétées et que les estimations de modèles paramétriques sont évaluées sur des données variées (des correspondances 2D-2D, 2D-3D). Un seuil fixe n'est pas idéal car il est insensible au bruit particulier de chaque configuration rencontrée. En utilisant l'estimation de modèle *a contrario*, nous réalisons chaque estimation de modèle paramétrique de manière indépendante et adaptative. Un seuil δ sera déterminé automatiquement par des méthodes statistiques pour chaque situation rencontrée, évitant les effets d'efficacité relative liés à l'usage d'un seuil fixe.

En spécialisant la définition du NFA générique de la section 4.3, nous proposons d'adapter la chaîne décrite avec les procédures 6 et 7. Nous remplaçons l'utilisation de RANSAC par *AC-RANSAC* sur les lignes marquées par une étoile. Nous obtenons une chaîne de reconstruction séquentielle *a contrario* que nous désignerons par l'acronyme *AC-SfM*. Elle réalise la reconstruction 3D d'une scène en étant adaptative aux données. Les opérations suivantes sont réalisées *a contrario* :

La mise en correspondance de points saillants : La vérification de l'existence d'un groupe de correspondances valides par géométrique épipolaire est réalisée *a contrario* pour chaque paire d'images.

L'estimation de la reconstruction initiale : Le groupe de correspondances utilisé pour réaliser la reconstruction initiale est déterminé *a contrario*.

L'ajout de caméra : Chaque estimation de matrice de projection caméra est réalisée *a contrario*. La précision identifiée *a contrario* est utilisée comme valeur de confiance pour les nouvelles traces à trianguler. La reconstruction est ainsi étendue en conservant une erreur résiduelle n'excédant pas la confiance de l'image que l'on vient d'ajouter.

Contrairement aux méthodes habituelles qui utilisent un seuil global fixe lié au type de modèle paramétrique :

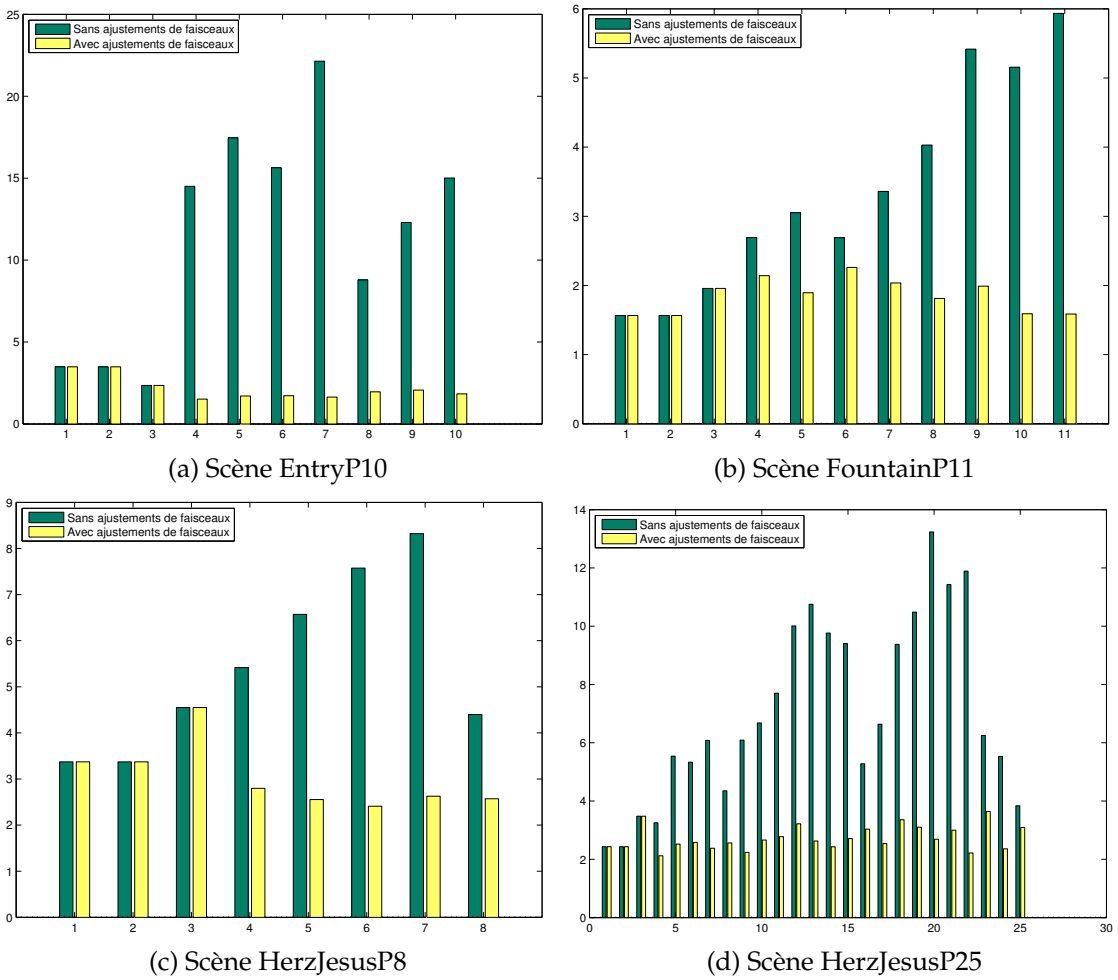
- nous libérerons l'utilisateur du choix de seuils heuristiques,
- nous calculons des estimations robustes adaptatives aux bruits des données.

Pour rappel : Bundler utilise les seuils suivant : $\delta\mathbf{F} = 9$ pixels, $\delta\mathbf{H} = 6$ pixels et $\delta\mathbf{P} = 4$ pixels.

Nous allons maintenant vérifier que notre chaîne *AC-SfM* est bien réactive aux bruits des données. Puis nous évaluerons, sur plusieurs jeux de données, en terme quantitatif la qualité d'estimation des poses de caméra.

5.3.1 Une chaîne adaptative aux bruits des données

Afin de vérifier l'adaptabilité de notre chaîne *AC-SfM* aux bruits des données, nous proposons l'expérience suivante. Par nature, l'utilisation d'un processus séquentiel accumule des erreurs et des étapes d'ajustement de faisceaux sont employées pour réduire ce biais. Nous proposons d'utiliser la chaîne de reconstruction *AC-SfM* avec et sans ajustement de faisceaux et d'analyser l'impact de seuils δ déterminés *a contrario*. Nous travaillons pour cela à graphe épipolaire identique : les correspondances initiales utilisées sont les mêmes. Sans ajustements de faisceaux on s'attend à ce que le cumul d'erreur ajoute du bruit à la scène et à ce que notre chaîne *AC-SfM* s'adapte aux données en identifiant des valeurs δ croissantes. Au contraire si les étapes d'ajustement de faisceaux sont réalisées alors le seul *a contrario* devrait varier avec beaucoup moins d'amplitude. Les résultats de l'expérience sont illustrés figure 5.4 et figure 5.5.



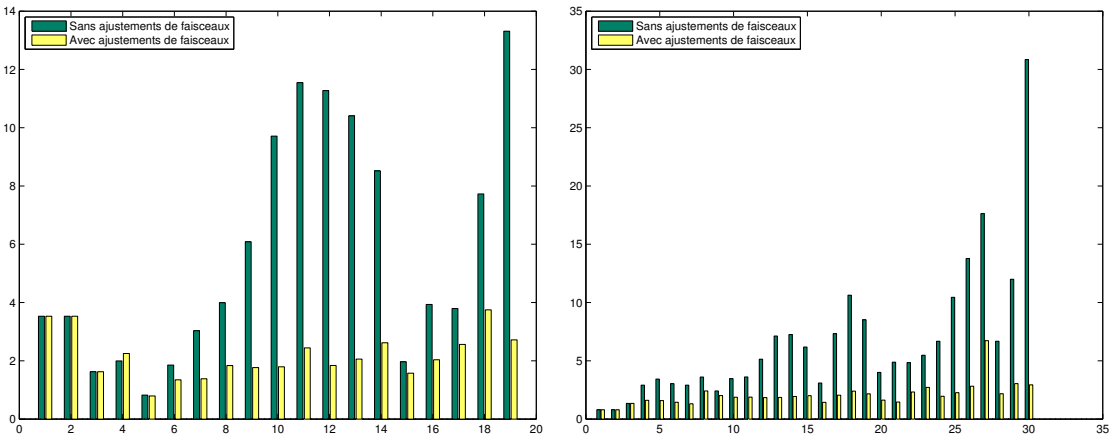
(a) Scène EntryP10

(b) Scène FountainP11

(c) Scène HerzJesusP8

(d) Scène HerzJesusP25

FIGURE 5.4 – Test de la chaîne *AC-SfM* avec et sans ajustements de faisceaux. Les seuils δ (pixels) déterminés par estimation robuste *a contrario* sont affichés dans l'ordre d'ajout des caméras. On retrouve ainsi de gauche à droite : deux seuils identiques pour la paire initiale, seuil issu de l'estimation robuste *a contrario* de la matrice essentielle, puis les seuils déterminés pour les estimations de poses successives.



(a) Scène CastleP19 (b) Scène CastleP30

FIGURE 5.5 – Test de la chaîne *AC-SfM* avec et sans ajustements de faisceaux. Les seuils δ (pixels) déterminés par estimation robuste *a contrario* sont affichés dans l'ordre d'ajout des caméras. On retrouve ainsi de gauche à droite : deux seuils identiques pour la paire initiale, seuil issu de l'estimation robuste *a contrario* de la matrice essentielle, puis les seuils déterminés pour les estimations de poses successives.

Les graphiques permettent de constater que notre chaîne est bien réactive au bruit. Á partir de la troisième caméra ajoutée on distingue des changements. Les deux premières caméras portent le même seuil car elles sont liées à l'estimation de la matrice essentielle pour réaliser la reconstruction initiale. On note que sans ajustements de faisceaux, les seuils calculés *a contrario* grandissent avec les erreurs accumulées. Avec ajustement de faisceaux ils sont réguliers, ils varient avec une faible amplitude, mais ne sont plus globalement croissants.

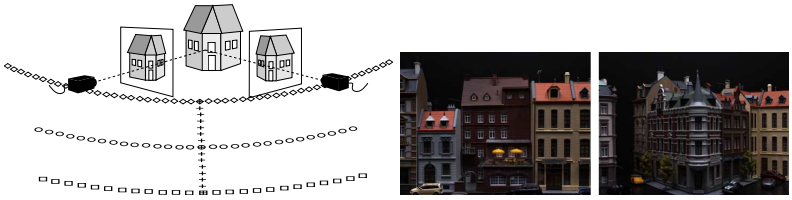
5.4 Résultats et évaluations

Nous avons vérifié que la chaîne *AC-SfM* est bien adaptative à chaque situation rencontrée. Pour évaluer notre approche de manière quantitative, nous comparons la précision de la reconstruction obtenue à deux autres méthodes séquentielles, **Bundler** et **VisualSfM**. Bundler est une solution open-source en relation avec le papier de (SNAVELY et al. 2006) qui suit le schéma général de calibration séquentielle à seuils fixes que nous avons présenté. VisualSfM (WU 2013) est une solution logicielle utilisant des seuils et des heuristiques différentes pour réaliser des calculs plus rapides (le calcul des points saillants SIFT (WU 2007) et les ajustements de faisceaux (WU et al. 2011a) sont réalisés en utilisant la carte graphique).

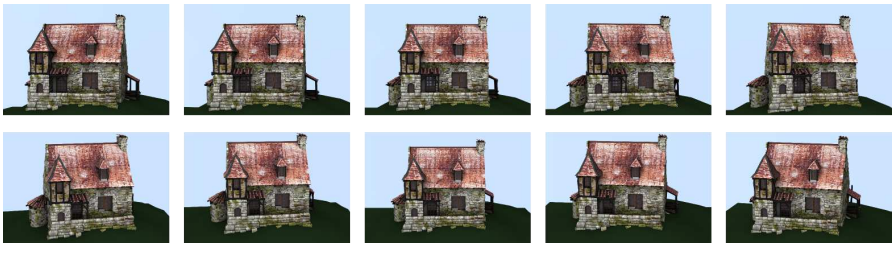
Pour évaluer la qualité de l'estimation de pose des caméras, nous avons réalisé des expériences sur des jeux de données comportant une vérité terrain très précise pour les paramètres intrinsèques et extrinsèques. Le positionnement des caméras est connu dans un espace métrique. Trois jeux de données utilisant des images sans distorsion optique sont considérés :

STRECHA et al. (2008) : 6 jeux de données de 8 à 30 images. Pour chaque scène, des marqueurs ont été posés et un nuage laser acquis. La position des caméras a été estimée par ajustement de faisceaux entre la position des marqueurs et leur position sur le terrain déterminée par scan 3D laser. Les noms des jeux de données sont : entryP10, FountainP11, HerzJesusP8, HerzJesusP25, CastleP19 et CastleP30.

AANÆS et al. (2012) : 60 scènes photographiées avec un robot positionnant un appareil photographique numérique à 119 positions calibrées à l'avance. La précision de positionnement est estimée être précise avec un écart type de 0.1 millimètre. Seules deux scènes représentant des maquettes architecturales sont utilisées (jeux de données DTU-Set001 et DTU-Set004).



MOULON et al. (2013a) : 1 jeu de données généré de manière synthétique sous le logiciel Blender. Une scène 3D est créée. Puis des images de la scène sont obtenues par lancer de rayons à différentes positions connues dans l'espace. Ce dataset est appelé HouseDataset.

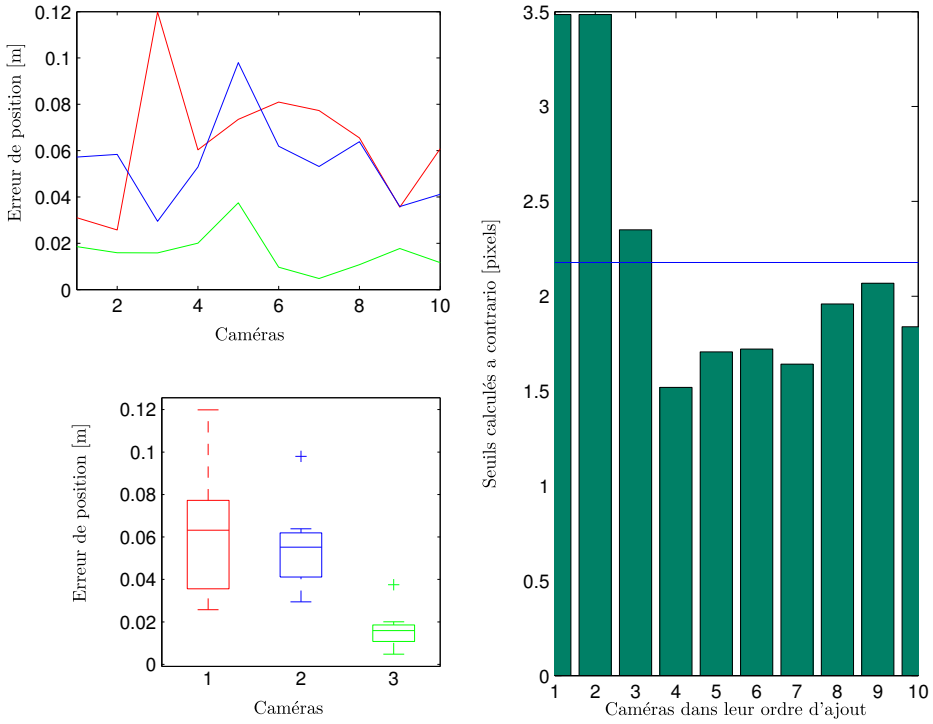


L'évaluation est réalisée en considérant les mêmes données d'entrées : des points SIFT validés par critère NN-DR avec un ratio de 0.6, et la même paire initiale. La matrice de paramètres intrinsèques est fournie pour toutes les méthodes évaluées. La qualité d'estimation des paramètres extrinsèques est calculée par une transformation rigide de degré 7 (échelle, rotation, translation) entre la vérité terrain et les positions de caméra estimées (cf. section 5.2). Les résultats de cette évaluation sont présentés dans le tableau 5.2.

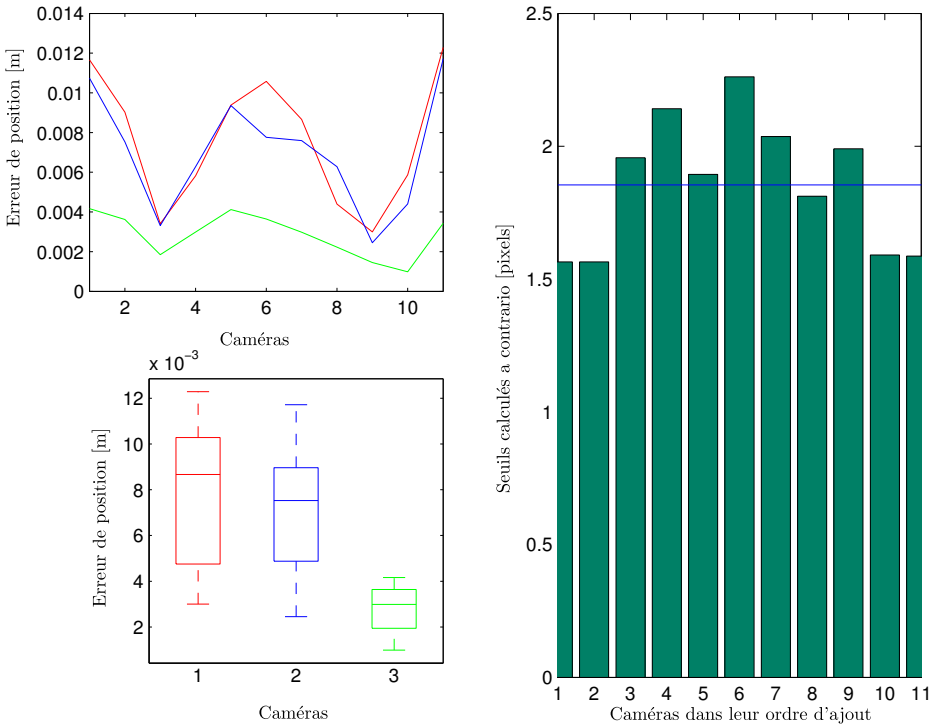
Scène		*AC-SfM*	Bundler	VisualSfM
entryP10	erreur	**16.2**	55.1	63.0
	rang	**1**	2	3
FountainP11	erreur	**2.86**	7.03	7.64
	rang	**1**	2	3
HerzJesusP8	erreur	**6.71**	16.4	19.3
	rang	**1**	2	3
HerzJesusP25	erreur	**8.85**	21.5	22.4
	rang	**1**	2	3
CastleP19	erreur	**223**	344	258
	rang	**1**	3	2
CastleP30	erreur	**69.1**	300	522
	rang	**1**	2	3
DTU-Set001	erreur	**0.71**	0.72	1.22
	rang	**1**	2	3
DTU-Set004	erreur	0.70	0.49	**0.46**
	rang	3	2	**1**
HouseDataset	erreur	**11.6**	11.7	211
	rang	**1**	2	3

TABLE 5.2 – Évaluation des erreurs de localisation moyenne des caméras en millimètres par rapport à la vérité terrain pour trois méthodes séquentielles sur différents jeux de données. Notre solution *AC-SfM*, Bundler (SNAVELY et al. 2006) et VisualSfM (WU 2013).

Constat. On observe que la solution *AC-SfM* donne dans 90% des cas la meilleure solution et qu'en moyenne la solution identifiée est 2.1 fois plus précise que Bundler et 4.6 fois plus précise que VisualSfM. Les expériences confirment le fait que l'utilisation de seuils variables adaptables pour chaque situation permet de sélectionner de meilleures correspondances et donc d'obtenir une reconstruction 3D plus fiable. Les résultats détaillés par scène sont décrits dans sur les figures 5.6, 5.7, 5.8, 5.9 et 5.10. Nous comparons les qualités d'estimation de pose par caméra sous forme de courbe et boîte à moustaches pour les méthodes *AC-SfM* (vert), Bundler (bleu) et VisualSfM (rouge). Les seuils *a contrario* déterminés pendant la reconstruction sont affichés sur les parties droites dans l'ordre d'ajout des caméras ainsi que la valeur moyenne des seuils en bleu.

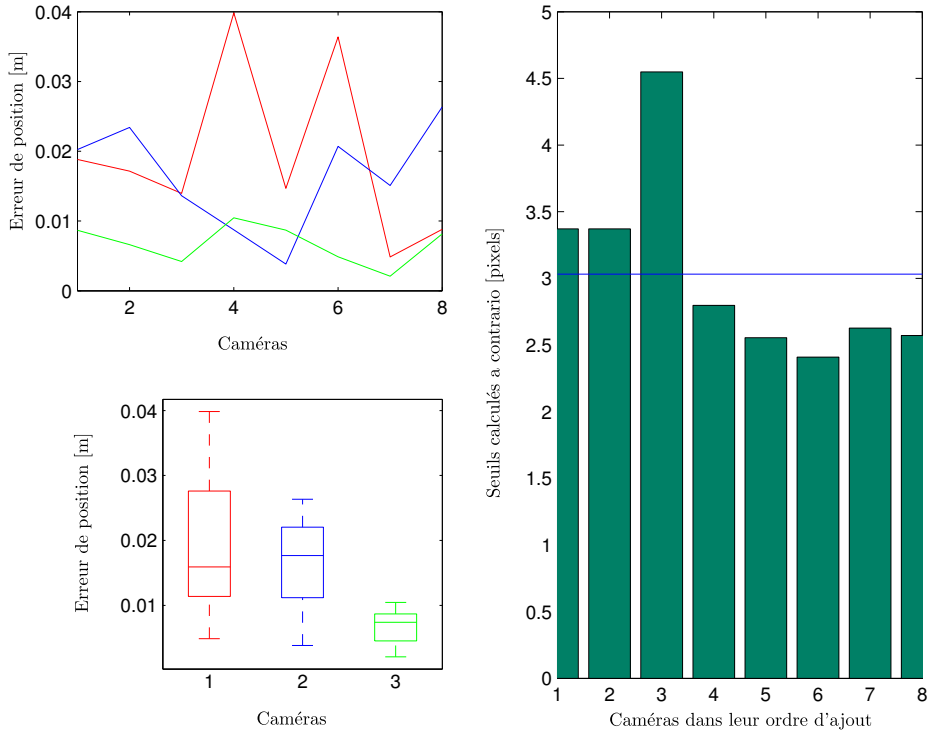


(a) Scène EntryP10

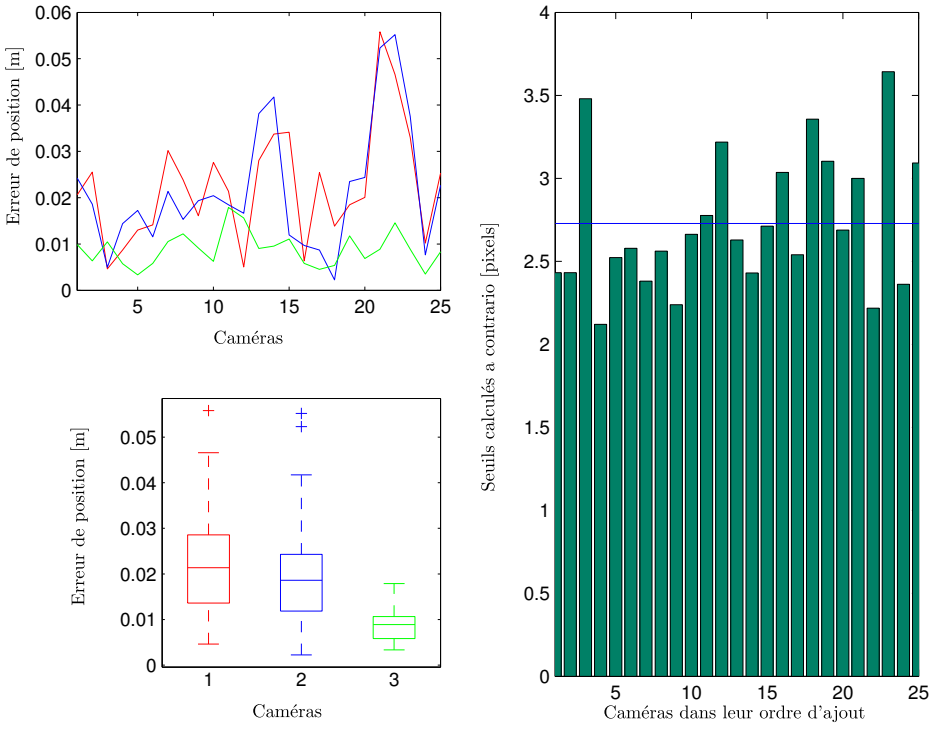


(b) Scène FountainP11

FIGURE 5.6 – A gauche : les qualités d'estimation de pose par caméra dans leur ordre initial de numérotation sous forme de courbe et boîte à moustaches. Légende : En vert *AC-SfM*, en bleu Bundler et en rouge VisualSfM. A droite : les seuils déterminés *a contrario* suivant l'ordre d'ajout des caméras, en bleu leur moyenne.

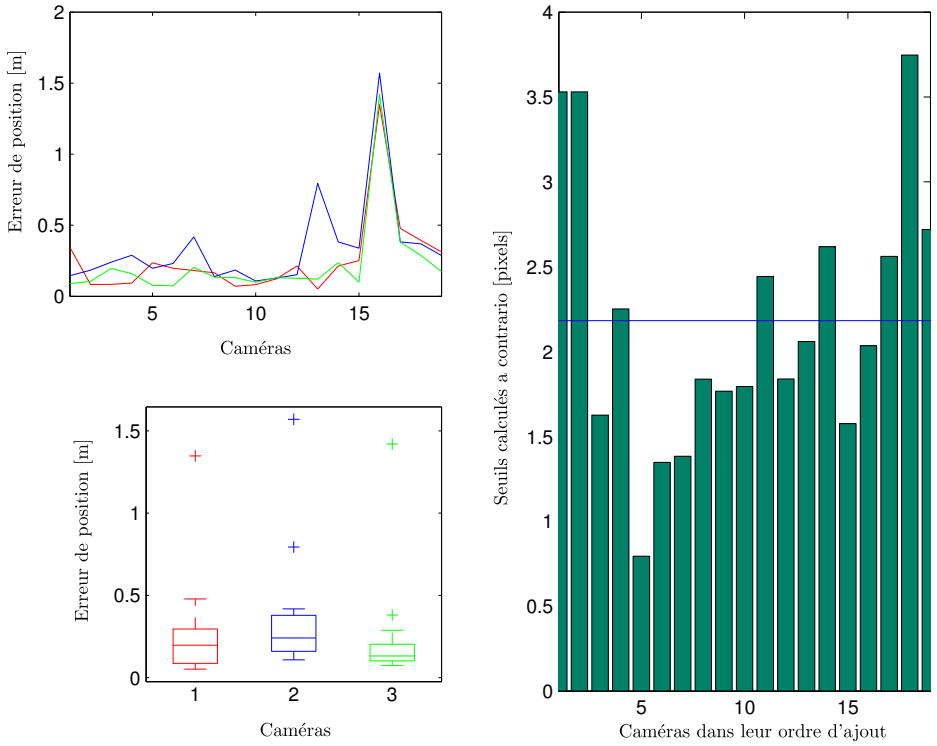


(a) Scène HerzJesusP8

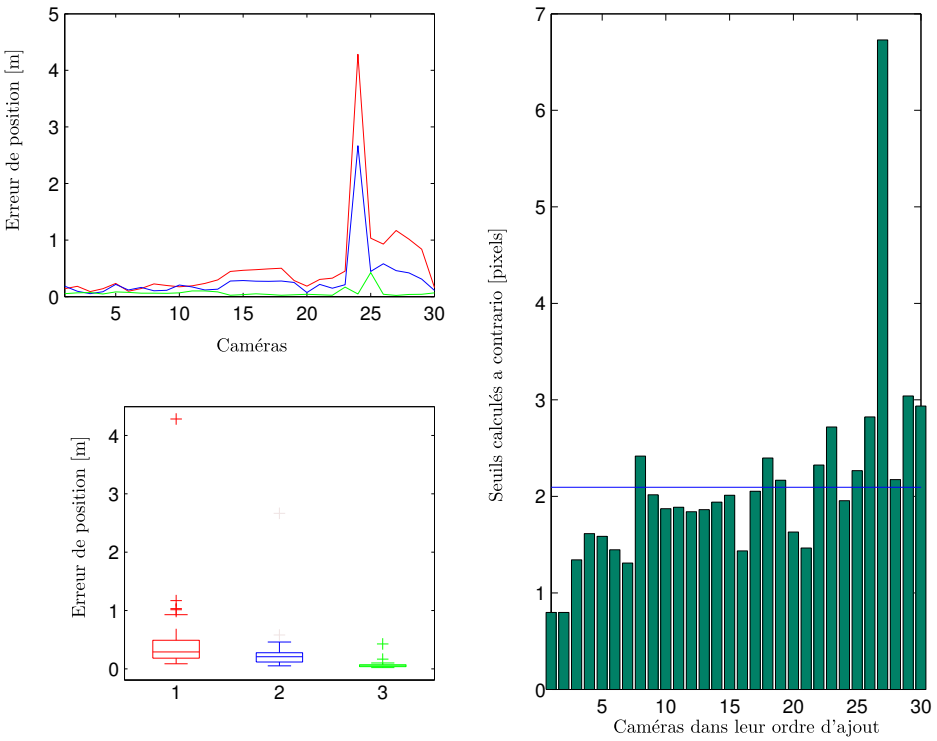


(b) Scène HerzJesusP25

FIGURE 5.7 – A gauche : les qualités d'estimation de pose par caméra dans leur ordre initial de numérotation sous forme de courbe et boîte à moustaches. Légende : En vert *AC-SfM*, en bleu Bundler et en rouge VisualSfM. A droite : les seuils déterminés *a contrario* suivant l'ordre d'ajout des caméras, en bleu leur moyenne.

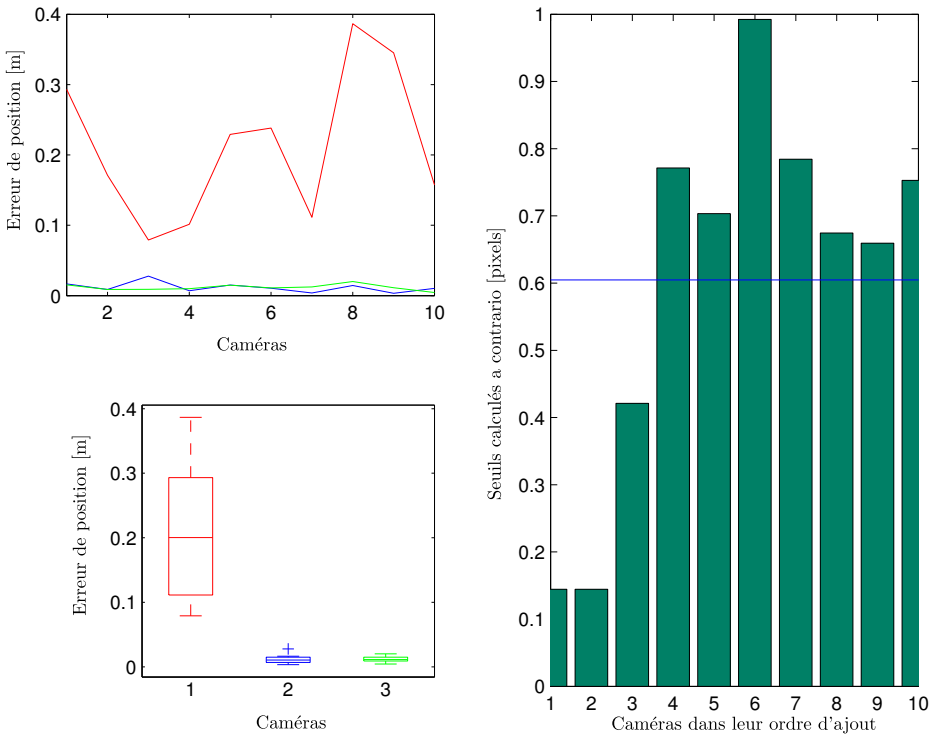


(a) Scène CastleP19



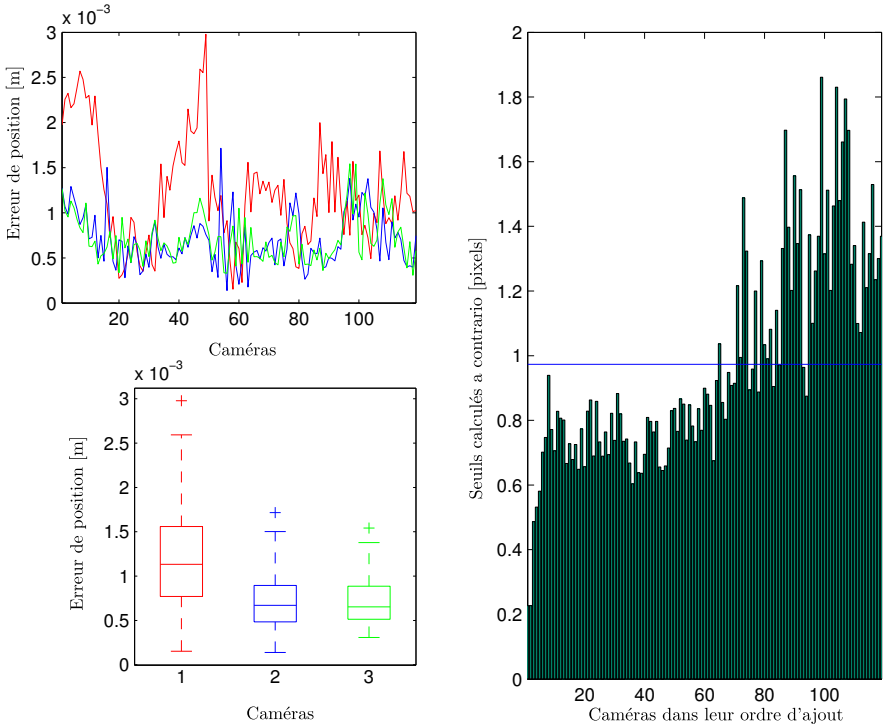
(b) Scène CastleP30

FIGURE 5.8 – A gauche : les qualités d'estimation de pose par caméra dans leur ordre initial de numérotation sous forme de courbe et boîte à moustaches. Légende : En vert *AC-SfM*, en bleu Bundler et en rouge VisualSfM. A droite : les seuils déterminés *a contrario* suivant l'ordre d'ajout des caméras, en bleu leur moyenne.

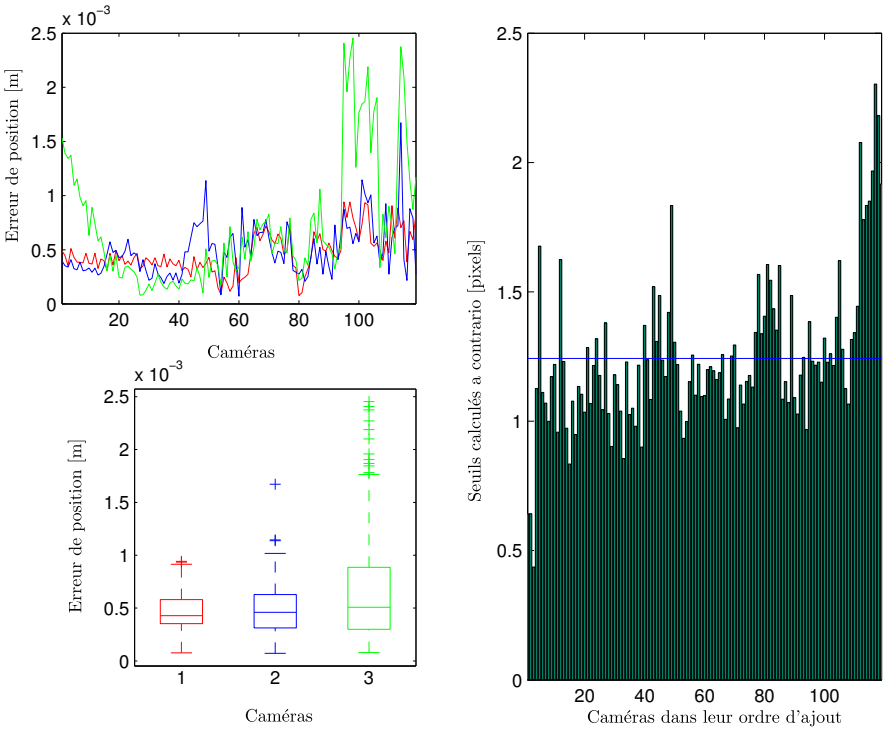


(a) Scène HouseDataset

FIGURE 5.9 – A gauche : les qualités d'estimation de pose par caméra dans leur ordre initial de numérotation sous forme de courbe et boîte à moustaches. Légende : En vert *AC-SfM*, en bleu Bundler et en rouge VisualSfM. A droite : les seuils déterminés *a contrario* suivant l'ordre d'ajout des caméras, en bleu leur moyenne.



(a) Scène Set001-ready



(b) Scène Set004-ready

FIGURE 5.10 – A gauche : les qualités d'estimation de pose par caméra dans leur ordre initial de numérotation sous forme de courbe et boîte à moustaches. Légende : En vert *AC-SfM*, en bleu Bundler et en rouge VisualSfM. A droite : les seuils déterminés *a contrario* suivant l'ordre d'ajout des caméras, en bleu leur moyenne.

Les résultats et les typologies d'acquisition et de scènes permettent certaines conclusions intéressantes :

Impact de la localisation des points saillants

On note que la scène synthétique, figure 5.9, est la scène présentant le seuil *a contrario* moyen le plus faible (0.6 pixels). Cela s'explique par le fait qu'avec des images de synthèse, présentant moins de bruit que des images naturelles, les appariements de points homologues sont réalisés avec une meilleure précision.

L'utilisation d'images à fort recouvrement améliore la reconstruction

Lorsque les images présentent de forts recouvrements, les traces de point saillants permettent d'avoir des erreurs résiduelles d'un même point 3D sur de nombreuses images (DTU-Set001, DTU-Set004). Les ajustements de faisceaux convergent alors vers de meilleures solutions. Des précisions millimétriques sont alors atteignables.

Trajectoire d'acquisition

Des prises de vues convergentes vers l'objet que l'on souhaite reconstruire donnent les meilleures reconstructions (FountainP11, HerzJesusP8, HerzJesusP25). Contrairement à des trajectoires rectilignes et fronto-parallèle à l'objet (EntryP10).

Cycle de grande taille

Lorsque l'utilisateur tourne complètement autour d'un objet ou à l'intérieur d'un bâtiment l'image de début est liée à l'image de fin (ex. cour de château : CastleP19 et CastleP30). Avec les risques de dérive dûs à l'accumulation d'erreurs, (*drift error*), on risque de pas relier les points saillants représentant des points homologues associés à des mêmes points 3D désignés en début et fin de la séquence d'images mais d'en «halluciner» de nouveaux. Lorsque cette situation arrive, on parle de non fermeture de boucle : une ou plusieurs contraintes de cyclicité du graph épipolaire ne sont pas respectées. Lorsque les boucles ne sont pas respectées les erreurs ne sont pas équi-réparties sur les arêtes des cycles. Ainsi on observe alors des reconstructions plus ou moins bonnes par partie. Que ce soit avec des cycles :

- de large amplitude. Cas d'images acquises en réalisant une rotation le long des murs de la cour d'un bâtiment (cf. figure 5.8),
- nombreux : Lorsque les images observent toutes le même objet (cf. figure 5.10).

5.5 Contributions de ce chapitre

Dans ce chapitre nous avons présenté la chaîne classique utilisée pour réaliser une calibration séquentielle. En modifiant cette chaîne pour utiliser la méthodologie *a contrario* afin de réaliser les différentes estimations robustes sans utiliser de seuils a priori nous avons réduit le nombre de paramètres à utiliser et amélioré la précision des estimations de pose et reconstructions. Les contributions de ce chapitre sont :

- Nous avons étudié expérimentalement l'impact de l'efficacité relative des étapes d'estimations robuste en fonction d'un seuil fixe choisi de manière heuristique. Aucun choix de seuil ne permet d'obtenir la meilleure reconstruction pour l'ensemble des jeux de données évalués.
- Nous avons présenté une version *a contrario* d'une chaîne de calibration séquentielle : *AC-SfM*. Cette chaîne supprime le choix par l'utilisateur d'un seuil arbitraire et s'adapte de manière dynamique et automatique aux données. Cette réaction dynamique au bruit a été vérifiée expérimentalement et nous avons démontré que cela permettait d'estimer plus précisément la position des caméras qu'avec les deux autres solutions testées.
- Nous avons montré que les seuils identifiés par *AC-RANSAC* sont différents pour chaque estimation robuste, ce qui montre que chaque estimation est bien indépendante. La réaction dynamique aux données permet d'éviter les situations de sous- ou sur-évaluation et donc de mieux selectionner les données et les modèles paramétriques.

Ces travaux ont été publiés à la conférence ACCV (MOULON et al. 2013a) et une implémentation libre d'*AC-SfM* est disponible au sein de la librairie open-source openMVG (*open Multiple View Geometry*)(MOULON et al. 2013d).

5.6 Les problématiques posées par les méthodes de calibrations séquentielles

Les méthodes de calibration séquentielle ne garantissent en aucun cas de converger vers la solution optimale. Si l'on résume, les problémes rencontrés sont les suivants :

- Le choix de la paire d'images initiale,
- L'ordre d'ajout des images,
- Le risque de dérive par accumulation d'erreurs,
- Le coût des opérations d'ajustements de faisceaux (minimisation non linéaire).

Pour chacune des ces problématiques nous avançons quelques arguments :

Le choix de la paire d'images initiale

La paire d'images initiale a un impact déterminant sur la qualité de la reconstruction finale. De cette paire une reconstruction initiale est calculée puis agrandie. Comme avancé par (HARTLEY et ZISSERMAN 2000) avec la proposition suivante, le choix de cette paire initiale n'est pas évident.

Proposition 6. *"There are several strategies that may be used to obtain the initial reconstruction, though this area is still to some extent a black art."*

Ainsi une reconstruction finale légèrement différente sera observée pour diverses paires initiales choisies.

L'ordre d'ajout des images

Le processus étant séquentiel, ajouter les images à la reconstruction dans un ordre variable va ajouter un bruit variable à la scène et donc donner lieu à des reconstructions légèrement différentes.

Le risque de dérive par accumulation d'erreurs

Par nature un processus séquentiel accumule des erreurs : (*drift error*). Des dérives sont ainsi observées sur des scènes présentant de grands déplacements le long de la séquence d'images. Le risque inhérent est que les cycles de départ du graphe épipolaire ne puissent pas être respectés. On observe ainsi souvent des boucles non fermées (*failed loop closure*).

Le coût des opérations d'ajustement non linéaire

Chaque minimisation non linéaire est une opération coûteuse. Étant lancé à chaque ajout d'image avec de plus en plus de données à considérer, la taille du problème à résoudre est croissante. SNAVELY et al. (2006) proposent de ne pas considérer l'étape d'ajustement à chaque ajout d'image mais ajoute des images par groupes. Un nouveau seuil apparaît alors pour déterminer combien d'images sont ajoutées à chaque itération. WU (2013) proposent de réaliser d'alterner des ajustements de faisceaux sur une sous-partie des données ou sur la totalité des données. Deux nouveaux seuils apparaissent pour déterminer quand lancer ces adjustements de faisceaux complets ou partiels.

Pour certains de ces problèmes, des améliorations ont été proposées :

Le choix de la paire d'images initiale

NOZAWA et al. (2013) démontrent qu'une série d'heuristiques permet de déterminer si une géométrie essentielle est propre à mener à une reconstruction stable ou non. Cela aide à choisir une paire initiale fiable pour mener à bien une reconstruction.

Le risque de dérive par accumulation d'erreurs

Des méthodes de fermetures de boucles peuvent être réalisées a postériori sur la reconstruction (*loop closure*). Ces méthodes sont inspirées du milieu de l'odométrie visuelle (cf. GUILBERT et al. (2004) ; KLOPSCHITZ et al. (2008)).

Le coût des opérations d'ajustement non linéaire

Des implémentations efficaces du problème existent, permettant de résoudre le problème à l'aide de matrices creuses LOURAKIS et ARGYROS (2004). Des approches utilisant des processus parallèles ont aussi été proposées WU et al. (2011a). Des méthodes pré-conditionnent le problème (AGARWAL et MIERLE 2012) afin d'obtenir une solution plus rapidement. RODRIGUEZ et al. (2011) proposent de ne pas considérer la structure dans le problème de minimisation, *structure-less BA*, mais des trajectoires rectilignes et des rotations pures ne sont pas traitées. INDELMAN et al. (2012) proposent de réaliser des ajustements de faisceaux sur une sous-partie du problème.

Concernant le passage à l'échelle, en plus des améliorations des temps de calcul des méthodes d'ajustement de faisceaux, HAVLENA et al. (2010) et SNAVELY et al. (2008) proposent de calculer un ensemble d'images connecté représentatif de la scène permettant d'estimer une structure stable : un arbre représentant le mieux le graphe de connexion épipolaire. Une reconstruction de la scène pour cet arbre est obtenue, puis les images restantes sont ajoutées par estimation de pose. Une reconstruction est ainsi produite plus rapidement que si l'ensemble des images avait été considérés.

Bien que le passage à l'échelle ait été amélioré (Wu 2013), seules des astuces sont utilisées pour tenter de conserver des boucles fermées et faire que la reconstruction soit correcte. Il subsiste un certain nombre de seuils pour paramétrer les estimations robustes et réduire le nombre d'étapes d'ajustement global et partiel. L'usage de ces seuils et heuristiques n'ajoutent aucune garantie de converger vers la solution optimale.

Afin d'obtenir des reconstructions de meilleures qualités nous proposons désormais d'analyser les méthodes de calibration externe globale, qui en considérant le réseau d'images dans sa globalité permet de ne pas souffrir d'effet de dérive et de ne plus imposer le choix problématique de la paire initiale.

Chapitre 6

Une chaîne de calibration globale

Nous avons vu que les méthodes de reconstruction séquentielles sont simples mais elles présentent plusieurs défauts : la qualité de reconstruction est dépendante de la reconstruction initiale et de l'ordre d'ajout des images ; et les erreurs s'accumulent par nature à cause du processus séquentiel. Malgré l'utilisation de multiples ajustements de faisceaux, la solution optimale n'est pas trouvée. Le fait est qu'en considérant le parcours du graphe de connexité épipolaire sous forme d'arbre, une solution bonne locale peut être calculée. Des méthodes hiérarchiques permettent d'obtenir de meilleures solutions grâce à une couverture plus large du graphe. Mais, les effets d'accumulation d'erreurs persistent car la solution finale est dépendante des reconstructions initiales qui sont assemblées.

Pour obtenir une meilleure reconstruction il faut considérer le problème dans sa globalité. C'est à dire forcer le respect des contraintes de cycles, ce qui à pour effet de répartir les erreurs d'estimation sur l'ensemble du réseau de caméras.

Plutôt que de considérer globalement l'ensemble des points homologues dans une collection d'images, dont le traitement serait trop coûteux, nous montrons que l'estimation globale de la structure à partir du mouvement peut être traitée globalement sous forme d'un problème de fusion de mouvements relatifs. Cet élément clef nous permet un passage à l'échelle plus aisé et l'estimation d'une solution plus précise que les méthodes concurrentes.

Sommaire

6.1 État de l'art

Les chaînes globales traitent le problème d'optimisation de la structure à partir du mouvement en considérant l'ensemble du réseau de caméras. Rappelons que : "Ce réseau est représenté par un graphe où les nœuds représentent les différentes vues de la scène et les arêtes des points homologues en correspondance entre images." En considérant un mouvement relatif pour chaque arête du graphe, déterminer le mouvement global de chaque caméra est un problème de calcul de «mouvement moyen» sur l'ensemble du graphe : (*motion averaging*). La solution qui pour chaque cycle du graphe réalise une boucle fermée est recherchée. Autrement dit, la composition des mouvements relatifs de chaque cycle doivent être équivalente à une transformation identité (cf. figure 6.1).

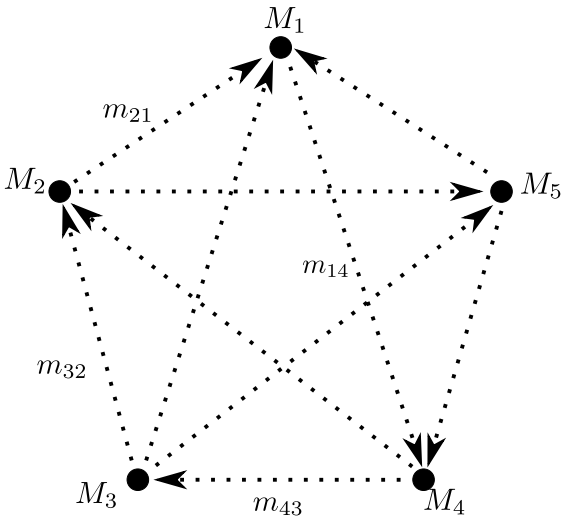


FIGURE 6.1 – Les transformations relatives d'un cycle sont cohérentes si leur composition est un mouvement nul (identité) : $m_{14} \circ m_{43} \circ m_{32} \circ m_{21} = Id$.

En considérant pour chaque arête des estimations robustes de matrices essentielles, on peut extraire une rotation et une translation relative par paire de caméras connectées. On note que les rotations peuvent être composées directement le long d'un cycle, mais pas les translations car ces dernières sont estimées à un facteur d'échelle près (cf. géométrie essentielle). De ce fait, il est fréquent de déterminer d'abord les rotations globales de chaque vue puis d'estimer la translation de chaque caméra et la structure de la scène dans un second temps.

Les méthodes de calibration globale tentent de répartir uniformément les erreurs d'estimation sur l'ensemble du graphe en contraignant les contraintes de cyclicité à être respectée. C'est ce point clef qui supprime les risques de dérive, que ne possèdent pas les méthodes séquentielles. L'utilisation des méthodes globales pose deux problèmes principaux :

1. Il est nécessaire de détecter les données erronées (mouvements relatifs aberrants) au sein du réseau de caméras sinon la reconstruction en sera pénalisée.
2. Le fait de traiter le problème de manière globale implique de considérer toutes les données à disposition ce qui fait que la taille du problème à résoudre croît rapidement avec la taille du jeu d'images.

Le problème étant posé, nous allons nous intéresser aux solutions proposées par la littérature pour les trois problémes principaux des chaînes de reconstruction globale :

- le calcul des rotations globales,
- la détection des rotations erronées,
- l'estimation des translations et de la structure.

Calcul des rotations globales

Étant données des rotations relatives R_{ij} entre les vues i et j, extraites de matrices essentielles, la tâche consiste à estimer la rotation globale de chaque vue R_i (par rapport à une référence arbitraire donnée). La cohérence de mouvement entre rotations relatives et rotations globales peut s'écrire $R_j = R_{ij}R_i$. Calculer des rotations globales consiste à résoudre le problème suivant :

$$\begin{aligned} \underset{\{R_k\}_k}{minimiser} \quad & d(R_j, R_{ij}R_i), \quad \forall i,j \\ & R_k \text{ orthogonale}, \quad \forall k \in \{1,...,n\} \end{aligned} \tag{6.1}$$

Plusieurs façons d'effectuer l'optimisation de l'équation 6.1 sont proposées dans la littérature : SHARP et al. (2001) distribuent les erreurs observées d'un cycle sur chaque rotation relative composant ce cyle. Les erreurs sont réparties de manière répétée sur un ensemble de cycles jusqu'à stabilisation de la solution. GOVINDU (2001) identifie une solution approchée par ajustement aux moindres carrés en utilisant les rotations représentées par des quaternions. MARTINEC et PAJDLA (2007) proposent d'identifier une solution approchée aux moindres carrés puis de rechercher les matrices orthogonales les plus proches à posteriori. ARIE-NACHIMSON et al. (2012) proposent de résoudre ce problème sous forme d'un problème d'optimisation semi-définie positive (*SOCP*) afin de contraindre les matrices de rotations recherchées à être orthogonales. GOVINDU (2004) et HARTLEY et al. (2013) proposent d'utiliser l'espace SO(3) et l'algèbre de Lie afin de réaliser le calcul des rotations globales. CRANDALL et al. (2011) utilisent les algorithmes de *Belief Propagation (BP)* pour estimer les rotations globales dans un espace discret. Les rotations ainsi calculées sont ensuite optimisées par minimisation non linéaire. Le système nécessitant des rotations globales a priori pour quelques images et un axe de rotation fixe supposé, la méthode n'est pas utilisable dans le cas général.

Nous invitons le lecteur à consulter (HARTLEY et al. 2013) pour un rapport complet sur le problème de *rotation averaging* et une analyse des différentes méthodes d'estimation de la littérature. Ces solutions fonctionnent avec du bruit de mesure parmi les rotations mais elles ne sont pas robustes à la présence de données aberrantes.

Calcul robuste des rotations globales

Étant donné que les rotations relatives R_{ij} peuvent contenir des données aberrantes (*outliers*), le calcul des rotations globales se doit d'être robuste. La tâche consiste à identifier les arêtes erronées sur le graphe composé des rotations relatives entre images (cf. figure 6.2). Deux méthodes sont principalement utilisées : l'analyse par arbres recouvrants, ou bien l'analyse de consistance de cycles.

Les approches utilisant un arbre recouvrant, (*spanning tree*), utilisent RANSAC pour déterminer un consensus de rotations globales s'ajustant le mieux au graphe. Une série

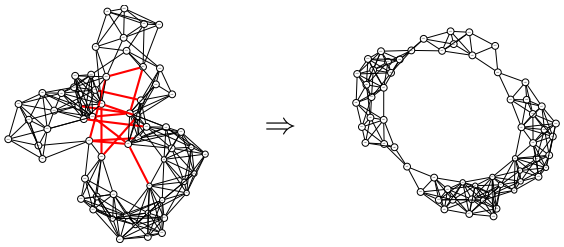


FIGURE 6.2 – Illustration de l'estimation robuste de rotations avec un graphe pour une séquence d'images acquises en tournant autour d'un bâtiment. Un cycle dominant devrait apparaître mais des mouvements relatifs aberrants (arêtes rouges) sont présent. (Ex. Des arêtes relient des côtés opposés de façade à cause de similitudes et de symétries locales dans les images). Après suppression de ces arêtes erronées, le cycle dominant apparaît.

d'arbres recouvrants sont tirés au hasard (cf. figure 6.3), ce qui permet le calcul des rotations globales pour la série d'images par composition des rotations relatives le long de ces arbres. Le consensus d'arêtes en adéquation avec chaque arbre est évalué. L'erreur d'angle de la matrice $R_i^T R_{ij} R_j$ représentant l'adéquation d'ajustement des rotations globales aux rotations relatives est estimé pour chaque arête. Si l'angle est inférieur à une valeur δ, l'arête est ajoutée au consensus, sinon elle est rejetée. L'arbre ayant le plus large consensus est retenu. Ces approches sont utilisées par GOVINDU (2006) ainsi que OLSSON et ENQVIST (2011). Cette approche stochastique ne fournit aucune garantie de déterminer la solution globale idéale (car l'espace des solutions n'est pas couvert de manière complète). Des seuils δ de valeur 0.25° et 1° sont respectivement utilisés par GOVINDU (2006) et OLSSON et ENQVIST (2011).

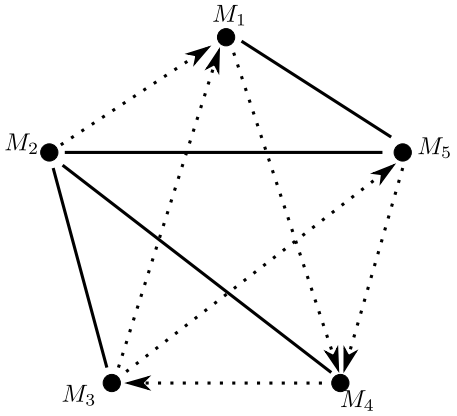


FIGURE 6.3 – Exemple d'arbre recouvrant : tous les sommets sont accessibles à partir de n'importe quel autre (dans la composante connexe) en passant par un chemin unique (les arêtes pleines).

ENQVIST et al. (2011) proposent de supprimer les cycles qui ne correspondent pas à une transformation unité. Soit un graphe ou chaque arête est pondérée par le nombre de correspondances relatives identifiées entre les deux sommets associés. Un arbre recouvrant de poids maximal est calculé pour calculer les rotations globales. Chaque arête n'appartenant pas à l'arbre crée un cycle. En composant les rotations le long de chaque cycle formé par ces arêtes, une erreur d'angle est calculée. Si l'erreur est faible, l'arête est conservée sinon elle est rejetée. Afin d'obtenir des mesures cohérentes en fonction de la longueur l des cycles considérés. Chaque erreur d'angle est pondérée par le fac-

teur $1/\sqrt{l}$. Le constat est le même que pour les méthodes précédentes, si l'arbre recouvrant de poids maximal est erroné, alors les rotations globales estimées seront également fausses.

ZACH et al. (2010) proposent de détecter les fausses géométries épipolaires par inférence bayésienne. Les erreurs de composition sur les cycles du graphe sont modélisées par deux probabilités. La probabilité que seul un bruit d'estimation soit présent le long du cycle et la probabilité qu'au moins une arête le long de ce cycle soit fausse. Une solution au problème est obtenue par optimisation convexe en posant chaque arête comme variable binaire validant ou rejetant les données de l'arête. Considérer l'ensemble de tous les cycles du graphe de caméras n'est pas envisageable car beaucoup trop grand en pratique. De ce fait seul un nombre de cycle limité est utilisé. Des cycles de longueur maximale 6 recouvrant l'ensemble du graphe sont considérés.

Une fois les rotations globales R_i estimées, les translations globales T_i peuvent être calculées. Deux approches existent en fonction du nombre d'éléments à calculer. L'une consiste à rechercher d'abord les translations, puis à calculer la structure par triangulation. L'autre consiste à calculer les translations et la structure simultanément.

Calcul des translations seules

Dans la première approche, l'idée est de considérer uniquement les translations globales comme inconnues. Le nombre de variables considéré est ainsi «minimal» pour le problème. Les déplacements de caméras sont calculées à partir des translations relatives. Le calcul de la structure et l'utilisation des correspondances de points homologues pour la triangulation des traces est ainsi repoussé à la fin de la chaîne.

GOVINDU (2001) propose de déterminer les translations T_i à partir des vecteurs $\mathbf{t}_{ij}$ (directions de translation extraites des matrices essentielles). Les équations de cohérence entre les directions de translations relatives et les translations globales peuvent s'écrire sous la forme suivante :

$$T_{ij} = T_j - R_{ij}T_i. \tag{6.2}$$

Étant donné que les translations relatives T_{ij} sont identifiées à un facteur d'échelle variable λ_{ij} près, nous pouvons écrire : $\mathbf{t}_{ij} = \lambda_{ij}(T_j - R_{ij}T_i)$. En considérant le produit vectoriel on peut écrire :

$$[\mathbf{t}_{ij}]_\times (T_j - R_{ij}T_i) = 0 \tag{6.3}$$

L'équation 6.3 permet d'identifier aux moindres carrés les translations globales T_i inconnues à partir de translations relatives T_{ij} de directions $\mathbf{t}_{ij}$ d'échelles inconnues λ_{ij} (cf. figure 6.4).

SIM et HARTLEY (2006) quant à eux optimisent des positions de caméras en minimisant des erreurs d'angles entre des vecteurs directeurs $\mathbf{t}_{ij}$ et les positions de n caméras globales $\{T_i\}_i = \{T_1, \ldots, T_i, \ldots, T_n\}$. Ainsi avec des vecteurs directeurs issus de tenseurs bifocaux comme la matrice essentielle, l'équation suivante est minimisée :

$$\underset{\{T_k\}_k}{\text{minimiser}} \quad \max_{i,j} \quad tan\theta_{ij}, \quad \forall i,j \tag{6.4}$$

θ_{ij} étant l'angle entre le vecteur $\mathbf{t}_{ij}$ et directions de translation $T_i - T_j$ entre les positions de caméras $\{T_i, T_j\}$ (cf. figure 6.5). En considérant des tenseurs tri-focaux on obtient des résultats plus précis. Ce problème de minimisation est résolu par utilisation d'un *Second*

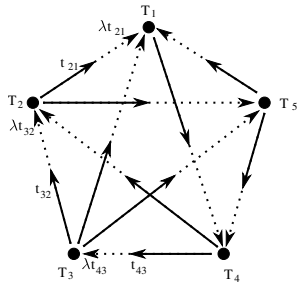


FIGURE 6.4 – Etant donné un ensemble de directions de translations relatives $\mathbf{t}_{ij}$, les positions globales T_i des caméras sont recherchées.

Ordre Cone Program, (SOCP) et d'une bissection pour rechercher sous quelle précision le programme quadratique est réalisable.

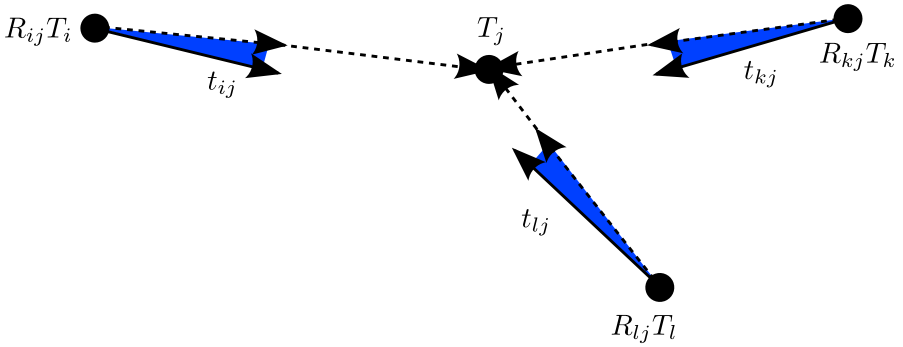


FIGURE 6.5 – Un ensemble de translations globales $\{T_i\}_i$ est recherché en fonction des directions de translations relatives $\mathbf{t}_{ij}$ par minimisation de la plus grande des erreurs angulaires observées (angles bleus). Pour lisibilité seules les contraintes que portent T_j sont représentées.

ARIE-NACHIMSON et al. (2012) utilisent une minimisation aux moindres carrées de l'équation de cohérence épipolaire pour estimer les translations globales inconnues (cf. équation 6.5). Soit $R_j\hat{\mathbf{x}}_k^j$ le point de l'image j visualisant le point 3D k en repère caméra local. Le principal problème de cette méthode est qu'aucune correspondance de points erronés ne doit être présente car toutes les correspondances de points homologues sont utilisées. De plus, cette équation est mise en défaut lorsque les déplacements de caméra sont en ligne droite ou en rotation pure (cf. RODRIGUEZ et al. (2011)).

$$(\mathbf{t}_i - \mathbf{t}_j)^T (R_i\hat{\mathbf{x}}_k^i \times R_j\hat{\mathbf{x}}_k^j) = 0 \tag{6.5}$$

Limitations des approches présentées.
Ces méthodes présentent l'avantage de considérer un nombre très réduit de variables pour estimer les translations globales mais ont pour désavantage d'être sensibles aux erreurs d'estimation des translations relatives pour GOVINDU (2001) et SIM et HARTLEY (2006), ou à la présence de correspondances aberrantes parmi les traces pour ARIE-NACHIMSON et al. (2012).

Calcul simultané des translations et de la structure

HARTLEY et SCHAFFALITZKY (2004b) définissent l'estimation des translations $\mathbf{t}_k$ et des points 3D X_j à partir de rotations globales R_i connues par la minimisation des erreurs résiduelles observées dans les images. Le problème est posé sous le nom : *translation and structure from known rotation*. On minimise pour les vues considérées l'erreur quadratique de re-projection ρ :

$$\rho(\mathbf{t}_i, X_j) = \left\| (x_j^i(1) - \frac{R_i^1 X_j + \mathbf{t}_i^1}{R_i^3 X_j + \mathbf{t}_i^3}, x_j^i(2) - \frac{R_i^2 X_j + \mathbf{t}_i^2}{R_i^3 X_j + \mathbf{t}_i^3}) \right\|_2, \tag{6.6}$$

avec R_i^1 désignant la première ligne de la matrice de rotation globale de la caméra i et $\mathbf{t}_i^k$ la k^e composante du vecteur de translation globale de la caméra i et $(x_j^i(1), y_j^i(2))$ l'observation du point 3D X_j par l'image i. Ce problème impliquant des contraintes quadratiques est résolu par un programme d'optimisation semi-définie positive (*SOCP*). Cependant, afin de garantir une solution valide, l'ajout de contraintes est nécessaire :

- Les points 3D sont forcés à se situer devant les caméras : la profondeur de chaque points 3D est définie supérieure ou égale à 1,
- Les translations globales sont définies à une translation constante près. L'ambiguïté est levée en imposant un repère local arbitraire pour la première caméra : $\mathbf{t}_1 = (0,0,0)^T$.

Le problème étant quasi-convexe, une optimisation globale directe n'est pas possible. Mais l'ajout d'une variable additionnelle permet de rechercher l'existence d'une solution sous une précision γ donnée :

$$\begin{aligned} \underset{\{\mathbf{t}_k\}_k, \{X_l\}_l, \gamma}{\text{minimiser}} \quad & \gamma \\ \text{tel que} \quad & \rho(\mathbf{t}_i, X_j) \leq \gamma, \quad \forall i, j \\ & R_i^3 X_j + \mathbf{t}_i^3 \geq 1, \quad \forall i, j \\ \text{et} \quad & \mathbf{t}_1 = (0,0,0). \end{aligned} \tag{6.7}$$

Une dichotomie sur les valeurs γ (algorithme de bissection) permet d'identifier la plus petite valeur γ jusqu'à laquelle les contraintes du programme linéaire sont toujours réalisables. Une suite de programmes quadratiques doit être réalisée pour identifier les variables $\{\mathbf{t}_k\}_k, \{X_l\}_l$ pour lesquelles l'erreur de reprojection est la plus faible : la plus petite valeur de γ donnant lieu à une solution. La plus grande erreur résiduelle observée étant inférieure à γ, on obtient ainsi une solution optimale au problème sous la norme l_∞. Nous invitons le lecteur à consulter (KAHL et HARTLEY 2008) pour plus de détails.

AGARWAL et al. (2008) démontrent expérimentalement, à travers une étude comparative de différents algorithmes de bissection, que l'algorithme de bissection GUGAT (1996) est la méthode permettant de calculer la valeur γ optimale le plus rapidement. Cependant il est important de noter que cette minimisation considère la totalité des données et qu'elle est sensible à la présence de données aberrantes. L'expression des erreurs résiduelles implique de considérer l'ensemble des points homologues de la scène (traces), l'ensemble des translations et l'ensemble des points 3D de la scène. Le problème à résoudre est donc d'autant plus gourmand en temps de calcul et en mémoire que la scène est grande.

DALALYAN et KERIVEN (2012) proposent une méthode robuste éliminant dans un premier temps les données aberrantes, puis dans un second temps calcule une formulation simplifiée du problème 6.7. La méthode repose sur deux programmes linéaires (*LP*) :

1. On identifie les correspondances de points homologues en contradiction avec la géométrie de la scène et le second détermine les translations et la structure de la scène.
2. On résout la formulation 6.7 en utilisant une expression des erreurs résiduelles en norme l_∞ :

$$\rho(\mathbf{t}_i, X_j) = \left\| (x_j^i(1) - \frac{R_i^1 X_j + \mathbf{t}_i^1}{R_i^3 X_j + \mathbf{t}_i^3}, x_j^i(2) - \frac{R_i^2 X_j + \mathbf{t}_i^2}{R_i^3 X_j + \mathbf{t}_i^3}) \right\|_\infty . \tag{6.8}$$

   Cela permet de résoudre le problème en utilisant une suite de programmes linéaires (*LP*) qui sont plus rapides à résoudre que des programmes quadratiques (*SOCP*). De plus, les correspondances aberrantes étant détectées et supprimées à l'étape 1, une problème de moindre taille est à traiter.

OLSSON et ENQVIST (2011) proposent de déterminer les points homologues aberrants ainsi que la structure et le mouvement de translation globale des caméras de manière simultanée. Leur formulation implique l'ajout de *slack-variables* pour chaque erreur résiduelle afin de mesurer si une erreur doit être compensée ou non pour calculer la meilleure solution globale. En minimisant la somme des *slack-variables*, les *outliers* sont identifiés comme les variables étant «pénalisées». La méthode fonctionne dans la pratique bien que théoriquement il n'y ait pas de garantie de succès. Cette solution impliquant l'ajout d'une variable par erreur résiduelle fait que le programme linéaire à résoudre croît de manière encore plus rapide, par rapport au cas sans slack variables, avec la taille des jeux de données considérés.

ZACH et POLLEFEYS (2010) utilisent les méthodes proximales pour calculer une solution robuste à la formulation 6.7. Il en résulte un algorithme basé sur un gradient qui permet de calculer plus rapidement la solution que par l'utilisation d'une séquence de *SOCP*. Cependant, seule une solution approchée est trouvée car les auteurs précisent que la condition d'arrêt idéale de l'algorithme itératif de minimisation n'a pas pu être formulée.

Autres approches

MARTINEC et PAJDLA (2007) proposent de supprimer les fausses géométries relatives par itération. Ils utilisent de manière répétée une chaîne de calibration globale pour supprimer à chaque itération la paire présentant les plus larges résidus moyens. Le principal problème est que la condition d'arrêt n'est pas définie, MARTINEC et PAJDLA (2007) réalisent alors la procédure un certain nombre de fois variable en fonction de la scène considérée. Afin de conserver des temps de calcul raisonnables l'équation 6.7, est optimisée avec une sous-partie des correspondances de points homologues. Chaque sous partie repose sur un minimum de 4 points en correspondance par paire d'images ayant validé une géométrie épipolaire.Les calculs sont alors accélérés, mais il y a une perte en précision sur la solution pour la structure et les positions de caméra calculées.

COURCHAY et al. (2012) utilisent une paramétrisation de tenseurs tri-focaux construits à partir de matrices fondamentales. Des homographies permettant de fusionner les tenseurs tri-focaux paramétriques respectant au mieux les contraintes de cyclicité locale sont calculées par programmation linéaire. Cependant la méthode est limitée à l'utilisation d'un seul cycle de tenseurs tri-focaux et ne supporte pas de relations épipolaires érronées.

6.2 Une approche pour le passage à l'échelle basée sur des triplets

Étant donné une série d'images $\{I_1, \ldots, I_n\}$ et une calibration interne $\mathbf{K}_i$ supposée connue pour chaque caméra, notre but est de déterminer de manière robuste la position et l'orientation globale de chaque caméra à partir de mouvements relatifs entre images. Nous venons de voir que les méthodes de calibration globales reposent généralement sur trois points principaux :

- la détection des rotations relatives erronées,
- le calcul des rotations globales,
- l'estimation des translations et de la structure.

Considérer l'estimation robuste des translations et de la structure simultanément compromet le passage à l'échelle. L'utilisation d'un programme linéaire ou quadratique de grande taille possédant un nombre important de variables requiert de longues optimisations. Des heures et beaucoup de mémoire vive sont alors nécessaire pour résoudre les translations et la structure de la scène. (OLSSON et ENQVIST 2011) rapportent 7 heures de calcul pour 480 caméras et 77182 points 3D.

Comme GOVINDU (2001) ; SIM et HARTLEY (2006) ; ARIE-NACHIMSON et al. (2012) nous avons choisi de découpler l'estimation des translations et de la structure. Ce choix est primordial pour assurer un passage à l'échelle et une parallélisation supplémentaire de la chaîne. Notre chaîne estime donc tout d'abord les rotations, puis les translations globales à partir de mouvements relatifs, et détermine enfin la structure. Le nombre d'inconnues à identifier et la taille des données considérées sont ainsi réduits pour chaque étape. Les calculs impliquant le plus de données, pour la détermination et l'optimisation de la structure, sont ainsi repoussés à la fin du processus.

Nous avons choisi de faire reposer notre chaîne de calibration sur une primitive de base, le triplet d'images, pour les raisons suivantes :

- Un triplet forme naturellement un cycle permettant de vérifier et de forcer la composition de mouvements en un mouvement identité,
- Un triplet permet d'évaluer la géométrie tri-focale. Grâce à une métrique de type point-point-point entre trois points images, la suppression des correspondances aberrantes est très efficace. Les fausses correspondances établies à tort par géométrie épipolaire (les correspondances points-lignes le long des lignes épipolaires) sont filtrées et ainsi évitées.
- Des facteurs d'échelles entre les translations relatives sont estimés pour un tenseur tri-focal ce qui permet de calculer des positions de caméra même lors de mouvements rectilignes.

Notre chaîne est une généralisation robuste et rapide de la chaîne présentée par SIM et HARTLEY (2006). Plutôt que de considérer une séquence d'images unique formant un réseau de caméras à boucle simple sans mesures aberrantes, nous considérons ici des réseaux de caméras complexes pouvant être fortement contaminés par des mouvements relatifs erronés et l'estimation de la structure.

Nous proposons dans un premier temps d'analyser les éléments clefs de notre chaîne de calibration globale basés sur des triplets. Puis nous montrons comment l'assemblage de ces briques de base permet de former une chaîne de reconstruction globale (cf. section 6.3) réalisant le nettoyage et l'utilisation d'un graphe de mouvements relatifs pour le positionnement d'un réseau de caméras. Cette section démontre les points suivants :

1. Nous montrons que l'utilisation itérative de la méthode d'inférence bayésienne de (ZACH et al. 2010) et l'utilisation de la normalisation des erreurs de rotations par la longueur de cycle de (ENQVIST et al. 2011) supprime la plupart des arêtes aberrantes du graphe de connexité épipolaire. Ces modifications permettent une estimation plus stable des matrices de rotations globales R_i.
2. Afin d'obtenir des translations relatives stables à fusionner, nous présentons une estimation robuste et rapide de tenseur tri-focal réduit *a contrario* sous la norme l_∞. Ces translations relatives sont ensuite fusionnées par optimisation convexe pour déterminer les translations globales T_i sous la norme l_∞.
3. Une structure possédant très peu d'*outliers* est obtenue et optimisée de manière conjointe avec les caméras par ajustement de faisceaux. Nous montrons que la fusion de toutes les structures locales validées par géométrie tri-focale permet d'établir une structure globale extrêmement stable comportant peu, voir, aucun point 3D aberrant. Une introduction progressive de variables (rotations, translations et points 3D) en fonction de leur confiance est utilisée pour éviter de compenser des variables entre elles et ainsi éviter une convergence vers une solution locale.

6.2.1 Inférence de graphes de rotations relatives

Étant donné un graphe de rotations relatives, nous souhaitons estimer les rotations globales R_i pour chaque caméra de manière robuste. Pour cela, nous détectons et supprimons les rotations relatives aberrantes, puis effectuons le calcul des rotations R_i à partir des rotations R_{ij} stables (cf. figure 6.6).

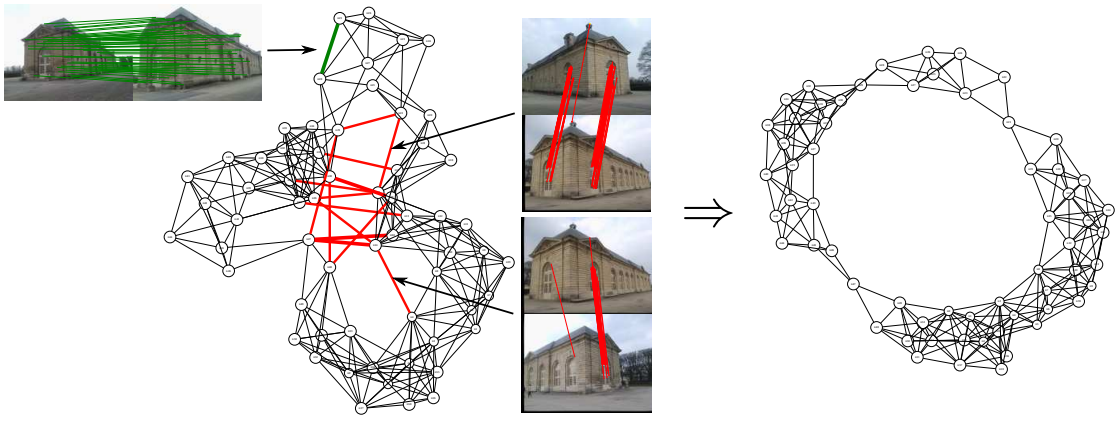


FIGURE 6.6 – Résultat de l'inférence des rotations relatives sur un graphe représentant des images acquises en tournant autour d'un bâtiment. Des arêtes relient des côtés opposés de façade à cause de similitudes et symétries locales dans les images. Ces rotations relatives sont localement connectées avec les points retenus validés par estimation robuste, mais elles ne sont pas correctes dans un contexte global.

Afin de détecter les estimations relatives erronées nous utilisons une variante de la méthode proposée par ZACH et al. (2010) (inférence traitée par optimisation convexe et modélisation bayésienne). Notre variante normalise les erreurs de cycles en fonction

de leur taille en utilisant les résultats rapportés par ENQVIST et al. (2011) : un facteur de normalisation $1/\sqrt{l}$ vient pondérer les erreurs de composition en fonction de la longueur l du cycle.

Cependant nous avons constaté expérimentalement que cette méthode ne calculait qu'une solution approximative. Pour palier ce problème nous proposons d'effectuer l'inférence de manière itérative. Tout d'abord, la première itération retire les estimations les plus grossières, puis des estimations aberrantes de moins grande amplitude sont rejetées d'itération en itération. Lorsque le bruit moyen des compositions de rotations relatives arrive autour de 2°, l'inférence cesse de rejeter des arêtes. Cela est cohérent avec le bruit de base supposé par la méthode bayésienne (ZACH et al. 2010). Finalement nous supprimons les arêtes appartenant aux triplets ayant une erreur de composition supérieure à 2° sont supprimées pour garantir l'estimation de rotations globales stables.

Calcul des rotations globales

Les rotations globales sont ensuite calculées par une minimisation aux moindres carrés de l'erreur entre rotations globales et rotations relatives $d(R_j, R_{ij}R_i)$ sur le graphe de rotations relatives nettoyé. Les matrices R_i considérées dans la minimisation n'étant pas orthogonales, les matrices de rotations les plus proches sont calculées en utilisant les bases orthonormales de la décomposition en valeur singulière (MARTINEC et PAJDLA 2007).

Évaluation de l'inférence de rotations relatives

L'itération de l'inférence a été motivée suite à l'observation du comportement de l'inférence sur différents jeux de données (cf. tableau 6.1). On note que la moitié des rotations relatives aberrantes peuvent être encore restantes à l'itération suivante.

Jeux de données \ #Iterations	1	2	3	2° vérification
OrangerieP61	8	4	1	9
OperaP160	7	3	—	125
PanthéonP126	9	2	—	7
AntonyP29	41	2	4	231

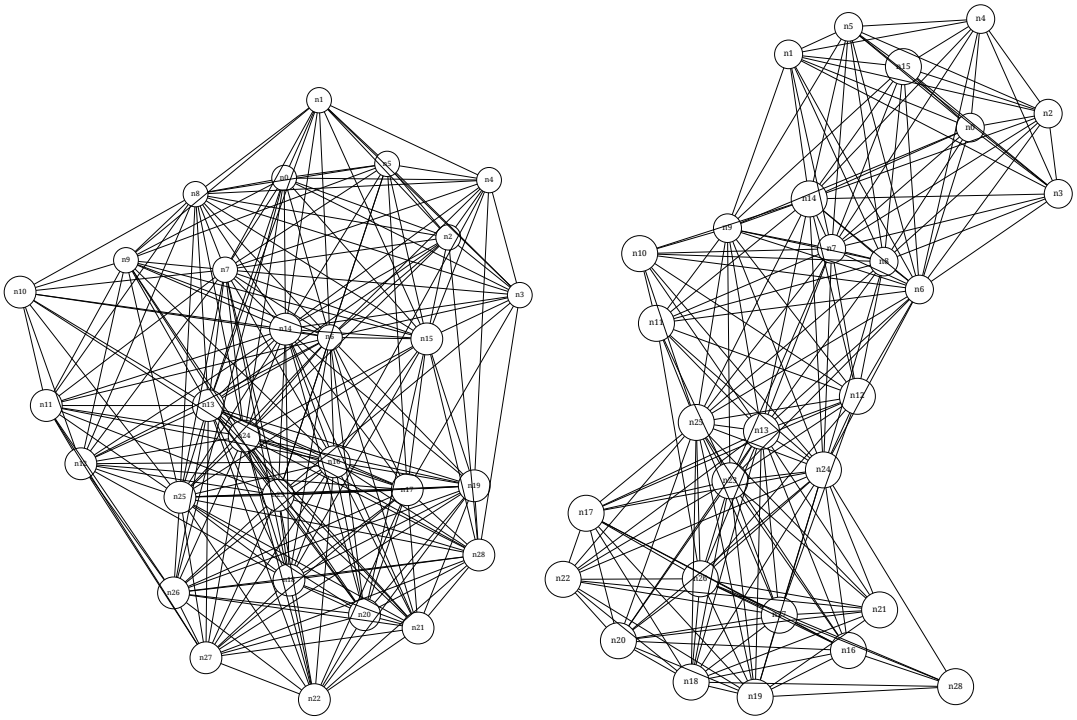
TABLE 6.1 – Nombre d'arêtes rejetées par l'inférence bayésienne à chaque itération et nombres de triplets rejetés par vérification de consistance.

Des exemples de paires dont la rotation relative a été détectée comme aberrante par le processus d'inférence sont illustrés sur les figures 6.7, 6.8, 6.9 pour les jeux de données utilisés du tableau 6.1. On remarque :

- que les paires présentent des correspondances de points qui sont valides dans un contexte local mais invalides dans un contexte global.
- que les points homologues représentent des correspondances entre des éléments géométriques répétés,
- que certaines correspondances géométriquement invalides sont présentes. Elles représentent cependant des correspondances valides pour la métrique point-ligne de la contrainte épipolaire.

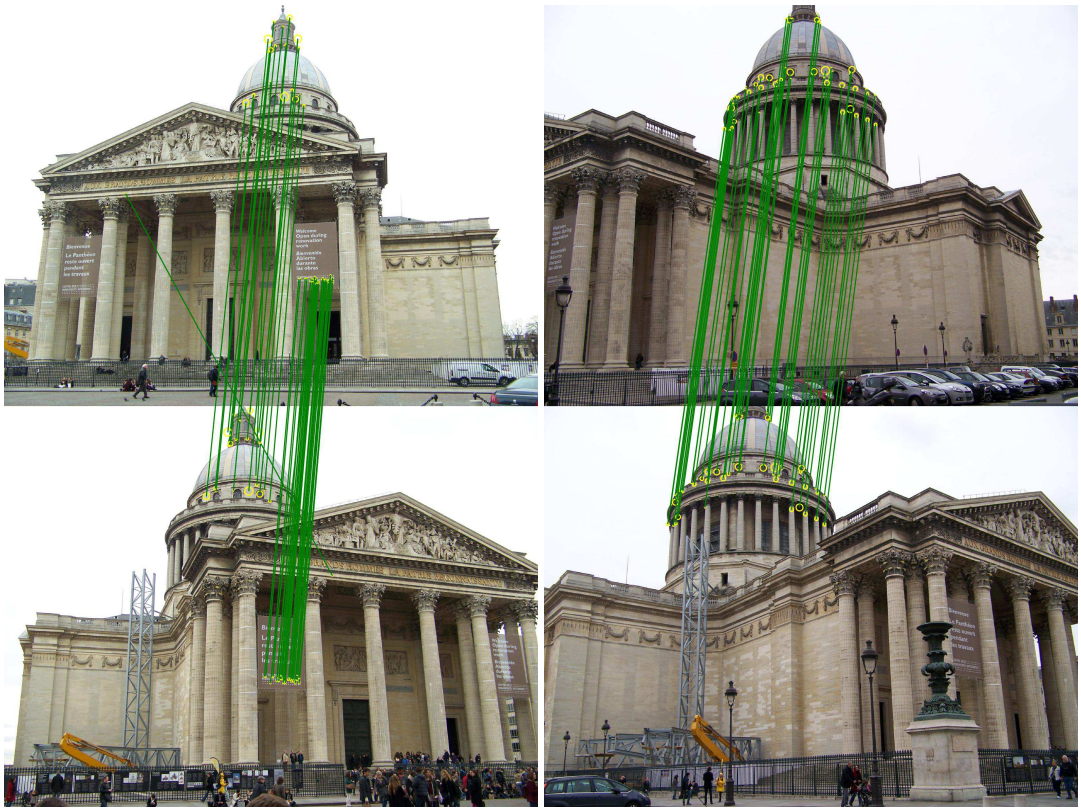
Les graphes avant et après inférence sont affichés et illustrent clairement le travail réalisé par l'inférence (cf. le résultat de la figure 6.9).

(a) Paires identifiées comme aberrantes : on constate des motifs locaux similaires, mais dans un contexte global ces paires sont identifiées comme invalides. On note à gauche que les correspondances de points homologues sont impossibles à réfuter sans contexte global. On note à droite de fausses correspondances valides pour la géométrie épipolaire détectée.

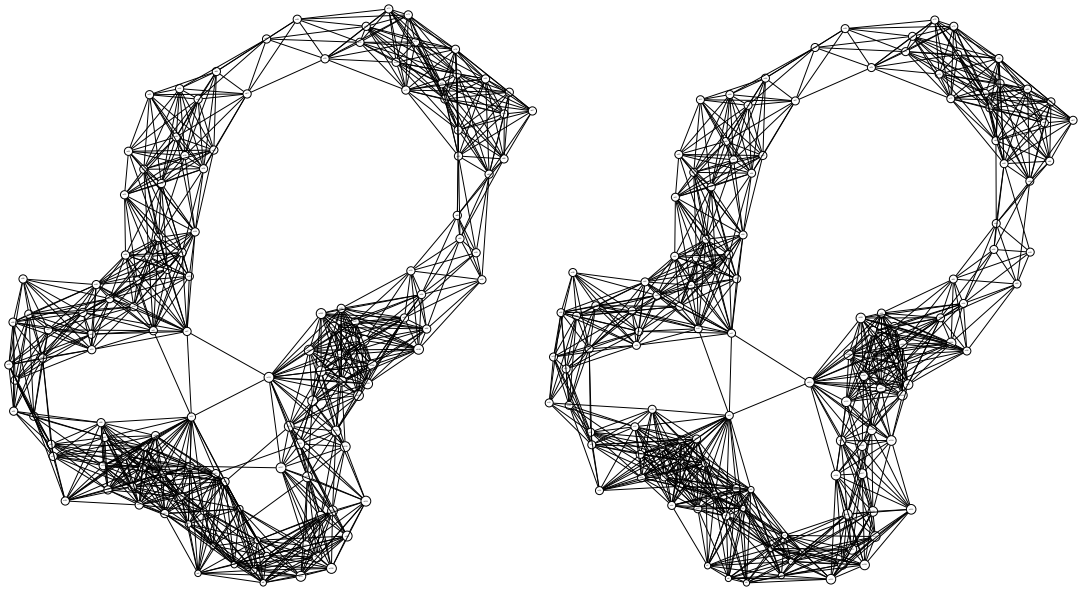


(b) De gauche à droite : le graphe épipolaire avant et après inférence des rotations relatives.

FIGURE 6.7 – Scène AntonyP29 : des façades très similaires sur deux pans d'une même maison donnent lieu à de nombreuses confusions.

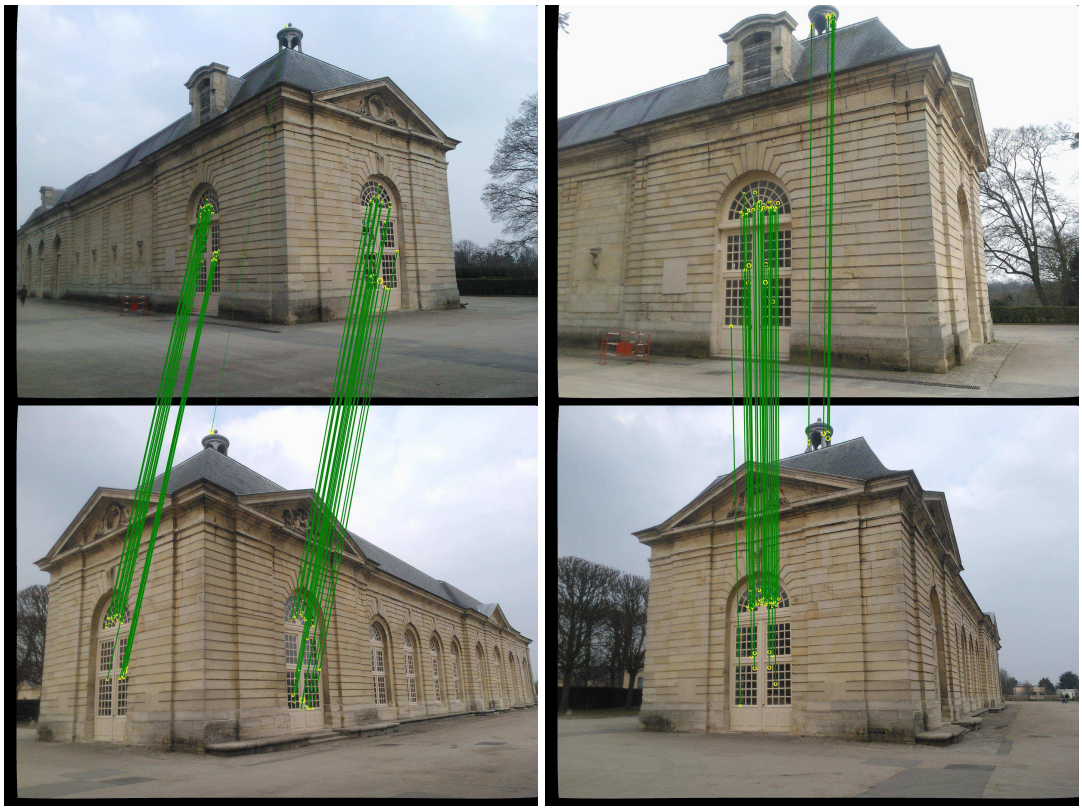


(a) Paires identifiées comme aberrantes : on constate des motifs locaux similaires, mais dans un contexte global ces paires sont identifiées comme invalides.

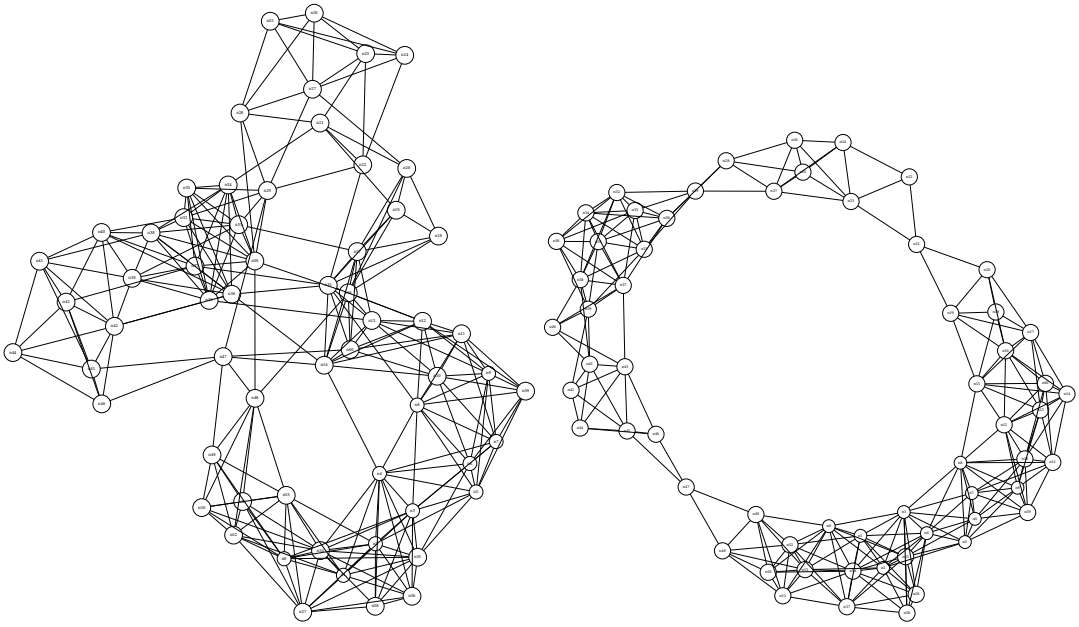


(b) De gauche à droite : le graphe épipolaire avant et après inférence des rotations relatives.

FIGURE 6.8 – Scène PanthéonP126 : des motifs similaires sont présents tout autour du bâtiment et sur des installations temporaires (affiches) et établissent de fausses géométries à l'échelle globale du graphe.



(a) Paires identifiées comme aberrantes : on constate des motifs locaux similaires, mais dans un contexte global ces paires sont identifiées comme invalides.



(b) De gauche à droite : le graphe épipolaire avant et après inférence des rotations relatives.

FIGURE 6.9 – Scène OrangerieP61 : Images acquises en rotation autour d'un bâtiment, on constate que le cycle dominant après inférence, les relations relatives aberrantes dans un contexte global ont bien été supprimés.

6.2.2 Calcul de translations relatives stables par l'utilisation de tenseurs tri-focaux réduits

Plutôt que de considérer des translations relatives issues de matrices essentielles, nous proposons de calculer les translations relatives pour des triplets d'images. Étant donné que les rotations globales R_i sont connues, nous proposons pour chaque triplet l'estimation d'un tenseur tri-focal dit réduit : $\mathbf{T} = \{X_j, \mathbf{t}_i\}_{i,j} | \{R_i\}_{i \in \{1,2,3\}}$. Nous montrerons que les translations relatives identifiées par géométrie tri-focale sont bien plus précises que des estimations réalisées par géométrie épipolaire et qu'elles peuvent être calculées rapidement.

Une solution optimale mais un problème de passage à l'échelle.

Au lieu de considérer une minimisation du tenseur tri-focal réduit par une erreur algébrique (SIM et HARTLEY 2006) aux moindres carrés, nous proposons d'utiliser une minimisation de l'erreur géométrique des erreurs résiduelles de la structure $\{X_j\}_j$ aux mesures images $\{x_j^i\}_{i,j}$ sous la norme l_∞ au sein des 3 images i considérées. Ce problème consiste à résoudre le problème de *translation and structure from known rotation* pour 3 vues (cf. le programme 6.7). Cependant cette solution optimale pose deux inconvénients :

1. La complexité du problème augmente de manière polynomiale avec le nombre de variables considérées (SEO et HARTLEY 2007),
2. Rendre la méthode robuste impose d'utiliser la méthode d'OLSSON et ENQVIST (2011) qui ajoute autant de *slack variables* que d'erreurs considérées dans les images ou bien d'utiliser la formulation de DALALYAN et KERIVEN (2012) qui utilise deux étapes basées sur l'utilisation de programmes linéaires. Pour OLSSON et ENQVIST (2011) l'ajout de ces *slack variables* fait qu'en pratique une solution ne peut pas être trouvée dans toutes les situations. Pour DALALYAN et KERIVEN (2012) la complexité reste polynomiale et impose la résolution de programmes linéaires avec de nombreuses variables et contraintes.

Afin de ne pas être dépendant d'un problème de complexité croissante avec la taille des données à traiter (le nombre de traces) nous proposons d'évaluer la validité d'hypothèses, générées à partir d'une formulation minimale du problème, *a contrario*. Nous proposons de calculer chaque tenseur tri-focal en utilisant :

1. *AC-RANSAC* comme méthode d'estimation robuste,
2. Un générateur d'hypothèse de tenseur tri-focaux précis, rapide et compact reposant sur :
   - une formulation convexe, pour garantir la précision des solutions calculées,
   - l'utilisation d'un faible nombre de traces, pour garantir une complexité minimale et un temps de calcul optimal.

La section suivante présente la méthode d'estimation *a contrario* utilisée et une série d'expériences qui démontre que notre solution pourra en temps quasi-constant déterminer une solution robuste et précise.

Estimation robuste de tenseur tri-focal réduit *a contrario*

Afin de pouvoir estimer *a contrario* notre tenseur tri-focal réduit nous devons définir les trois éléments suivants : une mesure de similarité, un modèle paramétrique et une mesure de significativité.

La mesure de similarité ρ est définie comme l'erreur de re-projection maximale de la structure X_j observée à travers les 3 images du triplet :

$$\rho(\mathbf{t}_i, X_j) = \left\| (x_j^i(1) - \frac{R_i^1 X_j + \mathbf{t}_i^1}{R_i^3 X_j + \mathbf{t}_i^3}, x_j^i(2) - \frac{R_i^2 X_j + \mathbf{t}_i^2}{R_i^3 X_j + \mathbf{t}_i^3}) \right\|_\infty, \quad \forall i \in \{1,2,3\} \tag{6.9}$$

avec R_i^1 désignant la première ligne de la matrice de rotation globale de la caméra i, $\mathbf{t}_i^1$ la première composante du vecteur de translation local de la caméra i et $(x_j^i(1), x_j^i(2))$ les projections images de la trace du point X_j dans les 3 images i du triplet.

Le modèle paramétrique (le tenseur tri-focal réduit) consiste à résoudre le problème de *translation and structure from known rotation* de l'erreur maximale ρ observée en utilisant le programme linéaire suivant, une minimisation sous la norme l_∞ :

$$\begin{aligned} \underset{\{\mathbf{t}_k\}_k, \{X_l\}_l, \gamma}{\text{minimiser}} \quad & \gamma \\ \text{tel que} \quad & \rho(\mathbf{t}_i, X_j) \leq \gamma, \quad \forall i,j \in \{1,2,3,4\} \\ & R_i^3 X_j + \mathbf{t}_i^3 \geq 1, \quad \forall i,j \\ \text{et} \quad & \mathbf{t}_1 = (0,0,0). \end{aligned} \tag{6.10}$$

L'existence d'une solution, ayant une erreur résiduelle maximale γ est évaluée pour identifier les translations $\{\mathbf{t}_i\}_{i\in\{2,3\}}$ et les points 3D $\{X_j\}_j$ pour un quadruplet de points. Au lieu d'estimer le paramètre γ optimal par bissection pour chaque quadruplet de traces demandé par *AC-RANSAC*, nous le fixons à une valeur fixe $\gamma = 0.5$ pixels. Cette borne fixe est utilisée comme un critère permettant de rejeter facilement des groupes de traces aberrantes car le programme linéaire sera identifié comme non réalisable. La faisabilité du problème (obtention d'une solution sous la borne γ) permet ainsi de fournir ou non à *AC-RANSAC* une hypothèse à évaluer. La formulation la plus compacte du programme linéaire s'exprime avec 4 traces (points 3D) et c'est celle que nous utilisons. Cette formulation nous garantit l'obtention d'une solution avec la plus faible complexité possible. Chaque hypothèse de tenseur tri-focal réduit $M = \{\mathbf{t}_i, X_j\}_{i\in\{1,2,3\}, j\in\{1,\dots,4\}}$ sera évaluée de manière statistique à toutes les n correspondances par la mesure de significativité.

Note. Sachant que $\|\boldsymbol{x}\|_\infty = \max\left(|x^1|, \dots, |x^n|\right)$ et qu'une valeur absolue est exprimable par le biais de 2 inégalités tel que $|x^n| \leq \gamma$ est équivalent à $(-\gamma \leq x^n \leq \gamma)$. Nous pouvons réécrire la contrainte $\rho(\mathbf{t}_i, X_j) \leq \gamma$, pour le programme linéaire, sous la forme de 2 inégalités pour les dimensions des variables considérées (x, y), soit 4 contraintes :

$$\rho(\mathbf{t}_i, X_j) \leq \gamma \Leftrightarrow \left\{ \begin{array}{l} X_j(R_i^k + R_i^3(\gamma - x_j^i(k))) + \mathbf{t}_i^k + \mathbf{t}_i^3(\gamma - x_j^i(k)) \geq 0 \\ X_j(R_i^k - R_i^3(\gamma + x_j^i(k))) + \mathbf{t}_i^k - \mathbf{t}_i^3(\gamma + x_j^i(k)) \leq 0 \end{array} \right\}, \quad \forall k \in \{1,2\}.$$

La mesure de significativité est évaluée par la spécification du NFA générique du chapitre 4.3. Ce NFA nous permet de mesurer l'adéquation de k traces parmi les n du triplets avec un tenseur tri-focal réduit M en cours hypothèse :

$$\text{NFA}(M,k) = (n-4)\binom{n}{k}\binom{k}{4} e_k(M)^{k-4}. \tag{6.11}$$

M étant le tenseur tri-focal réduit obtenu à partir de 4 traces, k le nombre potentiel de traces valides (*inliers*) en cours d'hypothèse et $e_k = \varepsilon_k / \max(w,h)$ la k^{e} erreur définie telle que :

$$\varepsilon_k(M) = k^{\text{e}} \text{ plus petit élément de } \{\max_j \rho(\mathbf{t}_i, X_j)\}_{i\in\{1,2,3\}, j\in\{1,\dots,n\}}. \tag{6.12}$$

où w et h désignent la largeur et hauteur des images considérées, et e_k la probabilité qu'un point ait une erreur résiduelle d'au plus ε_k. Les points X_j sont évalués par triangulation des points images correspondants : $\{(x_j^i, y_j^i)\}_{i=1,2,3}$ et du tenseur hypothèse M. Dans l'équation 6.11, $e_k(M)^{k-4}$ représente la probabilité que $k-4$ correspondances uniformément distribuées et indépendantes aient une erreur d'au plus ε_k dans les trois images. Un tenseur tri-focal réduit M est considéré significatif (non observé par chance) si :

$$\mathrm{NFA}(M) = \min_{5 \leq k \leq n} \mathrm{NFA}(M,k) \leq 1. \tag{6.13}$$

On remarquera que l'index k identifié *a contrario* par le NFA joue, indirectement, le rôle de la variable γ recherché par bissection dans la formulation utilisant toutes les correspondances pour le problème de *translation and structure from known rotation*. Sauf que dans notre cas, nous réalisons une estimation robuste et ce à un coût de calcul bien moindre.

En pratique nous avons observé que 300 itérations sont suffisantes à *AC-RANSAC* pour évaluer à partir de quadruplets de traces des tenseurs tri-focaux réduits stables. Si aucune hypothèse ne vérifie l'équation (6.13), le triplet est rejeté, sinon le triplet avec le plus petit NFA est conservé. L'utilisation de l'optimisation locale d'*AC-RANSAC* réalise 30 itérations de tirages parmi les *inliers* (10% du nombre maximal d'itérations) pour optimiser le premier tenseur tri-focal significatif calculé. Nous réalisons ainsi dans le pire des cas 300 itérations et dans le meilleur des cas 31 itérations pour déterminer une solution. Finalement la solution est affinée par ajustement de faisceaux avec les j traces validées *inliers*.

Nous allons maintenant évaluer la réaction des temps de calcul de l'algorithme aux nombres de traces d'un triplet et la précision de la solution identifiée à travers deux expériences.

Évaluation des temps de calcul

Un jeu de données synthétique est simulé avec des points 3D uniformément réparti dans un volume de dimension $[-1,1]^3$. Un tenseur tri-focal est généré avec trois caméras (de focale 1000 et résolution image 1000*1000) placées le long d'un cercle de rayon 5 aux angles 0°, 20° et 40° et permet de créer des traces parfaites (cf. figure 6.10). Les traces sont perturbées par un bruit uniforme de plus ou moins 1 pixel et 2% d'*outliers* sont introduits afin de vérifier que les estimations sont robustes.

Pour un nombre de traces n de plus en plus grand des tenseurs tri-focaux réduits sont estimés. Les temps de calcul et les précisions des directions de translations pour notre méthode d'estimation *a contrario* et la méthode d'estimation globale utilisant des *slack variables* (Olsson et Enqvist 2011) sont comparés. Les précisions angulaires des translations relatives des tenseurs tri-focaux réduits sont évaluées par rapport aux valeurs exactes en degrés. Le tableau 6.2 récapitule les résultats de l'expérience (chaque ligne présente la moyenne de 100 estimations).

On constate que les deux méthodes fonctionnent : elles identifient une solution précise et ont donc bien éliminés les données aberrantes. Cependant on peut noter que la méthode globale prend un temps croissant substantiellement avec la taille de l'ensemble de traces alors que notre méthode basée sur *AC-RANSAC* détermine une solution à temps quasi-constant. De plus *AC-RANSAC* s'ajuste au bruit avec un comportement plus stable (amplitude plus faible de l'erreur) et détermine des solutions plus précises.

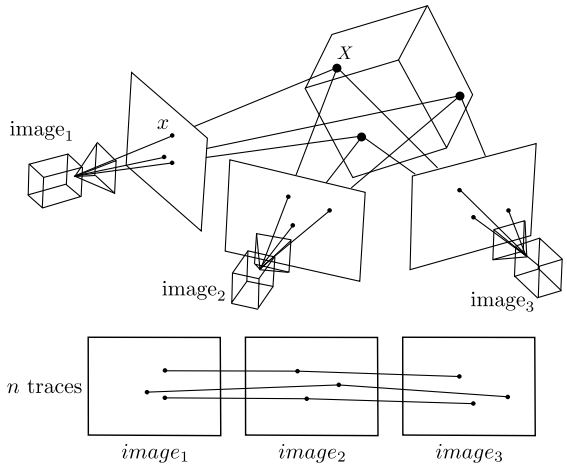


FIGURE 6.10 – Configuration de tests pour évaluer les temps de calcul d'estimation robuste tenseur tri-focaux.

#Points 3D	Temps de calcul (s)		Précision angulaire (°)	
	Global +slack	AC	Global +slack	AC
200	1.37	**0.09**	0.07	**0.03**
400	4.06	**0.11**	0.06	**0.03**
600	7.94	**0.13**	0.04	**0.02**
800	13.1	**0.15**	0.03	**0.02**
1000	19.6	**0.16**	0.03	**0.02**

TABLE 6.2 – Temps de calcul et précision (angle moyen entre les translations relatives vérité terrain et les calculées) pour l'estimation robuste de tenseur tri-focal réduit avec la méthode globale utilisant les *slack variables* (OLSSON et ENQVIST 2011) et notre méthode *a contrario* (combinaison d'un programme linéaire et d'*AC-RANSAC*). Meilleurs temps et précision en gras.

Évaluation de la précision des translations relatives en fonction de la *baseline*

Nous considérons l'expérience menée par (ENQVIST et al. 2011) pour comparer la stabilité de l'évaluation de translations relatives en fonction de l'écartement des caméras (*baseline*). Cependant nous généralisons l'expérience pour comparer la précision de translations relatives estimées à partir de matrices essentielles et de tenseurs tri-focaux.

Nous montrons que l'utilisation d'un tenseur tri-focal permet d'estimer des translations relatives bien plus précises. Une fois le tenseur tri-focal $\{R_i, \mathbf{t}_i, X_j\}_{i,j}$ estimé (ou la matrice essentielle $\mathbf{E} = \{R_1, R_2, \mathbf{t}_1, \mathbf{t}_2\}$) avec leurs translations locales, les translations relatives sont extraites comme suit :

$$\begin{aligned} R_{ij} &= R_j R_i^T \\ \mathbf{t}_{ij} &= \mathbf{t}_j - R_{ij}\mathbf{t}_i. \end{aligned} \tag{6.14}$$

Afin d'évaluer la stabilité d'estimation des translations relatives d'un tenseur bi-focal $\mathbf{E}$ et tenseur tri-focal en fonction de l'écartement des caméras, nous utilisons le cadre de l'expérience précédente où les positions de caméras peuvent varier le long d'un cercle. La première caméra est positionnée à une position fixe 0° et les caméras 2 et 3 aux angles respectifs α et 2α (cf. figure 6.11). Des caméras avec un changement de

baseline, avec un écartement faible à moyen sont simulées en faisant varier α de 1° à 20°. Les projections de ces points 3D, perturbées par bruit uniforme $\in [-1,1]$, sont utilisées pour estimer et comparer la précision des translations relatives issues d'*AC-RANSAC* avec la méthode des 5 points pour identifier la matrice essentielle et d'*AC-RANSAC* avec notre tenseur tri-focal réduit. L'erreur angulaire d'estimation en degré par rapport à la direction de translation exacte est estimée pour les trois directions relatives et moyennée sur 100 expériences.

Les résultats sont présentés figure 6.11 et montrent que lorsque l'écart relatif des caméras augmente, la précision s'améliore pour les deux méthodes. On note que :

- pour la matrice essentielle, les estimations sont de mauvaises qualités avec des angles relatifs de faible amplitude.
- pour le tenseur tri-focal de bien meilleures performances sont obtenues, même pour des valeurs faibles d'écartement des caméras.

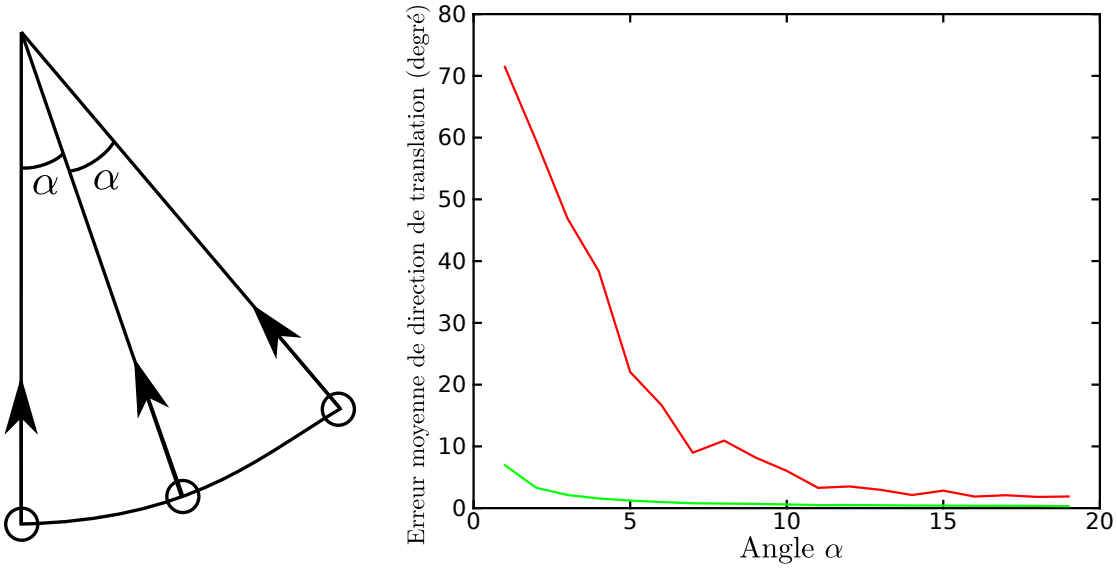


FIGURE 6.11 – A gauche la configuration de caméra utilisée pour l'expérience. A droite les erreurs moyennes de directions de translations pour des matrices essentielles (rouge) et pour notre tenseur tri-focal réduit (vert) en fonction de l'angle inter-caméra α.

Nous venons de présenter une méthode d'estimation robuste de tenseurs tri-focaux réduits basée sur l'estimation de multiples programmes linéaires de taille minimale. Utilisé avec la méthodologie *a contrario AC-RANSAC* nous avons démontré expérimentalement la précision des translations relatives obtenues et la complexité quasi-constante en fonction du nombre de traces considérés par le tenseur en cours d'estimation. Tandis que les translations relatives à deux vues procurent des résultats instables à faible écartement de caméras, nos translations relatives à trois vues sont bien plus précises et ce quelque soit la *baseline* utilisée.

Étant désormais capable d'estimer des translations relatives précises pour l'ensemble d'un graphe où les rotations sont connues nous désirons réconcilier ces vecteurs directeurs dans un système de coordonnées commun pour estimer les translations globales du réseau de caméras.

6.2.3 Fusion de translations relatives sous la norme l_∞ pour le positionnement global rapide d'un réseau de caméras

Étant donné un ensemble de mouvements relatifs $\{R_{ij}, \mathbf{t}_{ij}\}$ (les rotations et directions de translation), nous souhaitons trouver les positions globales $(T_1, \ldots, T_n)$ de toutes les caméras. Il s'agit d'un problème de fusion où l'on recherche les translations globales T_i optimales réconciliant les translations relatives et leurs facteurs d'échelles λ_{ij} associés (cf. illustration 6.12).

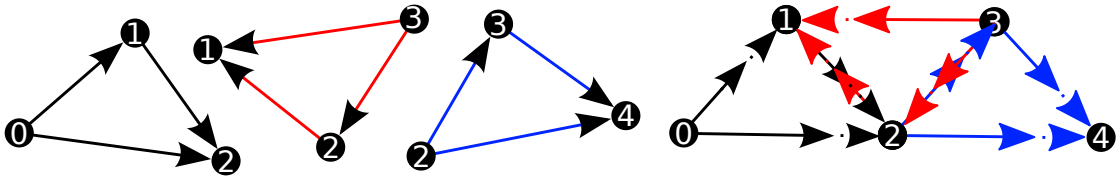


FIGURE 6.12 – Gauche : 3 tenseurs tri-focaux et leur translations relatives respectives. Droite : La fusion optimale des translations relatives et les facteurs λ_{ij} identifiés.

Nous avons vu que GOVINDU (2001) (cf. équation 6.3) propose une solution pour résoudre le problème au moindres carrés. Cependant certaines configurations de vecteurs relatifs $\mathbf{t}_{ij}$ peuvent amener la solution aux moindres carrés à choisir des valeurs sous optimales pour le problème car une taille minimale des vecteurs directeurs $\mathbf{t}_{ij}$ ne peut être imposée (cf. figure 6.13).

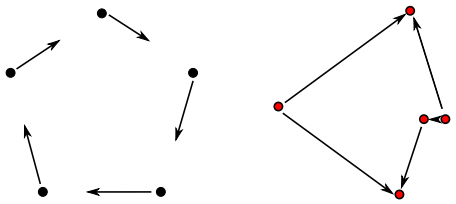


FIGURE 6.13 – A gauche les vecteurs translations $\mathbf{t}_{ij}$ normé initiaux. A droite le résultat déterminé par moindres carrés avec la méthode GOVINDU (2001). La taille des vecteurs n'étant pas contrainte, une solution dégénérée au problème est calculée.

Fusion stable de translations relatives

Nous proposons alors une formulation pour résoudre ce problème de manière stable par ajout de contrainte au problème initial. Étant donné l'équation de consistance entre les translations globales et les translations étendues de leur vecteur échelle on peut écrire :

$$\lambda_{ij}\mathbf{t}_{ij} = T_j - R_{ij}T_i. \tag{6.15}$$

Une solution à cette équation est invariante par translation et par changement d'échelle. Mais l'indétermination d'échelle peut être supprimée en forçant les facteurs λ_{ij} à une valeur définie positive minimum. Et l'indétermination de translation de la solution peut être contrainte en fixant une des translations globales à la position origine. On peut alors écrire :

$$\begin{aligned} \underset{\{T_i\}_i, \{\lambda_{ij}\}_{i,j}}{\text{minimiser}} \quad & \left\| T_j - R_{ij}T_i - \lambda_{ij}\mathbf{t}_{ij} \right\|_2 \\ \text{sujet à} \quad & \lambda_{ij} \geq 1, \quad \forall i, j \\ \text{et} \quad & T_1 = (0,0,0). \end{aligned} \tag{6.16}$$

Cependant on ne peut résoudre ce problème aux moindres carrés car les contraintes ne sont pas exprimables sous cette forme. Nous reformulons alors le problème d'optimisation avec une variable additionnelle γ et une erreur résiduelle de type l_∞ pour obtenir le programme linéaire suivant :

$$\begin{aligned} \underset{\gamma,\{T_i\}_i,\{\lambda_{ij}\}_{i,j}}{\text{minimiser}} \quad & \gamma \\ \text{tel que} \quad & \left\|T_j - R_{ij}T_i - \lambda_{ij}\mathbf{t}_{ij}\right\|_\infty \leq \gamma \\ & \lambda_{ij} \geq 1, \quad \forall i,j \\ \text{et} \quad & T_1 = (0,0,0), \end{aligned} \tag{6.17}$$

avec la contrainte $\left\|T_j - R_{ij}T_i - \lambda_{ij}\mathbf{t}_{ij}\right\|_\infty < \gamma$ formulée par 2 inégalités pour les 3 dimensions (X,Y,Z) des variables considérées, soit 6 contraintes :

$$\left\|T_j - R_{ij}T_i - \lambda_{ij}\mathbf{t}_{ij}\right\|_\infty \leq \gamma \Leftrightarrow \left\{ \begin{array}{l} T_j(k) - R_{ij}^k T_i(k) - \lambda_{ij}\mathbf{t}_{ij}(k) \leq \gamma \\ T_j(k) - R_{ij}^k T_i(k) - \lambda_{ij}\mathbf{t}_{ij}(k) \geq -\gamma \end{array} \right\}, \quad \forall k \in \{1,2,3\}.$$

L'optimisation convexe (6.17) permet de déterminer les translations globales en réconciliant les estimations relatives par le biais de la minimisation sous la norme l_∞ des distances résiduelles entre les translations globales et les vecteurs directeurs augmentés par leur facteur d'échelle λ respectif (cf. figure 6.14). La valeur γ trouvée représente la distance résiduelle maximale pour laquelle une solution a été déterminée (cf. figure 6.14).

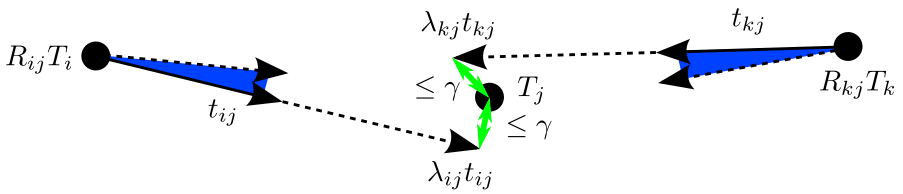


FIGURE 6.14 – Notre approche minimise le maximum des distances euclidiennes (distance verte) tandis que SIM et HARTLEY (2006) minimisent le maximum des erreurs angulaires indiquées en bleu.

Une solution valide est désormais identifiée pour la situation qui était en échec pour la méthode de GOVINDU (2001) (cf. figure 6.15).

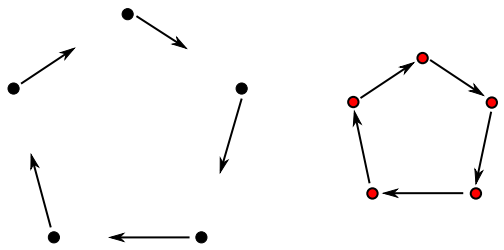


FIGURE 6.15 – A gauche les vecteurs translations $\mathbf{t}_{ij}$ initiaux et les positions de caméras C_i. A droite le résultat déterminé par notre méthode. Ce cas parfait de 5 caméras le long d'un cercle à distance régulière calcule une solution avec des vecteurs λ_{ij} de valeur 1. La solution au problème est à un facteur d'échelle qui demeure inconnu de la vérité terrain.

Cette formulation convient pour un réseau composé de translations relatives indépendantes où chaque translation globale est représentée dans un minimum de deux équations relatives. Cette fusion peut donc impliquer des translations relatives issues de matrices essentielles ou de tenseurs tri-focaux. Cependant, étant donné que les tenseurs tri-focaux impliquent des translations relatives plus précises nous proposons une solution dédiée, permettant de conserver les facteurs d'échelles liés à chaque triplet.

Spécialisation pour des translations relatives de tenseurs tri-focaux

Nous avons vu au début de ce chapitre que chaque tenseur tri-focal estime des facteurs d'échelles pour ses translations relatives. Cet avantage nous permet d'estimer avec une grande précision des structures locales et ce même pour des trajectoires rectilignes. Cependant si l'on considère, lors de la fusion, des λ_{ij} indépendants pour chaque translation relative de ces tenseurs, des déformations affines pour chaque sous-structure peuvent apparaître. Ce problème est évité en utilisant un facteur d'échelle λ^τ unique pour les translations relatives issues d'un même tenseur tri-focal τ, (cf. illustration de la figure 6.12 où chaque triplet τ, désigné par une couleur, sera contraint par un unique facteur λ^τ).

On obtient le programme linéaire suivant :

$$\begin{aligned} \underset{\gamma,\{T_i\}_i,\{\lambda^\tau\}_\tau}{\text{minimiser}} \quad & \gamma \\ \text{tel que} \quad & \left\| T_j - R_{ij}T_i - \lambda^\tau \, \mathbf{t}_{ij}^\tau \right\|_\infty \leq \gamma, \quad \forall \tau, \forall (i,j) \in \tau \\ & \lambda^\tau \geq 1, \quad \forall \tau \\ \text{et} \quad & T_1 = (0,0,0), \end{aligned} \tag{6.18}$$

avec la contrainte $\left\| T_j - R_{ij}T_i - \lambda^\tau \, t_{ij}^\tau \right\|_\infty < \gamma$ formulée par 2 inégalités pour les 3 dimensions (X,Y,Z) des variables considérées, soit 6 contraintes :

$$\left\| T_j - R_{ij}T_i - \lambda^\tau \, \mathbf{t}_{ij}^\tau \right\|_\infty \leq \gamma \Leftrightarrow \left\{ \begin{array}{l} T_j(k) - R_{ij}^k T_i(k) - \lambda^\tau \, \mathbf{t}_{ij}^\tau(k) \leq \gamma \\ T_j(k) - R_{ij}^k T_i(k) - \lambda^\tau \, \mathbf{t}_{ij}^\tau(k) \geq -\gamma \end{array} \right\}, \quad \forall k \in \{1,2,3\}.$$

Le fait de considérer des translations par triplets réalise en pratique une réduction drastique du nombre de variables λ à estimer, le nombre de triplets étant souvent inférieur au nombre d'arêtes du graphe.

La table 6.3 présente le nombre minimal de triplets suffisant pour recouvrir les arêtes des graphes relatif à différents jeux de données. On note que sur l'ensemble des scènes, le nombre de triplets est inférieur au nombre d'arêtes. Le fait d'utiliser des triplets permet ainsi de réduire le nombre de variables de λ et donc de formuler un programme linéaire plus compacte qui sera plus rapide à résoudre.

Plusieurs points théoriques et pratiques confirment que notre méthode présente de meilleures garanties de convergence et de temps de calcul que les méthodes exposées plus tôt.

1. Contrairement à l'approche originale de GOVINDU (2001), nous sommes certains d'identifier une solution non dégénérée et optimale grâce aux contraintes additionnelles formulées. Confère la comparaison entre le cas des figures 6.13 et 6.15.
2. Contrairement à SIM et HARTLEY (2006), nous employons des contraintes simples et identifions une solution par une seule itération d'une programme linéaire. Cette solution est identifiée plus rapidement qu'avec la séquence de *SOCP* utilisée par SIM et HARTLEY (2006) pour minimiser les erreurs d'angles.

Scène	#Images	#Triplets $\#\lambda^{\tau}$	#Arêtes $\#\lambda_{ij}$
CastleP19	19	**39**	69
CastleP30	30	**103**	175
EntryP10	10	**28**	43
FountainP11	19	**39**	69
HerzJesusP8	8	**13**	21
HerzJesusP25	25	**102**	156
MayaHeadP50	50	**149**	218
PanthéonP126	126	**582**	958
HotelCujasP182	182	**840**	1319
PavillonP421	421	**5612**	8041

TABLE 6.3 – Comparaison du nombre d'arêtes et du nombre minimal de triplets permettant de recouvrir les graphes de différentes collections d'images. En gras la configuration qui donnera lieu au moindre nombre de variables λ.

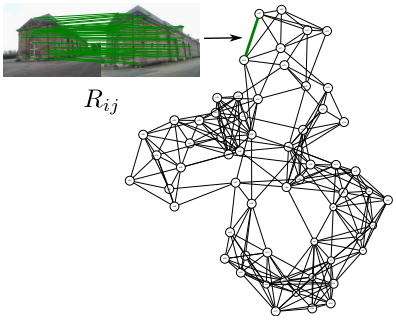
Étant désormais capable de calculer les orientations et les positions des caméras à partir de mouvements relatifs nous proposons de structurer une chaîne de calibration externe globale réalisant l'estimation d'une structure comportant très peu de données aberrantes.

6.3 Mise en place de la chaîne de reconstruction

Nous expliquons maintenant l'approche utilisée pour réaliser notre chaîne de calibration globale. Notre méthode consiste en la succession des étapes suivantes :

1. Calcul du graphe de connexité épipolaire et estimation des matrices essentielles,
2. Vérification de la cohérence des rotations relatives et calcul des rotations globales,
3. Calcul des translations relatives par tenseurs tri-focaux réduits,
4. Calcul des translations globales par fusion des translations relatives,
5. Calcul de la structure et ajustements de faisceaux.

Étape 1 : Calcul du graphe de connexité épipolaire et estimation des matrices essentielles. Les paires d'images en correspondance sont identifiées par mise en correspondance de leurs points saillants (cf. procédure 8). Nous utilisons le détecteur et descripteur SIFT (LOWE 1999), suivi de la politique de rejet DR (cf. section 3.9). Ces correspondances photométriques sont ensuite filtrées de manière géométrique en estimant de manière robuste des matrices essentielles avec *AC-RANSAC*. Afin de d'obtenir des correspondances relatives sous une précision raisonnable, nous limitons la précision maximale possible identifiée *a contrario* sous 4 pixels : $\delta_{ac} \in [0,4]$. Chaque composante connexe du graphe de géométrie épipolaire sera ensuite traitée de manière indépendante (puis elles seront fusionnées si possible).

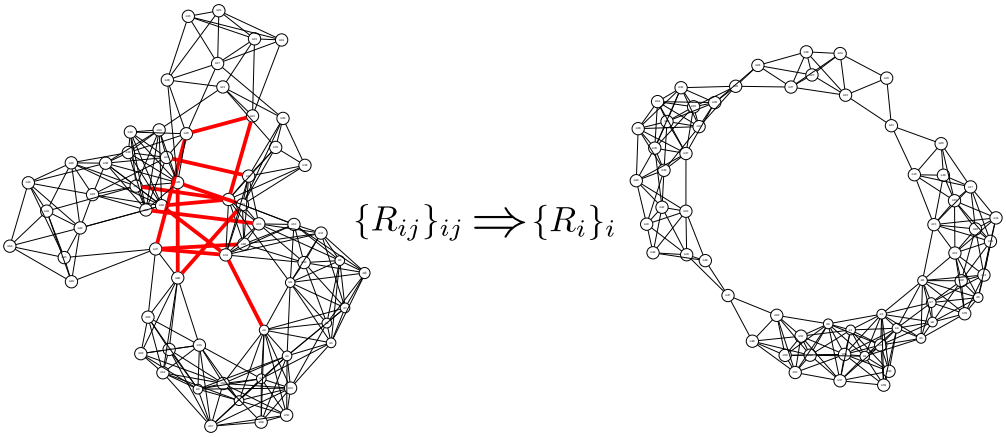


Procédure 8 Calcul des correspondances épipolaires géométriquement valides par géométrie essentielle

Entrée: Une collection d'images $\{I_i\}_i$, les paramètres intrinsèques $\{\mathbf{K}_i\}_i \, | i \in (1,...,n)$
Sortie: Un graphe de correspondances valides par géométrie fondamentale : $G_{\mathcal{R}_{ij}}$

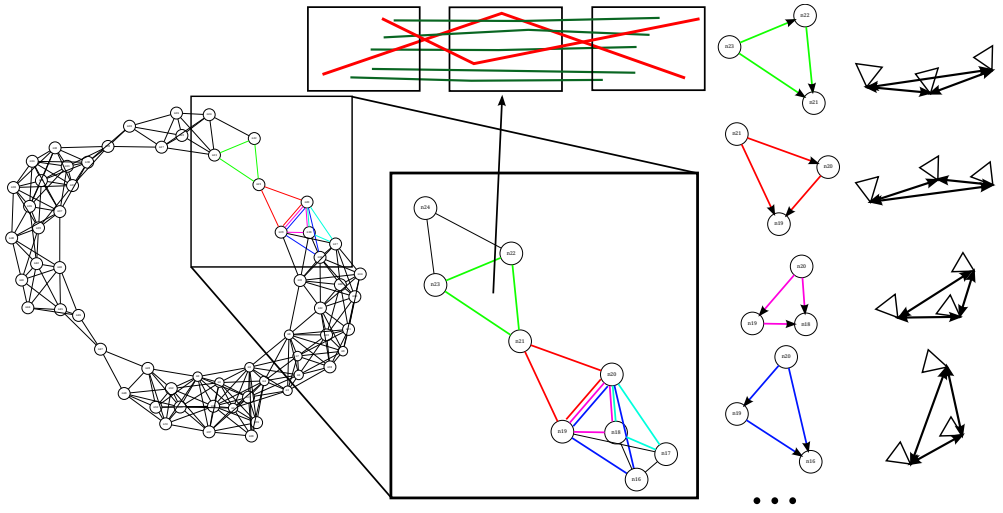
$G_{\mathcal{R}_{ij}} = \emptyset$
(1) **Calcul de la description locale des images :**
pour Toute image $i \in (1,...,n)$ **faire**
Détection et description des points saillants de l'image I_i
fin pour
(2) **Calcul des correspondances photométriques puis géométriques :**
pour Toutes paires d'images $(i,j) \in (1,...,n)$ **faire**
Calcul des correspondances photométriques entre I_i et I_j
$*$ Estimation robuste de la matrice essentielle $\mathbf{E}$, calcul de $(R_{ij}, \mathbf{t}_{ij})$
si Estimation réussie **alors**
Ajout de la paire (i,j) au graphe $G_{\mathcal{R}_{ij}}$ et R_{ij} en propriété de l'arête e_{ij}
fin si
fin pour

Étape 2 : Vérification de la cohérence des rotations relatives et calcul des rotations globales (cf. section 6.2.1) pour supprimer les relations relatives aberrantes. Nous utilisons de manière itérée l'inférence bayésienne de ZACH et al. (2010) jusqu'à stabilisation sur le graphe $G_{\mathcal{R}_{ij}}$. La cohérence des cycles de longueur 3 (les triplets) est vérifiée (cf. 6.2.1) puis les rotations globales R_i sont estimées. Un nouveau graphe $G_{\mathcal{R}}$ sans les rotations relatives aberrantes est obtenu.

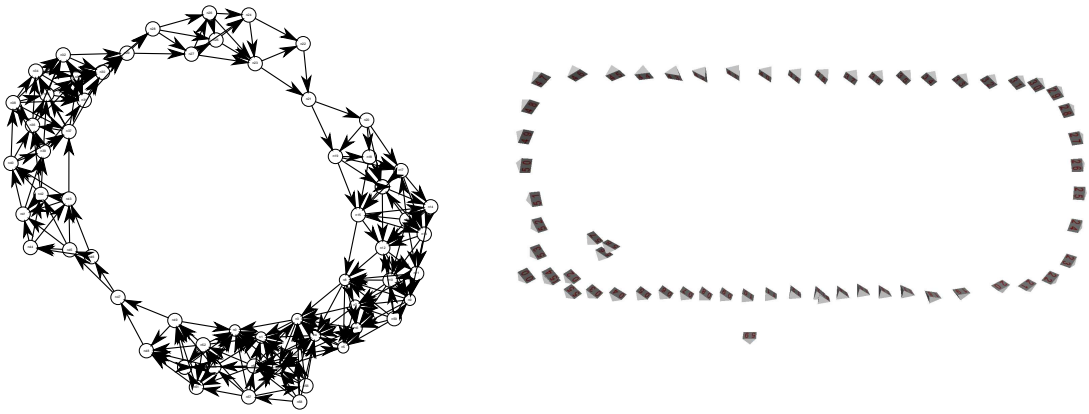


Étape 3 : Calcul des translations relatives par tenseurs tri-focaux réduits (cf. 6.2.2). Les translations relatives recouvrant le graphe $G_{\mathcal{R}}$ sont calculées. Plutôt qu'utiliser une approche naïve consistant à estimer tous les triplets du graphe $G_{\mathcal{R}}$, ayant une complexité linéaire en fonction du nombre de triplets du graphe, nous utilisons une méthode ayant une complexité sous-linéaire basée sur une procédure d'arêtes couvrantes (cf. 6.3.1). Un nouveau graphe $G_{\mathbf{t}^{\boldsymbol{\tau}}_{ij}}$ est obtenu où pour chaque triplet les translations relatives $\{\mathbf{t}^{\boldsymbol{\tau}}_{ij}\}_{ij}$ et le consensus de points validés *a contrario*, les traces $tr^{\boldsymbol{\tau}}$, sont conservées.

Le fait de considérer les traces à l'échelle des triplets procure plusieurs avantages :
- On évite de calculer les traces sur l'ensemble des correspondances du graphe $G_{\mathcal{R}}$.
- De nombreux petits problèmes de calcul des traces sont plus rapides à calculer qu'une fusion sur l'ensemble des correspondances du graphe.
- Considérer les traces à échelle locale plutôt que globale évite le rejet de traces qui seraient en conflit de correspondance épipolaire à échelle globale.
- Utiliser les traces locales permet de filtrer efficacement les incohérences locales qui seront fusionnée une fois les translations estimées.



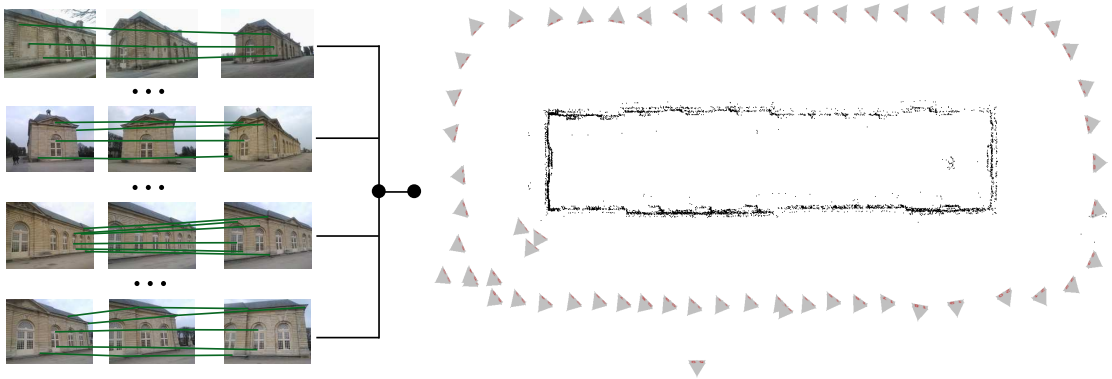
Étape 4 : Calcul des translations globales par fusion des translations relatives (cf. section 6.2.3). Les translations relatives sont fusionnées dans un repère commun global sous la norme l_∞ par optimisation convexe. Des poses de caméras sont déterminées sans avoir calculé la structure globale.



Étape 5 : Calcul de la structure et ajustements de faisceaux. Des étapes précédentes nous obtenons une solution initiale approchée des positions et orientations des caméras, ainsi que les traces (structures) cohérentes par triplet. Les traces locales validées par triplet τ sont reliées entre elles (calcul de traces de traces) puis les points en correspondances sont triangulés pour former la structure globale de la scène. Étant donné que les structures locales sont cohérentes par géométrie tri-focale on peut penser que la structure globale est quasi-exempte de correspondances aberrantes, ce que nous avons pu vérifier dans nos expériences. Bien que l'on observe expérimentalement que les mouvements de caméras et la structure initiale sont déjà d'excellente qualité à cette étape, nous réalisons une séquence d'ajustements de faisceaux pour équi-répartir les erreurs restantes au sein du graphe.

Nous optimisons tout d'abord la structure et les translations de caméras, puis les rotations, les translations et la structure. Réaliser cet ajustement de faisceaux en deux étapes, d'abord avec des rotations fixées, puis sur toutes les variables, est inspiré par les travaux d'OLSSON et ENQVIST (2011). Les variables sont relâchées en fonction de leur confiance. L'idée sous-jacente est d'éviter de compenser les erreurs d'estimation de translations par l'ajustement de rotations, étant donné que les rotations sont supposées stables et qu'elles ont servi pour estimer les translations.

On observe expérimentalement que les étapes d'ajustements de faisceaux convergent en un nombre très faible d'itérations ce qui atteste de la précision de nos estimations initiales.



On obtient la procédure 9 de calibration globale externe suivante :

Procédure 9 Calibration globale externe d'un réseau de caméra :

Entrée: Un graphe $G_{\mathcal{R}_{ij}}$, les paramètres intrinsèques $\{\mathbf{K}_i\}_i | i \in \{0,...,n\}$
Sortie: Une reconstruction 3D $\{X_{tr}\}$ et des poses de caméras $\{(R_i, T_i)\}_i$

(1) **Consistances des rotations relatives et calcul des rotations globales :**
$G_{\mathcal{R}_{ij}} = \text{InférenceRotationsRelatives}(G_{\mathcal{R}_{ij}})$
$(G_{\mathcal{R}_{ij}}, \{R_{ij}\}_{ij}) = \text{VérificationGraphe}(G_{\mathcal{R}_{ij}})$
$\{R_i\}_i = \text{RotationsGlobales}(\{R_{ij}\}_{ij})$

(2) **Estimation *a contrario* des translations relatives par tenseur tri-focaux réduit :**
$(G_{\mathbf{t}^\tau_{ij}}, \{\mathbf{t}^\tau_{ij}\}_\tau, \{tr^\tau\}_\tau) = \text{TranslationsRelativesParTriplet}(G_{\mathcal{R}_{ij}}, (\{R_i\}, \{\mathbf{K}_i\})_i)$
$(G_{\mathbf{t}^\tau_{ij}}, \{\mathbf{t}^\tau_{ij}\}_\tau, \{tr^\tau\}_\tau) = \text{VérificationGraphe}(G_{\mathbf{t}^\tau_{ij}})$

(3) **Calcul des translations globales par fusion des translations relatives :**
$\{T_i\}_i = \text{FusionTranslationsRelatives}(\{\mathbf{t}^\tau_{ij}\}_\tau)$

(4) **Calcul de la structure et raffinement non linéaire :**
$\{tr\} = \text{Traces}(\{tr^\tau\}_\tau)$ // Fusion des traces valides par triplet
$\{X_{tr}\} = \text{Triangulation}((\{R_i\}, \{T_i\})_i, \{tr\})$ // Triangulation initiale

$\{T_i\}_i, \{X_{tr}\} = \text{Ajustement de faisceaux}(\{T_i\}_i, \{X_{tr}\}) \,|\{R_i\}_i$
$(\{R_i\}, \{T_i\})_i, \{X_{tr}\} = \text{Ajustement de faisceaux}((\{R_i\}, \{T_i\})_i, \{X_{tr}\})$

6.3.1 Optimisation pour le passage à l'échelle

On notera que plusieurs étapes sont aisément parallélisables dans notre approche :

- le calcul des rotations relatives R_{ij},
- le calcul des triplets : les translations relatives $\mathbf{t}^\tau_{ij}$,
- le calcul de la structure 3D,
- les étapes de chaque itération d'ajustement de faisceaux.

Nous présentons quelques détails d'implémentation et détaillons la condition nécessaire pour qu'un graphe soit compatible pour la calibration globale par fusion de mouvements.

Évaluation rapide des translations relatives d'un réseau de caméras :

Une solution naïve pour calculer les translations relatives du réseau de caméra consiste à lister les triplets et tous les estimer. Cependant on va couvrir ainsi plusieurs fois les mêmes arêtes du graphe $G_{\mathcal{R}}$ inutilement.

Afin d'estimer les translations relatives recouvrant le graphe $G_{\mathcal{R}}$, nous utilisons la procédure 10 d'arêtes couvrantes. L'algorithme est le suivant. Pour chaque arête les triplets sont listés et ordonnés par nombre de traces décroissantes. Pour chaque arête on itère sur les triplets qui la contiennent (par nombres de traces décroissantes) en cherchant à estimer les translations relatives. Tant qu'une estimation robuste échoue, on continue d'itérer sur les triplets. Si tous les triplets pour cette arête échouent, on supprime l'arête du réseau de caméras : ses correspondances de points sont jugées incohérentes. A chaque triplet estimé avec succès, ses arêtes sont marquées comme estimées ; de une à trois arêtes sont marquées couvertes. Tant que le réseau de caméras n'a pas toutes ses arêtes couvertes, on itère. Cette approche permet de calculer les translations relatives et des structures locales couvrant le réseau de caméras avec une complexité sous-linéaire du nombre d'arêtes, de plus cet algorithme est facilement parallélisable.

Procédure 10 Calcul des translations relatives (basées triplets) d'un réseau de caméra :

Entrée: Un graphe $G = (\{I_i\}_{i\in 0,\dots,n}, \{e_{ij}\}_{ij})$
Sortie: $G_{\mathbf{t}^\tau_{ij}} = (\{I_i\}_{i\in 0,\dots,}, \{\mathbf{t}^\tau_{ij}\}_{ij})$: les translations relatives des triplets τ validés *a contrario*

(1) **Initialisation :**
$\{\tau_g\}_g$ = ListeTriplets(G) // chaque τ_g est un ensemble de 3 arêtes
(2) **Estimation des translations relatives par couverture des arêtes :**
tant que edge $e_{ij} \in \{e_{ij}\}_{ij}$ **faire**
(3) **Estimation de la translation de l'arête :**
$\{\tau_c\}_c = \{\tau_g : e_{ij} \in \tau_g\}$ // Récupération des triplets contenant l'arête e_{ij}
$[\tau_c]_c$ = Tri des triplets $\{\tau_c\}_c$ par nombre de traces décroissant
pour $\tau_c \in [\tau_c]_c$ **faire**
Suppression de e_{ij} dans $\{e_{ij}\}_{ij}$
Suppression de τ_c dans $\{\tau_g\}_g$
si $\{\mathbf{t}^\tau\} = AC\text{-}RANSAC(\tau_c) \neq \emptyset$ **alors**
(4) **Estimation du triplet réussie, on marque de 1 à 3 nouvelles arêtes :**
Ajout de $\{\mathbf{t}^\tau\}$ à $G_{\mathbf{t}^\tau_{ij}}$ // 3 translations relatives sont ajoutées
Suppression des $e_{ij} \in \tau_c$ dans $\{e_{ij}\}_{ij}$
Retour en (2)
fin si
fin pour
fin tant que

Filtrage d'un graphe de mouvements relatifs pour l'estimation de mouvements globaux

Nous avons vu que les étapes de notre chaîne de calibration globale passent par l'utilisation de plusieurs graphes et la suppression d'arêtes aberrantes : $G_{\mathcal{R}_{ij}}$, $G_{\mathcal{R}}$ et $G_{\mathbf{t}^\tau_{ij}}$. Afin que ces graphes soit utilisables entre les différentes étapes, nous vérifions certaines conditions : à chaque étape le graphe doit présenter une seule composante connexe et chaque sommet doit être accessible par deux chemins différents (cf. figure 6.16), c'est une condition nécessaire et suffisante pour qu'un graphe soit composé de cycles et donc permettre l'estimation moyenne d'un mouvement global en ces sommets.

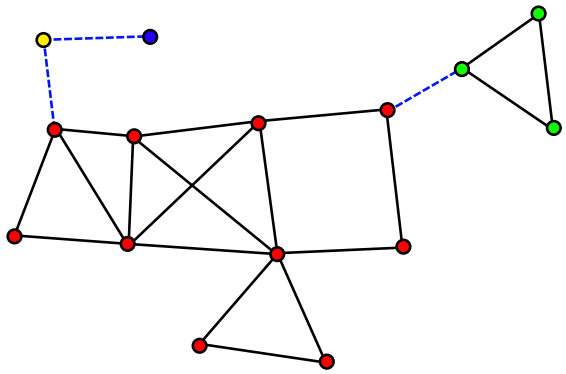


FIGURE 6.16 – Vérification d'un graphe pour la calibration globale par fusion de mouvements. Deux sous-graphes connexes sont utilisables pour l'estimation globale, une fois les bi-arêtes retirées (le graphe comportant des sommets rouges et le graphe contenant les sommets verts). Les autres sommets du graphe ne peuvent être utilisés pour de la calibration globale car leur connexité au graphe est assurée par un seul chemin.

Nous identifions les sous-graphes compatibles avec ces conditions, de manière séquentielle avec la procédure 11, en identifiant les composantes connexes après suppression des arêtes dites *bi-edges*. Les *bi-edges* sont des composantes de classes identifiées après utilisation d'une relation d'équivalence sur un graphe : deux sommets sont dans une même classe s'ils sont connectés par deux arêtes disjointes. Les arêtes reliant deux classes différentes sont les *bi-edges* recherchées.

Procédure 11 Identification du graphe composé de cycles comportant le plus de sommets

Entrée: Un graphe G
Sortie: G_c le plus grand sous-graphe composé de cycle

(1) **Identification et suppression des *bi-edges* :**
$\{e_{ij}\}_{ij}$ = ListeBiedges(G)
pour Toute edge $e_{ij} \in \{e_{ij}\}_{ij}$ **faire**
Suppression de e_{ij} dans G
fin pour

(2) **Identification du sous-graphe le plus large :**
$\{G_i\}_i$ = ListeComposantesConnexes(G)
$G_c = \underset{\{G_i\}_i}{\text{argmax}} \#Sommets(G_i)$

Lorsque plusieurs graphes sont présents, chaque sous-graphe peut être calibré de manière séparée puis les reconstructions peuvent être fusionnées. La fusion des reconstructions ou images est réalisée en fonction de la connexité au graphe considéré comme amer :

- **cas d'une arête seule :** l'image est ajoutée par estimation robuste de pose,
- **cas d'un graphe :** un calcul de rotation translation et échelle entre les points 3D de la structure commune permet de fusionner les reconstructions.

6.4 Résultats et évaluations

A travers différentes expérimentations sur plusieurs jeux de données, nous allons étudier les points suivants :

1. La sensibilité au bruit d'estimation parmi les rotations globales,
2. La précision et le temps de calcul de notre chaîne de calibration externe globale,
3. La capacité de notre chaîne à traiter des jeux de données présentant un grand nombre de fausses géométries relatives.

Sensibilité aux bruits d'estimation des rotations globales.

Afin d'illustrer l'effet d'estimations de rotations globales bruitées sur notre chaîne de calibration globale nous utilisons le jeu de données FountainP11 (STRECHA et al. 2008). Les rotations vérités terrain sont bruitées artificiellement par des petites rotations tirées uniformément sur une sphère entre 0° et α°. La figure 6.17 montre que l'erreur de reconstruction est quasi-linéaire en fonction du bruit des rotations, pour un bruit α variant de 0° à 1°. Des résultats similaires ont été observés par SIM et HARTLEY (2006).

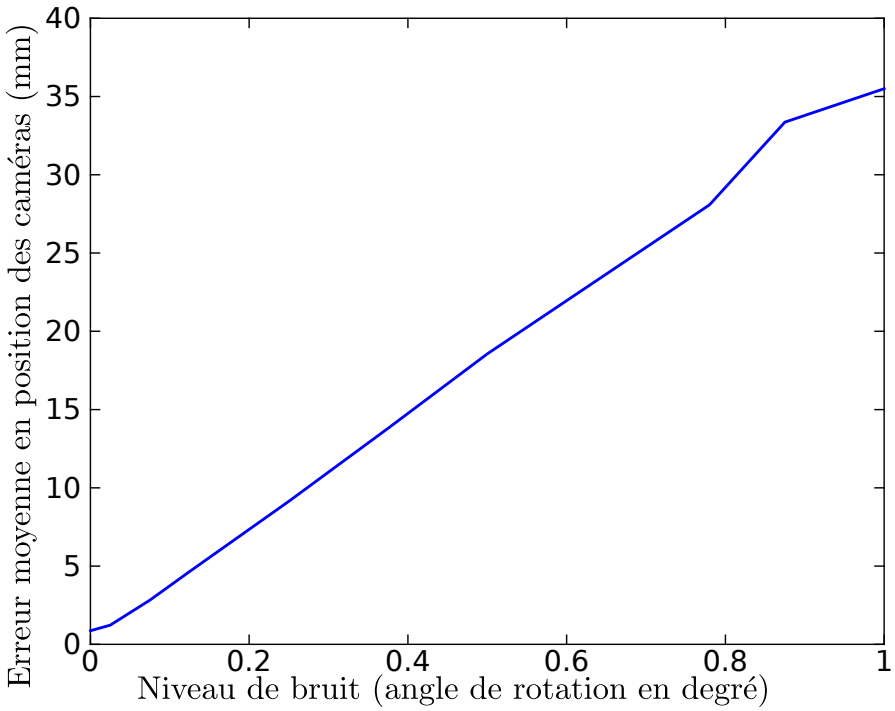


FIGURE 6.17 – Évaluation de la précision d'estimation moyenne de la pose des caméras, en mm, calculée en fonction du bruit introduit sur les rotations globales (moyenne des résultats obtenus sur 10 expériences pour chaque valeur de α évaluée).

Évaluation de la précision de notre chaîne de calibration externe globale

Le même test d'évaluation de la précision moyenne de l'estimation de localisation des caméras qu'en section 5.4 est réalisé pour la calibration globale. La qualité d'estimation moyenne des paramètres extrinsèques est calculée par une transformation rigide de degré 7 (échelle, rotation, translation) entre la vérité terrain et les positions de caméras estimées, pour différents jeux de données (cf. tableau 6.4). Les méthodes comparées sont :

Les méthodes globales :

notre méthode *a contrario* globale : *GlobalAC*,

la méthode globale de OLSSON et ENQVIST (2011) (code Matlab),

la méthode globale de ARIE-NACHIMSON et al. (2012) (résultats mentionnés dans leur publication).

Les méthodes séquentielles :

notre méthode *a contrario AC-SfM*,

Bundler : SNAVELY et al. (2006),

VisualSfM : WU (2013).

On remarque que les méthodes globales présentent toutes de meilleures précisions que les méthodes séquentielles grâce à une meilleure répartition des erreurs et à leur capacité à mieux rejeter les données aberrantes. On constate que pour la majorité des jeux de données notre méthode permet l'obtention d'une précision au moins équivalente à celle d'OLSSON, voire meilleure. Principalement pour les séquences Castle, on note une nette amélioration des résultats. Cela est dû au fait que notre méthode permet de rejeter plus efficacement les fausses géométries épipolaires et les mauvaises correspondances de points. OLSSON et ENQVIST (2011) précisent que des points aberrants

existent toujours après la résolution de la structure et des translations, ce qui pénalise la reconstruction. Ils ajoutent que la détection des mesures qui posent problème n'est pas facile à faire et que la solution qu'ils proposent n'est pas idéale sur ce plan.

		Méthodes globales			Méthodes séquentielles		
Scène		*GlobalAC*	Olsson	Arie	*AC-SfM*	Bundler	VisualSfM
entryP10	erreur	**5.96**	6.9	x	16.2	55.1	63.0
	rang	**1**	2	x	3	4	5
FountainP11	erreur	2.58	**2.2**	4.8	2.86	7.03	7.64
	rang	2	**1**	3	4	5	6
HerzJesusP8	erreur	**3.56**	3.9	x	6.71	16.4	19.3
	rang	**1**	2	x	3	4	5
HerzJesusP25	erreur	**5.36**	5.7	7.8	8.85	21.5	22.4
	rang	**1**	2	3	4	5	6
CastleP19	erreur	**25.6**	76.2	x	223	344	258
	rang	**1**	2	x	3	4	5
CastleP30	erreur	**21.9**	66.8	x	69.1	300	522
	rang	**1**	2	x	3	4	5
DTU-Set001	erreur	**0.32**	27.8	x	0.71	0.72	1.22
	rang	**1**	5	x	2	3	4
DTU-Set004	erreur	**0.45**	27.8	x	0.68	0.49	0.70
	rang	**1**	5	x	3	2	4
HouseDataset	erreur	**3.31**	8.94	x	11.6	11.7	211
	rang	**1**	2	x	3	4	5

TABLE 6.4 – Évaluation des erreurs de localisation moyenne des caméras en millimètres par rapport à la vérité terrain, pour six méthodes de calibration : trois méthodes globales et trois méthodes séquentielles. En gras la solution ayant la meilleure précision.

Évaluation des temps de calcul

La table 6.5 rassemble les temps de calcul pour différents jeux de données. La valeur mesurée correspond au temps nécessaire pour estimer les positions de caméras et la structure. Notre implémentation parallèle de notre chaîne *GlobalAC* sera notée *GlobalACP*. Le temps nécessaire pour calculer le graphe épipolaire n'est pas pris en compte.

On remarque que notre approche est bien plus rapide que la méthode globale d' OLSSON et ENQVIST (2011) sur l'ensemble des jeux de données testés. Notre approche est de 5 à 11 fois plus rapide pour sa version classique, et de 16 à 26 fois plus rapide lorsque l'on utilise la version parallélisée. Aucun mesure n'est précisée pour la méthode d'ARIE-NACHIMSON et al. (2012) car aucune implémentation n'est disponible.

	Méthodes globales			Méthodes séquentielles		Ratio de temps	
	GlobalAC (s)	*GlobalACP*	Olsson	Bundler	VisualSfM	Olsson /*GlobalAC*	Olsson /*GlobalACP*
HerzJesusP8	6	**2**	34	10	**2**	5.6	17
EntryP10	16	5	88	16	**3**	5.5	17
FountainP11	12	5	133	36	**3**	11.1	26
CastleP19	20	**6**	99	78	9	4.9	16
HerzJesusP25	47	**10**	221	100	12	4.7	22
CastleP30	55	**14**	317	300	18	5.7	22

TABLE 6.5 – Mesure des temps de calcul en secondes pour les différentes chaînes de calibration. La dernière colonne indique combien de fois notre méthode est plus rapide que celle d'Olsson et Enqvist. Les mesures en gras indiquent les méthodes les plus rapides.

Capacité de notre chaîne à traiter des jeux de données présentant un large nombre de fausses géométries relatives.

Afin d'obtenir un rendu visuel de l'impact des fausses géométries relatives et du filtrage de l'inférence bayésienne, nous avons fait les expériences suivantes. Des images sont acquises en tournant autour de bâtiments présentant de nombreux éléments répétitifs (façades similaires possédant des symétries axiales). Une fois les images mises en correspondances, un graphe épipolaire $G_{\mathcal{E}}$ est obtenu. On observe qu'au lieu d'obtenir un «beau» graphe en forme d'anneau on obtient un amas ou une forme de huit : de fausses géométries épipolaires apparaissent, qui relient à tort des éléments géométriques différents. Une fois les arêtes aberrantes détectées par l'inférence bayésienne, on note que le graphe $G_{\mathcal{R}}$ présente maintenant une «belle» boucle principale retraçant le trajet d'acquisition : les rotations relatives aberrantes ont bien été détectées et supprimées.

La figure 6.18 présente un cas d'échec pour le logiciel Bundler. Le graphe épipolaire étant trop contaminé par de fausses correspondances, des images sont positionnées aux mauvais endroits ou bien rejetées.

La figure 6.22 présente un cas à plus large échelle (160 images). On observe pour ce jeu de données des différences notables en temps de calcul. Bundler positionne les images dans l'espace en 3 heures (temps de calcul du graphe épipolaire décompté) tandis que nous estimons la pose des caméras en 7 minutes. Bundler passe en fait beaucoup de temps dans les nombreuses étapes d'ajustement de faisceaux. Au début, les itérations prennent quelques minutes, puis à la fin, chaque itération nécessite plus de 10 minutes.

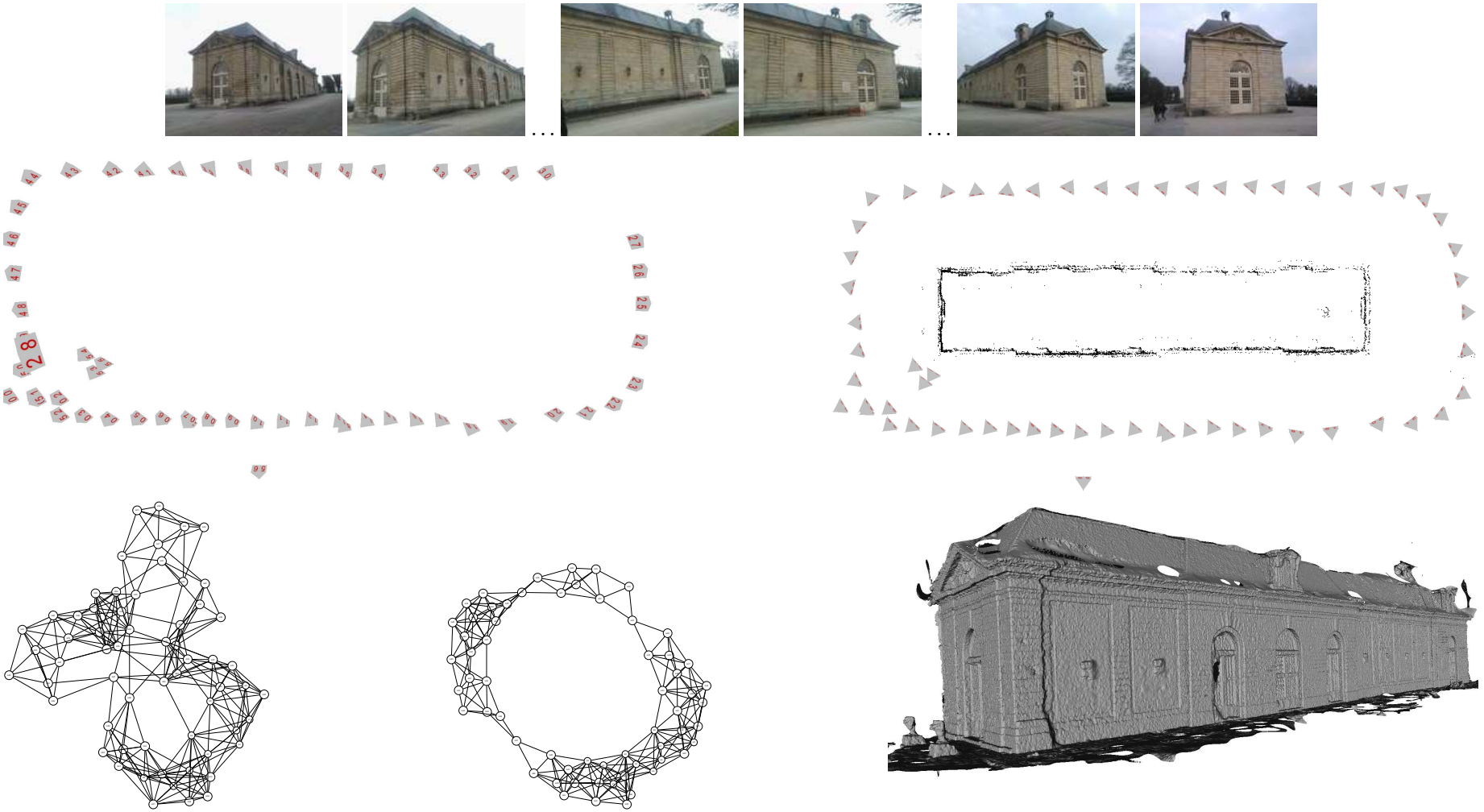


FIGURE 6.18 – OrangerieP61 : Echec de fermeture de boucle et mauvais positionnement avec Bundler. En haut : des images du jeux de données. Au milieu : à gauche les positions de caméras estimées par Bundler et à droite notre résultat. En bas : le graphe épipolaire $G_{\mathcal{E}}$ avant puis après inférence ($G_{\mathcal{R}}$), et le maillage reconstruit à partir de notre calibration externe.

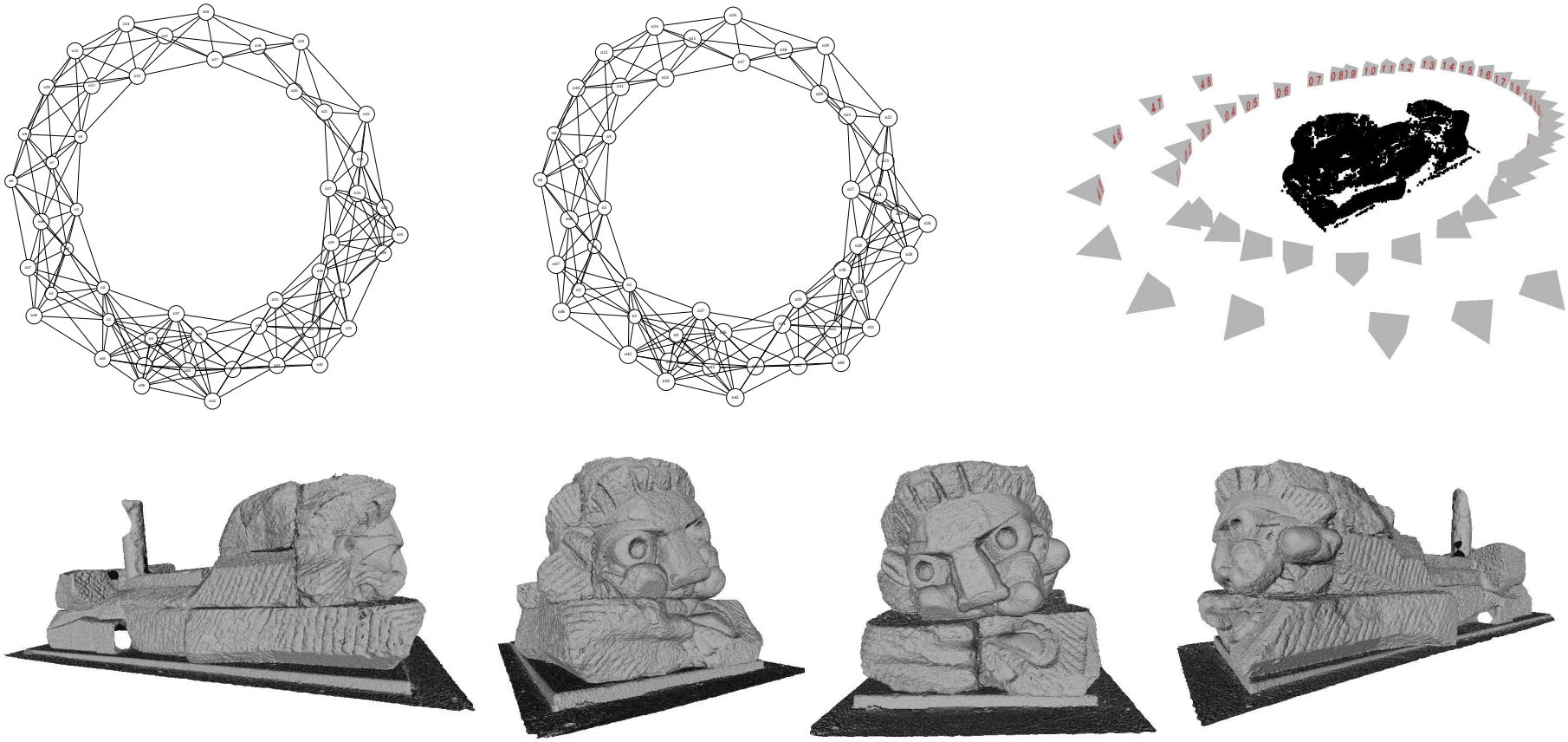


FIGURE 6.19 – MayaHeadP50. En haut , de gauche à droite : le graphe épipolaire $G_{\mathcal{E}}$ avant puis après inférence ($G_{\mathcal{R}}$). La structure et pose des caméras obtenues après calibration externe. En bas : des images du maillage reconstruit à partir de notre calibration externe.

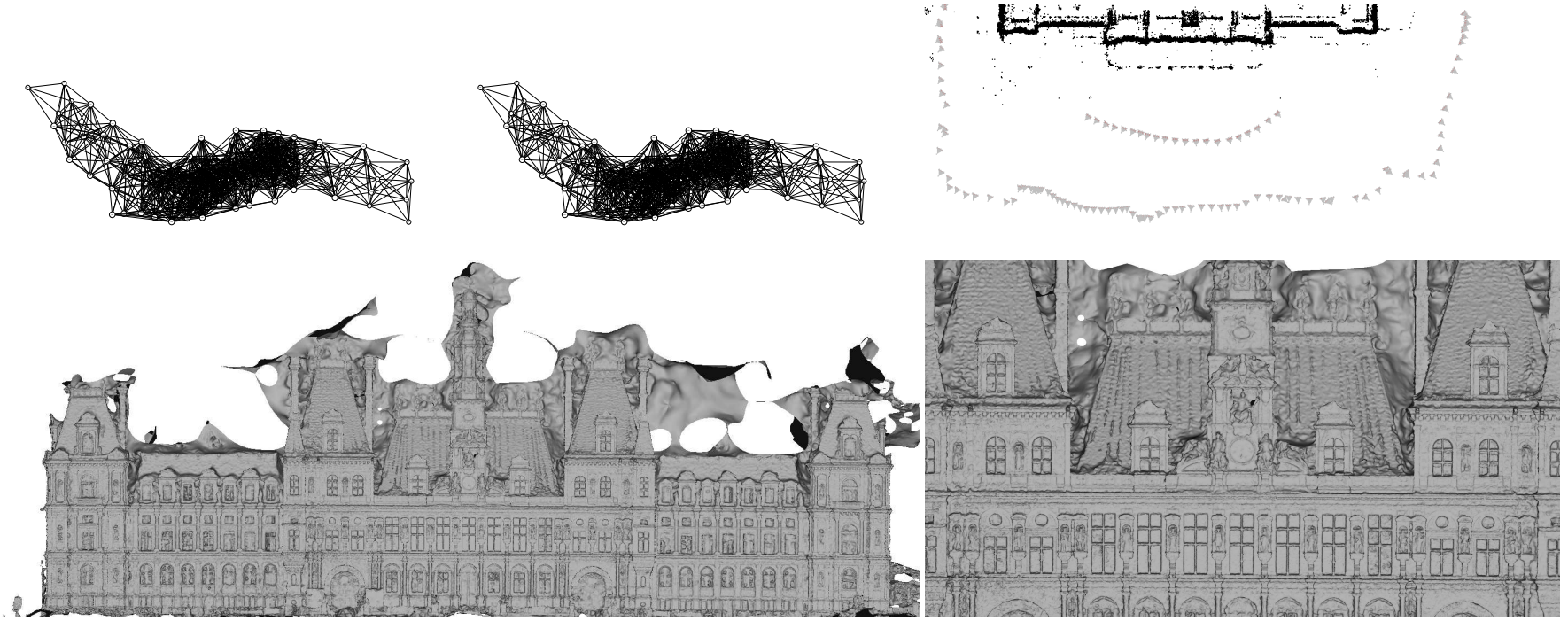


FIGURE 6.20 – HotelP100. En haut , de gauche à droite : le graphe épipolaire ($G_{\mathcal{E}}$) avant puis après inférence ($G_{\mathcal{R}}$). La structure et pose des caméras obtenues après calibration externe. En bas : des images du maillage reconstruit à partir de notre calibration externe de l'hôtel de ville de Paris.

FIGURE 6.21 – PanthéonP126. En haut , de gauche à droite : le graphe épipolaire ($G_{\mathcal{E}}$) avant puis après inférence ($G_{\mathcal{R}}$). La structure et poses des caméras obtenues après calibration externe. En bas : des images du maillage reconstruit à partir de notre calibration externe du Panthéon de Paris.

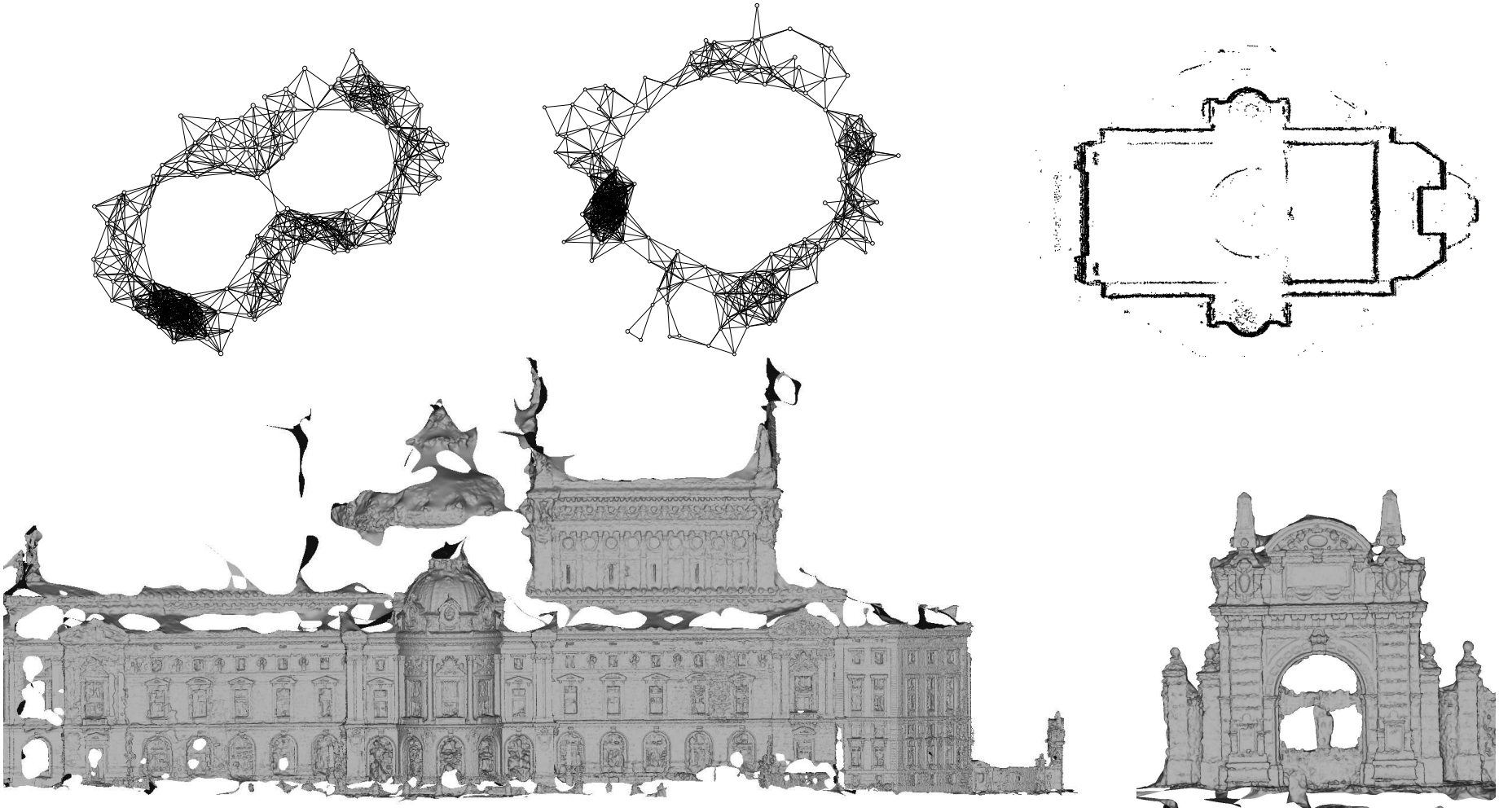


FIGURE 6.22 – OperaP160. En haut , de gauche à droite : le graphe épipolaire ($G_{\mathcal{E}}$) avant puis après inférence ($G_{\mathcal{R}}$). La structure obtenue après calibration externe (on ne relève aucune donnée aberrante). En bas : des images du maillage reconstruit à partir de notre calibration externe de l'opéra Garnier de Paris.

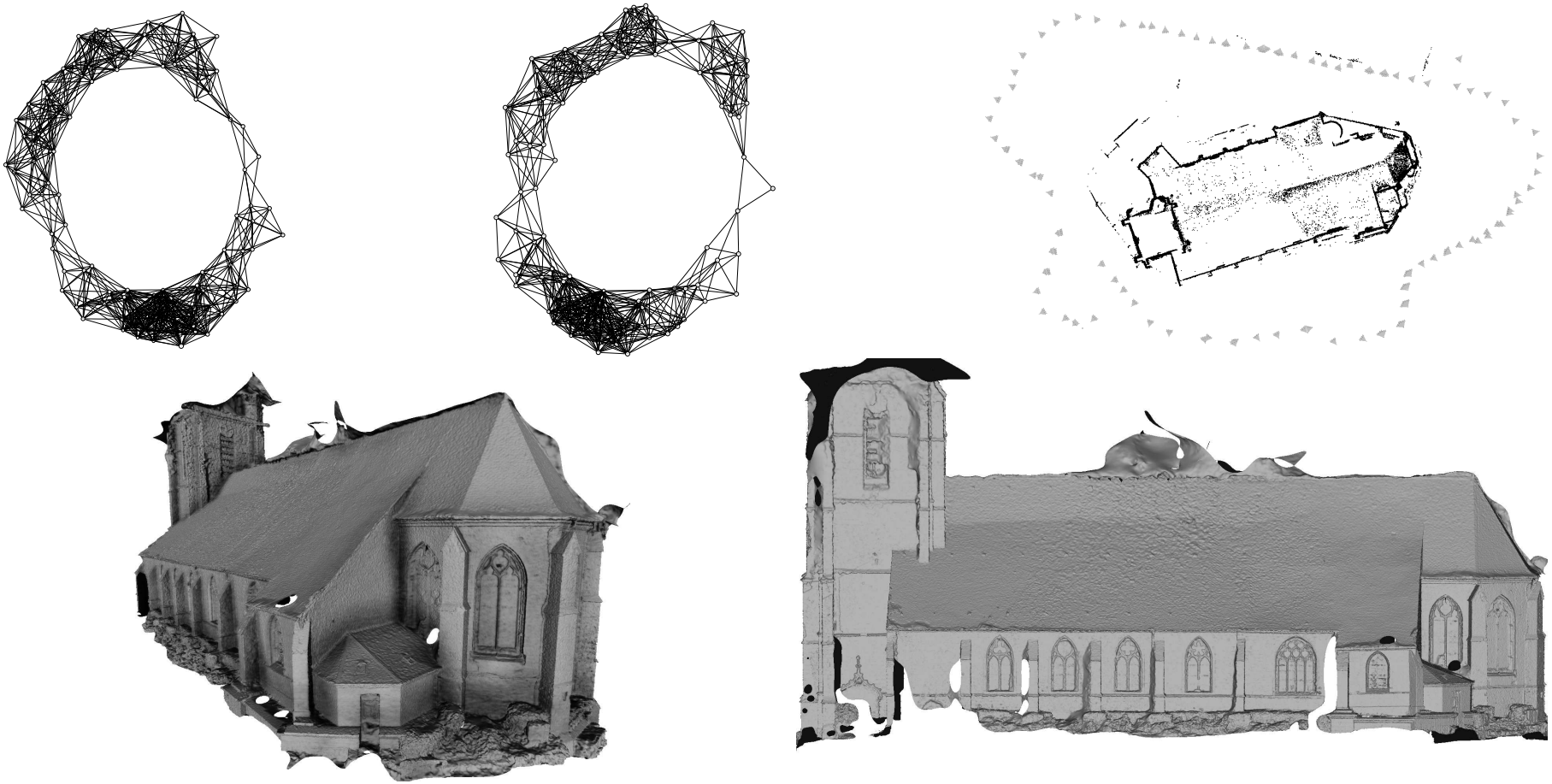


FIGURE 6.23 – AultP106. En haut , de gauche à droite : le graphe épipolaire ($G_{\mathcal{E}}$) avant puis après inférence ($G_{\mathcal{R}}$). La structure et pose des caméras obtenues après calibration externe. En bas : des images du maillage reconstruit à partir de notre calibration externe de l'église d'Ault. Temps de calibration (méthode globale : 20 min, méthode séquentielle : 3.5 h).

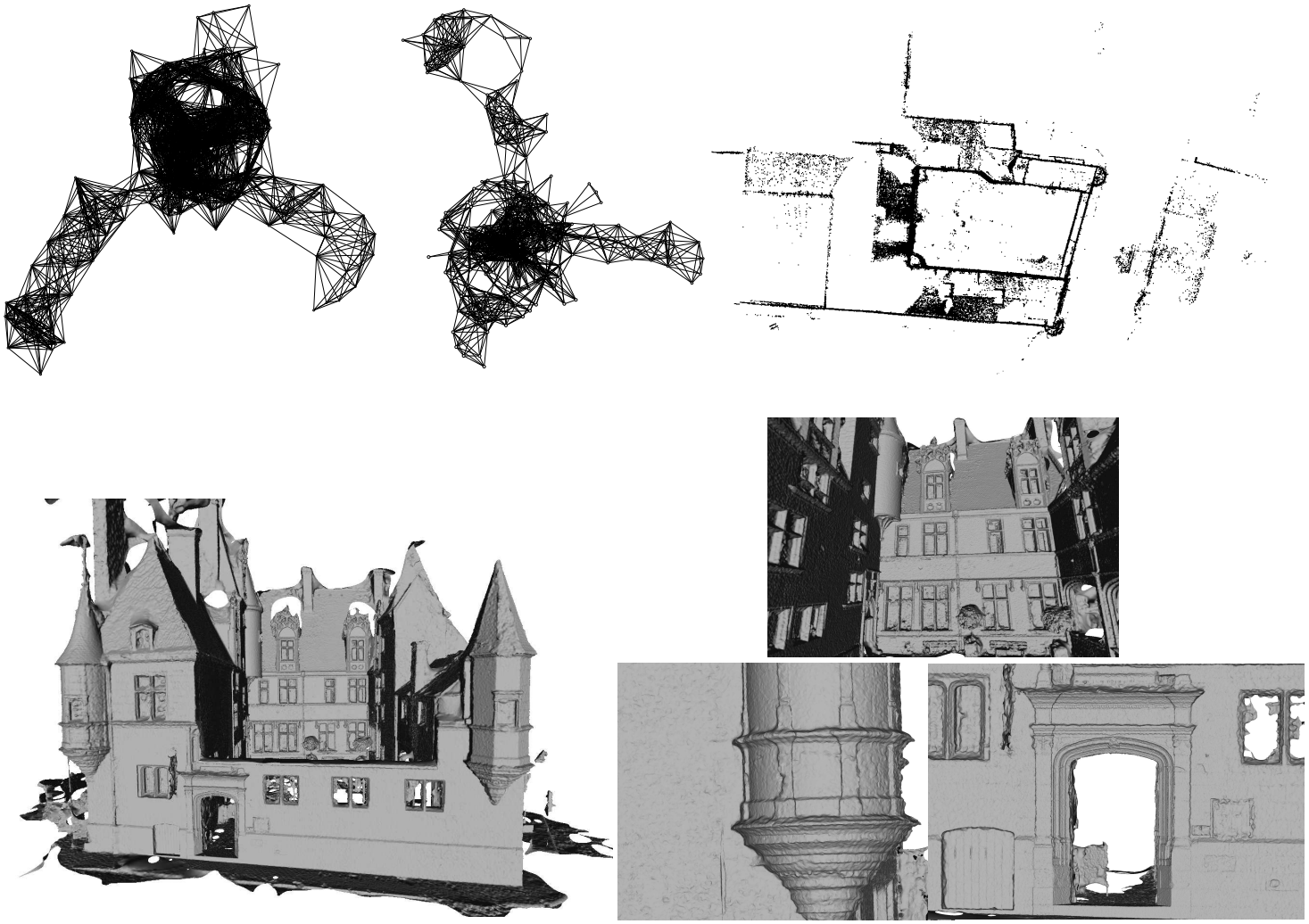


FIGURE 6.24 – HotelCujasP182. En haut , de gauche à droite : le graphe épipolaire ($G_{\mathcal{E}}$) avant puis après inférence ($G_{\mathcal{R}}$). La structure obtenue après calibration externe. En bas : Vues avec recul et sur des détails du bâtiment sur le maillage reconstruit de l'hôtel Cujas (Bourges) à partir de notre calibration externe. Temps de calibration (méthode globale : 15 min, méthode séquentielle : 2 h).

FIGURE 6.25 – PavillonP421. En haut , de gauche à droite : le graphe épipolaire ($G_{\mathcal{E}}$) avant puis après inférence ($G_{\mathcal{R}}$). La structure obtenue après calibration externe. En bas : images du maillage reconstruit à partir de notre calibration externe.

Analyse détaillée des étapes de la chaîne.

Nous présentons dans le tableau 6.6 des statistiques liées aux différentes étapes de la chaîne pour différents jeux de données allant de 8 à 421 images. Les informations présentées concernent :

- les triplets du réseau de caméra,
- les statistiques liées à la fusion des translations,
- les statistiques liées aux différentes étapes d'ajustement de faisceaux (BA),
- le nombre de points 3D reconstruit,
- le temps requis pour la reconstruction (temps de construction du graphe épipolaire exclu).

Commentaires sur l'estimation des triplets.
On note que le nombre de triplets résolu est bien plus faible que le nombre de triplets total du réseau de caméras. Cela montre que notre approche par couverture des arêtes permet de ne pas calculer tous les triplets du graphe et donc d'effectuer de manière efficace l'estimation des translations relatives couvrant le graphe.

Commentaires sur l'estimation avant ajustement de faisceaux.
On observe une relation directe entre le paramètre γ et la précision des points 3D initiaux triangulés. La variable γ identifiée indique dans quelle mesure les translations relatives s'accordent avec les translations globales, il est donc tout à fait cohérent d'observer une corrélation entre γ et les erreurs moyennes résiduelles de la structure globale initiale triangulée : $\bar{\rho}_{init}$.

Les faibles valeurs de γ confirment la précision des translations relatives calculées et la faible valeur des erreurs résiduelles moyennes confirment le fait que sans même avoir considéré la structure globale notre chaîne globale estime des poses de caméras précises.

Commentaires sur l'estimation des ajustements de faisceaux.
Après les étapes d'ajustements de faisceaux l'ensemble des scènes testées sont reconstruites avec une erreur moyenne résiduelle inférieure au demi-pixel et ce pour des structures comportant de nombreux points 3D et caméras. Ces résultats numériques attestent de la qualité des reconstructions obtenues (qualité visible à travers les détails que l'on peut observer sur les maillages calculés à partir de nos calibrations sur les figures précédentes).

Comme nous utilisons des minimisations l_∞ nous identifions une solution ou la plus large des erreurs est minimale pour le jeu de contraintes estimé. C'est pourquoi les ajustements de faisceaux convergent rapidement vers une solution optimale car seule une erreur de faible amplitude est à répartir sur l'ensemble du réseau de caméras. Ainsi on observe que le nombre d'itérations nécessaire pour les deux étapes finales d'ajustement de faisceaux (BA) est très faible (le plus souvent ≤ 10). Cela confirme que notre chaîne de calibration externe utilisant des séries de minimisation d'erreur sous la norme l_∞, est très précise et proche d'une solution optimale.

Les reconstructions étant toutes valides, les résultats démontrent que nous avons rejeté avec succès les données aberrantes présentes dans le graphe épipolaire (les rotations relatives et correspondances de points homologues erronées). On remarque que nos structures sont très «propres», très peu, voir aucun points 3D ne sont placés à une mauvaise position.

Jeux de données		Triplets			Fusion des translations				BA_1		BA_2		#3D pts	Temps
Name	#Ima.	#résolu	#possible	temps(s)	$\#t_{ij}$	temps(s)	γ	$\bar{\rho}_{init}$	#iter	$\bar{\rho}$	#iter	$\bar{\rho}$	#Traces	total(s)
FountainP11	11	28	78	2	84	< 1	5×10^{-4}	0.75	2	0.26	3	0.25	17216	5
EntryP10	10	28	105	2	84	< 1	1×10^{-3}	1.23	2	0.32	2	0.26	19056	5
HerzJesusP8	8	13	25	1	39	< 1	3×10^{-4}	0.73	2	0.36	2	0.34	7561	2
HerzJesusP25	25	102	522	4	306	< 1	5×10^{-4}	0.85	2	0.47	4	0.46	25990	10
CastleP19	19	39	113	1	117	1	3×10^{-3}	2.9	2	0.54	3	0.26	19759	6
CastleP30	30	103	540	6	309	1	3×10^{-3}	2.3	2	0.51	3	0.27	38589	14
DTU-Set001	119	4947	263341	550	14841	160	1×10^{-3}	0.92	6	0.51	9	0.40	22074	3353
DTU-Set004	119	4107	178588	366	12321	105	1×10^{-3}	0.82	7	0.49	10	0.45	19845	6990
HouseDatasetP10	10	33	120	22	99	< 1	$3 \times 10^{-}5$	0.14	1	0.13	4	0.13	24843	62
OrangerieP61	61	141	382	7	423	3	1×10^{-2}	8.09	5	1.0	7	0.57	7730	18
MayaHeadP50	50	149	402	10	447	2	1×10^{-3}	1.92	3	0.42	4	0.37	34960	41
HotelP100	100	773	6655	57	2319	30	2×10^{-3}	1.98	3	0.59	4	0.51	56481	208
PanthéonP126	126	582	3383	95	1746	25	3×10^{-3}	4.32	4	0.60	6	0.41	24252	189
OperaP160	160	588	3054	30	1764	41	1×10^{-2}	5.47	5	1.05	10	0.48	53703	207
AultP106	106	597	4462	284	1791	2	2×10^{-2}	3.06	5	3.73	10	0.32	309893	1214
HotelCujasP182	182	840	4432	95	2520	46	2×10^{-2}	5.2	7	0.52	7	0.25	233142	939
PavillonP421	421	5612	85149	1177	16836	409	2×10^{-1}	6.94	10	3.38	8	0.51	619405	15617

TABLE 6.6 – Temps (s) sur une machine 8 coeurs à 2.67GHz. Erreur de re-projection moyenne $\bar{\rho}$ (pixels) pour les différentes étapes de la calibration externe globale. Les temps relevés pour les scènes DTU-Set001, DTU-Set004, AultP106 et PavillonP421 s'expliquent par le nombre très important de SIFT détectés dans les images.

6.5 Contributions de ce chapitre et perspectives

Dans ce chapitre nous avons présenté les différentes méthodes existantes pour traiter le problème de la calibration externe de manière globale. Nous avons ensuite décrit une chaîne de calibration globale pouvant être rapide, robuste et précise. Cette chaîne repose sur trois idées principales :

1. l'utilisation d'une primitive de base, le triplet,
2. l'analyse et le nettoyage de graphe pour la fusion robuste de mouvements relatifs,
3. l'estimation découplée rapide des translations et de la structure.

A travers les différentes sections qui précèdent nous avons montré l'amélioration qualitative et quantitative des points suivants :

- l'inférence bayésienne pour la détection des rotations relatives aberrantes,
- l'optimisation rapide, précise et adaptative de tenseurs tri-focaux réduits en utilisant la méthodologie *a contrario* et une estimation convexe de modèle paramétrique,
- la fusion de translations relatives dans un repère global commun par optimisation convexe,
- la construction d'une structure globale très peu bruitée par fusion de traces validées localement par triplets.

L'assemblage de ces éléments permet de réaliser une chaîne de calibration globale qui estime des poses de caméra avec une précision millimétrique à centimétrique. L'ensemble des jeux de données testés démontre que par rapport à l'autre solution globale évaluée (OLSSON et ENQVIST 2011) nous sommes de 5 à 26 fois plus rapides tout en estimant une meilleure solution. Notre approche permet un meilleur passage à l'échelle. Ainsi sans à avoir à choisir de paire initiale (le point problématique des reconstructions séquentielles), ni se soucier de la configuration a priori de paramètres de précision, notre chaîne conduit à une meilleure reconstruction sur 90% des scènes évaluées.

Ces travaux ont été acceptés pour publication à la conférence ICCV (MOULON et al. 2013b).

Perspectives. Notre chaîne comporte plusieurs points pouvant être développés dans le futur :

1. L'amélioration de l'inférence des rotations relatives,
2. L'utilisation de notre méthode dans un contexte semi-temps réel,
3. L'application de la fusion de mouvements relatifs à d'autres modèles de caméra,
4. L'amélioration de la fusion de translations relatives convexes.

1. L'amélioration de l'inférence de rotations relatives. Nous avons observé à travers l'expérience de la figure 6.17 que la qualité des résultats est dépendante, de manière linéaire, de la précision de l'estimation des rotations globales. Ainsi, si les mouvements relatifs erronés de rotations sont détectés avec une meilleure précision, les résultats ne seront qu'améliorés. Après avoir noté à travers des expériences, sur de larges collections d'images, que cette étape devient la plus lente de tous le processus, nous pensons pour cela que les travaux de CHATTERJEE et GOVINDU (2013) sont prometteurs.

2. L'utilisation de notre méthode dans un contexte semi-temps réel. Notre chaîne possède l'avantage de pouvoir estimer la trajectoire globale de la caméra sans considérer la structure globale de la scène (sauf à la toute dernière étape), avec une très bonne précision. Nous pouvons alors envisager d'utiliser notre méthode dans le but de guider un utilisateur pendant son acquisition : imaginons que les images acquises par un utilisateur sont envoyées au fur et à mesure à un ordinateur. Notre chaîne de calibration pourra alors en un temps de réponse extrêmement rapide montrer le graphe de calibration ainsi que les photos ayant pu être positionnées dans l'espace. Ces indices visuels 2D et 3D pourront guider l'utilisateur pendant l'acquisition pour vérifier et ainsi l'aider à réaliser ces photos afin que les fermetures de boucles soient effectives et que la reconstruction porte une qualité suffisante sur les éléments d'intérêt. Un post-traitement pourra ensuite être lancé une fois que les ressources de calculs seront plus élevées pour déterminer une reconstruction complète et garantie avec succès.

3. L'application de la fusion de mouvements relatifs à d'autres modèles de caméra. Notre chaîne ne réalise aucun apriori sur le modèle de caméra utilisé autre que la connaissance de la calibration interne. Il est alors tout à fait possible d'étendre notre chaîne à l'utilisation de nouveaux modèles de caméras :

- **des caméras sphériques.** L'estimation *a contrario* de matrices essentielles pour les images panoramiques serait alors un point d'entrée. Des résultats préliminaires encourageants ont été obtenus.
- **des caméras couleur et profondeur (*RGB+Depth*).** Nous pensons qu'utiliser les cartes de profondeur pour estimer des rotations et translations relatives entre vues successives permettrait de fournir des informations stables pour établir des reconstructions denses avec notre chaîne de calibration. Seul le module d'estimation d'informations relatives serait à modifier dans notre chaîne de calibration. Cela permettrait de reconstruire des scènes ou la photogrammétrie basée photo échoue. Nous pensons notamment à des scènes d'intérieurs où certaines images ne présentent pas assez de points saillants pour établir des estimations relatives de poses.

4. L'amélioration de la fusion de translations relatives convexe. Deux points intéressants peuvent être explorés pour la méthode d'optimisation convexe réalisant la fusion des translations relatives en translations globales :

- **extension pour la prise en compte de positionnement absolu.** Grâce à la formulation du programme linéaire il est facile d'ajouter des contraintes de positionnement terrain GPS pour certaines vues. Une reconstruction métrique à l'échelle terrain serait alors directement obtenue.
- **extension pour des systèmes multi-caméras rigide.** Grâce à la formulation du programme linéaire il est facile d'ajouter des contraintes pour forcer la contrainte de rigidité entre les caméras du système d'acquisition. Il suffirait de forcer les translations relatives inter-caméras à être égales entre des acquisitions différentes.

La combinaison des deux types de contraintes pourrait être appliquée pour réaliser des calibrations globales pour des systèmes d'acquisition multi-caméras tel que ceux utilisés par des voitures de *mobile-mapping* (*Google car*).

Limitations. Notre approche présente deux limitations :

1. Notre approche suppose un graphe épipolaire couvert de triplets, cependant ce cas n'est pas toujours observé suivant les acquisitions images réalisées. Mais nous pensons qu'il ne s'agit pas vraiment d'un point bloquant car les techniques de mise en correspondance de points saillants ne cessent de s'améliorer en rapidité et répétabilité. Des graphes de plus en plus connexes et donc de plus en plus aptes à être traités par notre méthode pourront ainsi être obtenus.
2. Notre étape d'estimation de translations relatives est dépendante des rotations globales. Ainsi, si les rotations globales sont fausses, des mouvements aberrants de translation seront calculés. Bien que nous ayons vu que l'utilisation d'étapes d'ajustements de faisceaux robustes permette de rattraper des erreurs de faible amplitude, l'étape d'inférence de rotations et de calcul des rotations globales restent les points à optimiser de notre chaîne.

Chapitre 7

Amélioration de la consistance colorée

Il est de plus en plus facile de photographier un sujet avec nos appareils photographiques numériques. Cependant lors de l'acquisition d'images, la consistance colorée n'est pas garantie car l'appareil photographique numérique adapte ses paramètres d'acquisition à la scène visée. On observe alors un problème de consistance colorée entre les images.

Ces différences de tons colorés sont gênantes lorsque l'on souhaite assembler plusieurs images sur une surface commune. Des effets indésirables apparaissent ainsi pendant la création de mosaïque d'images (images panoramiques) ou lorsque l'on souhaite restituer l'apparence visuelle d'objets 3D.

Afin de restituer au mieux cette consistance colorée nous proposons d'estimer et corriger la transformation colorée entre des pixels références et les couleurs observées dans la série d'image. Deux problèmes sont alors à résoudre :

1. **la sélection** des pixels communs entre images,
2. l'estimation et **la correction** d'une transformation colorée entre ces pixels.

Sommaire

7.1 Introduction

L'acquisition d'une scène avec un appareil photographique numérique est désormais on ne peut plus facile. On allume l'appareil photographique et on réalise une série d'images correspondant à la zone d'intérêt que l'on souhaite. Ces effets sont majoritairement dûs à l'ajustement automatique des paramètres d'acquisition réglés en fonction du contenu photographié. L'appareil photographique optimise ces paramètres pour obtenir une image qui lui semble la plus nette et contrastée. Pour chaque image on peut donc se retrouver avec des paramètres légèrement différents qui vont plus ou moins impacter la couleur du sujet d'une image à l'autre. DEBEVEC et MALIK (1997) ont proposé de modéliser l'appareil photographique comme une chaîne de traitements (cf. figure 7.1) où les éléments principaux suivants contrôlent l'acquisition d'image :

- l'ouverture de la lentille, *lens aperture*
- temps d'ouverture, *shutter-speed or sensor exposure*
- balance des blancs + tons, *remapping : white-balance, tone-mapping*.

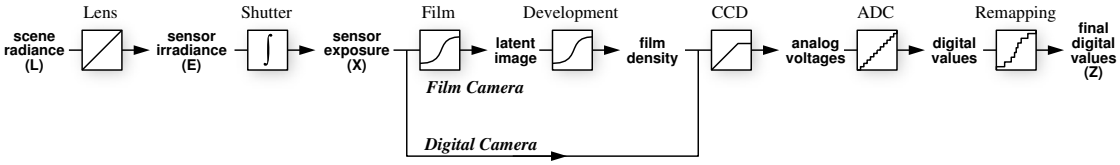


FIGURE 7.1 – Modèle de caméra de DEBEVEC et MALIK (1997).

On note que ces effets de changement de couleur sont observés avec modération sur des appareils hauts de gamme (reflex) à condition d'avoir fixé les paramètres d'acquisition et la balance des blancs. Des dérives plus larges seront observées sur des appareils photographiques grand public car les contrôles des paramètres d'acquisition sont peu ou pas accessibles. Cependant chaque constructeur d'appareil photographique proposant sa propre réponse colorée, on observe des tons colorés légèrement différents d'une marque à l'autre (cf. figure 7.2).

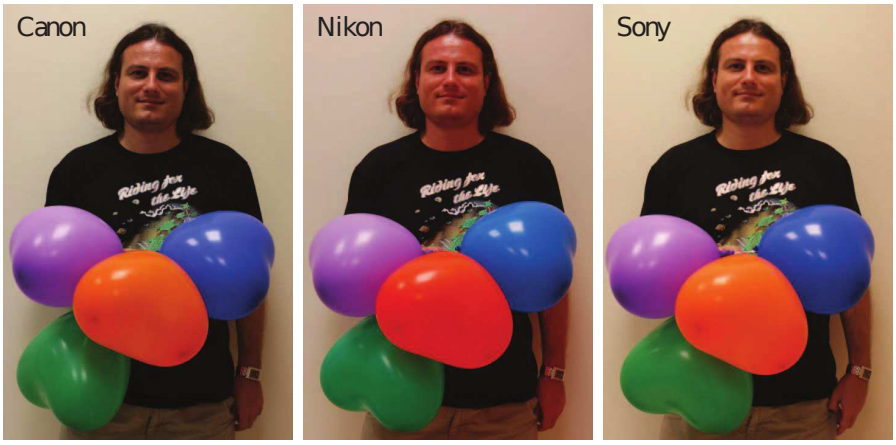


FIGURE 7.2 – Tons de couleurs en fonction de différents modèles de caméras pour des réglages identiques (Cf. KIM et al. (2012)).

Ce problème de consistance colorée est gênant lorsque l'on souhaite assembler différentes images afin de créer un panorama, ou bien que l'on souhaite restituer l'apparence visuelle d'un objet 3D. Afin d'obtenir la représentation photo-réaliste la plus

cohérente possible, il est souhaitable que ces différences colorées soient les plus faibles possibles. Le cas échéant, les différences lumineuses sont évidentes (cf. figure 7.3).

FIGURE 7.3 – Illustration de la problématique de la consistance colorée pour la construction d'images panoramiques (BROWN et LOWE 2007) et le cas de la restitution de l'apparence visuelle d'un maillage 3D (ALLENE et al. 2008).

7.2 État de l'art

Comme illustré dans l'étude comparative de XU et MULLIGAN (2010), de nombreuses méthodes existent pour corriger les effets de dérive de couleurs. Des méthodes locales et globales existent mais sont la plupart du temps appliquées sur des images présentant un fort recouvrement et testées sur des paires d'images et non des collections d'images. Nous présentons ci-dessous les méthodes nous semblant les plus à même de traiter des collections d'images et qui présenteront les notions utilisées pour notre méthode, expliquées en section 7.3.

Les méthodes séquentielles : REINHARD et al. (2001) proposent de rendre similaires les couleurs de deux images en alignant des distributions de pixels. Dans un premier temps le domaine des couleurs des images est modifié de RGB en $L\alpha\beta$ puis les images sont rendues similaires en alignant les distributions de pixels sur les 3 axes du domaine coloré en moyenne et variance.

DELON (2004) propose d'aligner au mieux les histogrammes des deux images considérées. Une fonction de passage permet de calculer un histogramme commun aux deux images : il s'agit d'un problème de *Midway equalization*. Cette méthode est basée sur l'analyse des différences des histogrammes cumulés des deux images. La méthode a ensuite été étendue pour réduire les effets de clignotements sur des séquences vidéos (DELON 2006) (cf. figure 7.4).

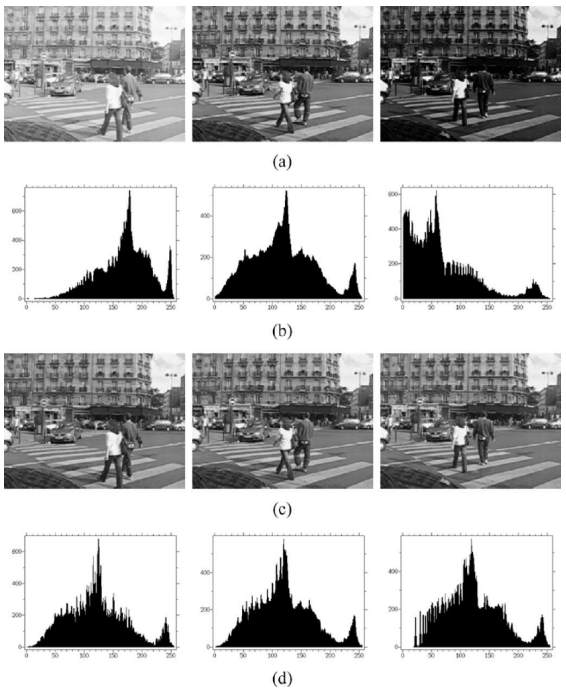


FIGURE 7.4 – (a) Images originales. (b) Histogrammes correspondants. (c) Image après *Midway equalization* (DELON 2006). (d) Les histogrammes après correction.

Afin de supporter des images présentant des sujets en mouvement ou des points de vues variables il est nécessaire de plus considérer les images dans leur globalité mais des zones locales en correspondances. HACOHEN et al. (2011) proposent de rechercher des régions présentant des caractéristiques similaires puis d'analyser la transformation couleur à réaliser pour ajuster les pixels en correspondance au mieux (cf. figure 7.5). La transformation colorée déterminée est ensuite appliquée à toute l'image. La méthode est appelée NRDC pour *Non Rigid Dense Correspondence*. La transformation colorée est modélisée par une courbe spline s'ajustant aux couleurs en correspondance pour chaque canal coloré (cf. figure 7.6).

Les méthodes globales : BROWN et LOWE (2007) proposent d'aligner les couleurs des images en utilisant une compensation de gain dans le cadre de construction d'images panoramiques. Ce calcul est réalisé en considérant un ensemble d'images en alignement sous contrainte homographique. Cette compensation de gain minimise la somme des différences colorées pour les pixels en correspondance dans la série d'images :

$$\underset{\{g_i\}_i}{minimise} \quad e = \frac{1}{2}(g_i I_i(u_i) - g_j I_j(u_j))^2, \quad \forall u_i \sim \mathbf{H}_{ij} u_j \tag{7.1}$$

avec g_i, g_j les gains et $u_i \sim \mathbf{H}_{ij} u_j$ les pixels en superposition sous la contrainte d'homographie.

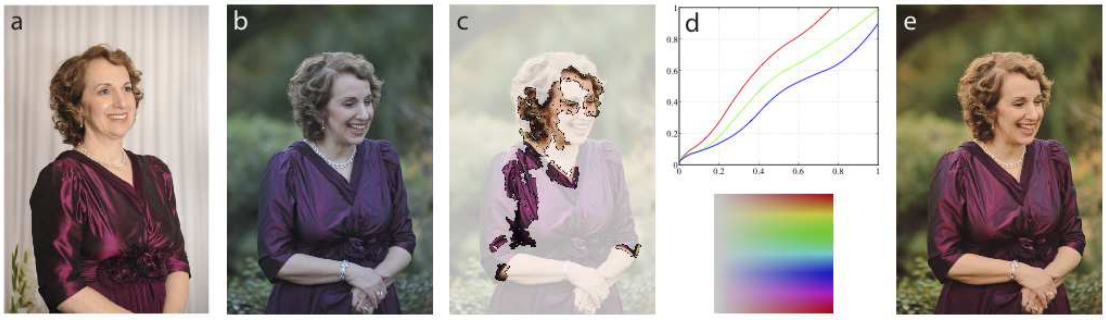


FIGURE 7.5 – (a) l'image de référence acquise en intérieur avec un flash, (b) image prise en extérieur avec un fond complètement différent, (c) les correspondances colorées identifiées, (d) le modèle paramétrique de transfert coloré. L'image (b) peut ainsi être modifiée pour ressembler à l'image référence (a).

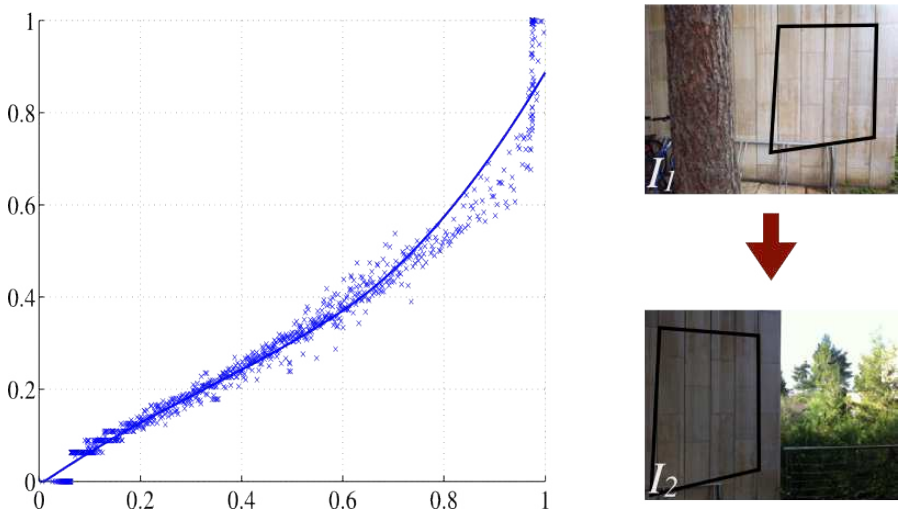


FIGURE 7.6 – Pour la zone mise en correspondance identifiée entre les images I_1 et I_2, la distribution des pixels en correspondances est analysée et simplifiée par une courbe spline à 7 points modélisant le passage d'une transformation colorée à l'autre (illustré pour la canal bleu ici).

Afin de limiter la taille des données à traiter, les auteurs ne considèrent pas l'ensemble des pixels en superposition mais des régions. Cette approximation permet de ne considérer en correspondance que les couleurs moyennes des régions R_{ij} communes entre les images i et j. La taille du problème à résoudre est alors moindre. Cependant on observe qu'une solution triviale au problème consiste à choisir des gains nuls : $\{g_i\}_i = \{0\}$. Des composantes de régularisation sur les gains et moyennes sont alors ajoutées pour forcer des variations modérées :

$$e = \frac{1}{2}\#R_{ij}((g_i M_i^{R_{ij}} - g_j M_j^{R_{ij}})^2/\sigma_N^2 + (1-g_i)^2/\sigma_g^2), \quad \forall R_{ij} \tag{7.2}$$

avec $M_i^{R_{ij}}$ la moyenne des pixels de la région R_{ij} commune entre les images i et j au sein de l'image i. σ_N^2 et σ_g^2 sont les variances des couleurs moyennes et des gains. Les valeurs $\sigma_N = 10.0$ et $\sigma_g = 0.1$ sont choisies.

Cette formulation peut être résolue aux moindres carrés en annulant le gradient de e de l'équation (7.2). Ainsi la consistance de région peut être rétablie avec un modèle de compensation de gain. Une approche de mélange fréquentiel, *multi-band-blending*, est

ensuite réalisée pour gommer les effets de vignette restants (décroissance lumineuse au bord des images) (BURT et ADELSON 1983) (cf. figure 7.7).

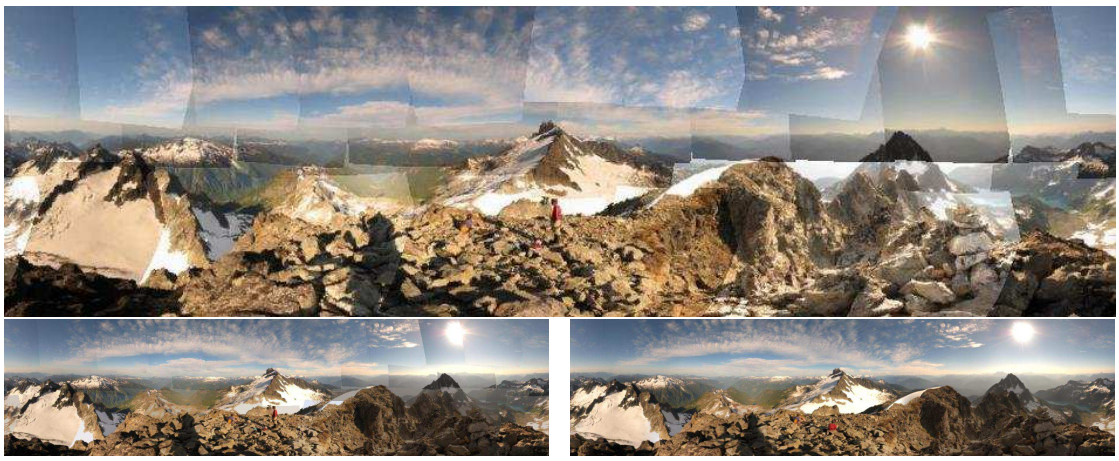


FIGURE 7.7 – En haut la mosaïque d'images d'origine, en bas à gauche l'ajustement de gain et en bas à droite le résultat du mélange fréquentiel (BROWN et LOWE 2007).

HACOHEN et al. (2013) proposent d'étendre leur méthode fonctionnant pour deux images (HACOHEN et al. 2011), à un contexte multiple-vues. A partir de transformations colorées, basées splines et NRDC, calculées pour chaque arête d'un graphe de connections images, un consensus d'ajustement global est identifié. Le critère d'ajustement recherche à minimiser la distance entre les courbes de corrections colorées sur l'ensemble des arêtes du graphe. Une pondération est utilisée pour prendre en compte les parties de dynamique de niveau de gris des images suivant leur support (nombre de points localement ajustés à la courbe paramétrique qui modélise la transformation couleur).

Les problèmes relatifs à l'existant

On note à travers les méthodes exposées que deux problèmes se posent :

1. la sélection des zones d'images présentant des couleurs communes,
2. l'évaluation de la fonction de la transformation colorée et sa correction.

Tandis que des méthodes considèrent des images représentant la même scène, d'autres considèrent des images prises de points de vues différents.Tandis que des méthodes utilisent des alignements d'histogrammes, d'autres considèrent l'alignement de valeurs moyennes de régions en correspondances.

Bien que HACOHEN et al. (2011) proposent une solution efficace pour identifier les zones en correspondance entre images, elle est réalisée sur des images miniatures pour garder des temps de calcul raisonnables. Un passage à l'échelle n'est alors pas réalisable, c'est pourquoi dans le cas de la généralisation de leur méthode à n images, ils utilisent plusieurs astuces pour réduire le nombre d'images à comparer (HACOHEN et al. 2013). Bien que BROWN et LOWE (2007) proposent une solution simple pour aligner des couleurs moyennes, la méthode dépend de nombreux paramètres.

Nous proposons une solution globale intermédiaire qui considère l'alignement de distributions de couleurs sous la forme d'une minimisation convexe, sans réglage de paramètres.

7.3 Une approche d'optimisation convexe pour améliorer la consistance colorée

Étant donné une série d'images $I : \{I_1, ..., I_n\}$ nous souhaitons aligner au mieux les couleurs des pixels correspondants. Afin d'éviter des phénomènes de dérives nous proposons une solution globale d'harmonisation colorée capable d'aligner des histogrammes de distributions colorées sous la norme l_∞.

Des zones cohérentes entre paires d'images sont identifiées et permettent de sélectionner les informations colorées représentant les mêmes objets de la scène sous forme d'histogrammes de distribution de pixels. En réalisant cette opération pour chaque arête d'un graphe de relation images il est estimé une série de 2-uplets d'histogrammes que l'on souhaite aligner au mieux de manière globale (cf. figure 7.8a,b).

L'harmonisation colorée est résolue en alignant les quantiles des histogrammes cumulés sous la norme l_∞. Un modèle de gain g_i et décalage o_i est utilisé par image pour modéliser les effets du temps d'acquisition et de l'ouverture de l'appareil photographique. Ce modèle permet de représenter des effets de compression, dilatation et translation des histogrammes. Ce processus est appliqué pour chaque canal couleur de manière indépendante et permet de construire une table de correspondance (*LUT*) pour réaligner les couleurs entre elles. Une solution est calculée grâce au programme linéaire suivant :

$$\begin{aligned} \underset{\{g_l, o_l\}_l, \gamma}{\text{minimise}} \quad & \gamma \\ \text{tel que} \quad & |(g_i Q^k_{ij} + o_i) - (g_j Q^k_{ji} + o_j)| \leq \gamma, \quad \forall i, j, k \\ & g_l \geq 0, \quad \forall l \\ & g_{\text{ref}} = 1, \\ \text{et} \quad & o_{\text{ref}} = 0, \end{aligned} \tag{7.3}$$

avec Q^k_{ij} le k^e quantile de l'histogramme cumulé représentant les pixels en commun pour la paire d'images (i, j). Afin d'éviter une solution dégénérée, les gains sont forcés à être positifs et une image de référence est choisie avec une transformation identité en imposant $g_{\text{ref}} = 1$ et $o_{\text{ref}} = 0$. Expérimentalement, utiliser 10 quantiles par histogramme est suffisant, en utiliser plus ne donne pas de différence visible. Le processus d'alignement est illustré sur la figure 7.8.

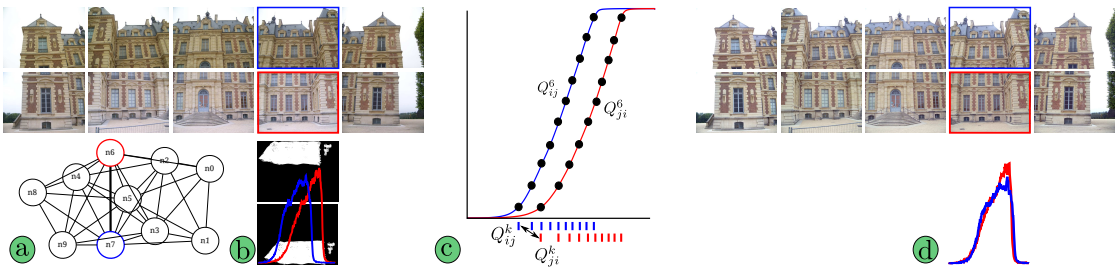


FIGURE 7.8 – Pour un graphe d'images (a), la consistance colorée est calculée comme suit : (b) les masques de données cohérentes sont utilisés pour extraire les couleurs communes pour chaque arête et identifier les histogrammes à aligner. (c) les quantiles des histogrammes cumulés sont alignés sous la norme l_∞. (d) les gains et décalages calculés sont utilisés pour corriger les images. Noter la meilleure consistance des images encadrées en bleu et en rouge et l'alignement effectif des histogrammes.

Le processus d'harmonisation est constitué de 3 étapes principales (décrites par la procédure 12) :

1. Récupération des distributions colorées en correspondance par paire d'images,
2. Alignement global des histogrammes cumulés pour chaque canal couleur,
3. Correction des images.

Procédure 12 Harmonisation couleur : Alignement d'histogrammes sous la norme l_∞

Entrée: $\{I\}_i$: les images à corriger,
$\{P\}_{ij}$: zones en correspondance par paires d'image (i, j),
ref : index de l'image de référence.
Sortie: $\{I_o\}_i$: les images corrigées.

(1) **Récupération des distributions colorées par paires d'images :**
pour $p \in \{P\}_{ij}$ **faire**
Calcul des histogrammes H_{ij} et H_{ji} pour les zones en correspondance entre I_i et I_j
Calcul des histogrammes cumulés Q_{ij} et Q_{ji}
fin pour

(2) **Alignement des histogrammes cumulés** $\{(Q_{ij}, Q_{ji})\}_{ij}$ **par canal couleur :**
pour $c \in \{R, G, B\}$ **faire**
$\{g_{lc}, o_{lc}\}_l, \gamma$ = Minimisation de l'équation 7.3 pour le canal couleur $c|$ref
fin pour

(3) **Correction des images :**
Calcul de LUT de correction en fonction des $\{g_{lc}, o_{lc}\}_l$ calculés
Transformation des images $\{I\}_i$ en images harmonisées $\{I_o\}_i$

Le fait d'utiliser une représentation sous forme d'histogrammes nous apporte les avantages suivants :

1. On utilise une place en mémoire constante quel que soit le nombre de pixels considérés, ce qui est important pour garantir un fonctionnement à large échelle de la méthode,
2. La répartition des couleurs est bien représentée (bien plus riche que les simples moyennes utilisées par BROWN et LOWE (2007)),
3. Le fait d'utiliser une distribution plutôt que des correspondances point à point vient lisser les effets d'erreurs d'alignement des régions en correspondance.
4. Aucune pré-normalisation des histogrammes n'est nécessaire pour réaliser l'alignement car on utilise les quantiles des distributions.

En fonction de l'usage, l'algorithme 12 réalise la sélection de régions en correspondances la plus appropriée (cf. figure 7.9) :

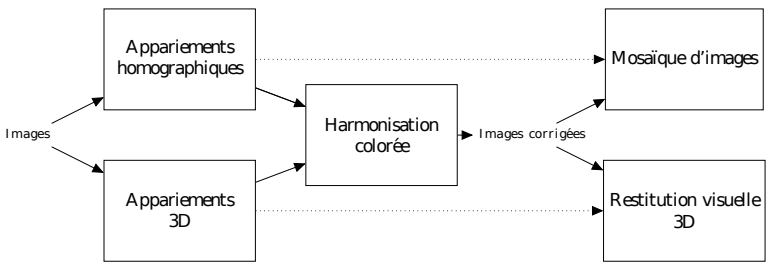


FIGURE 7.9 – Cadre d'utilisation de notre méthode d'harmonisation colorée.

Les régions en correspondance par paire d'images, suivant l'application, peuvent être construites de différentes manières (cf. figure 7.9) :

Harmonisation colorée pour construction de mosaïque d'images :
Pour chaque paire d'images il est réalisé un appariement robuste de points saillants et segments virtuels : des paires de points géométriquement cohérents sont identifiées par appariement de points SIFT et estimation robuste *a contrario* d'une matrice d'homographie. Cet ensemble de points assurant une faible couverture de la zone image (et donc un ensemble restreint d'informations colorées), nous étendons les zones en correspondances en calculant les segments VLD consistants entre les deux images (LIU et MARLET 2012). Ces segments en correspondances sont ensuite utilisés comme masque pour sélectionner les pixels et couleurs en commun entre images (cf. figure 7.10).

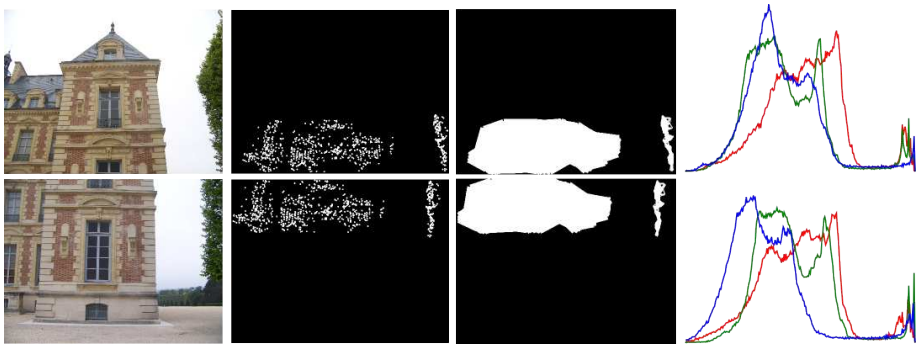


FIGURE 7.10 – De gauche à droite, les images d'origines, les points SIFT géométriquement cohérents, les segments VLD et les histogrammes extraits des masques. On observe pour la paire en question un déplacement de l'histogramme du canal vert.

Harmonisation colorée pour la restitution de l'apparence visuelle d'un modèle 3D :
Les informations de visibilité des points 3D reconstruits, les traces, sont utilisées pour créer les zones en cohérence entre paires d'images. Pour chaque paire, les points 3D reconstruits sont listés et leur projection images sont utilisées pour construire les masques de pixels communs (chaque projection apporte une contribution sous la forme d'un cercle de taille fixé à 10 pixels). Si aucune reconstruction 3D n'est disponible, les traces issues de points saillants mis en correspondance avec une estimation robuste *a contrario* de la matrice fondamentale peuvent être utilisées. Afin d'éviter des appariements aberrants seules les traces de longueurs supérieures ou égales à trois sont utilisées.

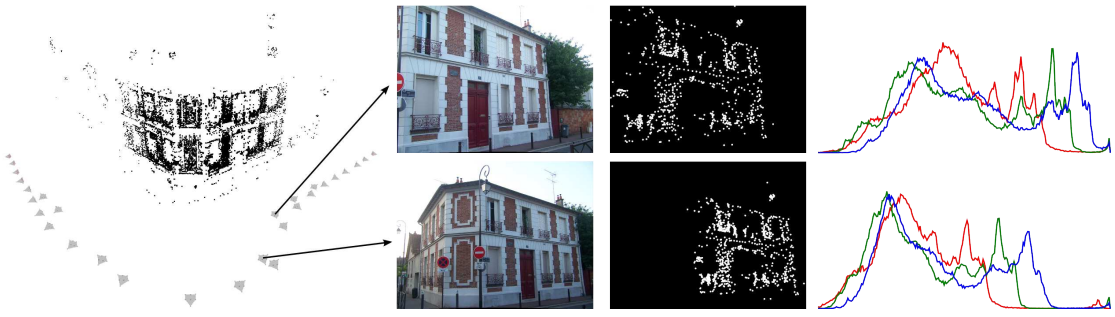


FIGURE 7.11 – De gauche à droite, la configuration de la scène, deux images de la scène, les points SIFT géométriquement cohérents et les histogrammes extraits des masques. On observe un décalage des trois histogrammes.

7.3.1 Évaluations

Nous proposons d'évaluer notre méthode globale sur trois cas d'utilisation différents :

1. des images parfaitement alignées,
2. des images acquises pour la création d'une mosaïque d'images,
3. des images utilisées pour calculer la restitution de l'apparence colorée d'un modèle 3D calculé par photogrammétrie.

1. Vérification du fonctionnement sur une séquence d'images alignées.
Nous allons évaluer notre méthode sur deux jeux de données, Chalk et BiWall, extraits de CHAKRABARTI et al. (2009). Les erreurs moyennes d'alignements des histogrammes avant et après notre méthode d'harmonisation colorée sur ces deux jeux de données seront comparées. Les masques de sélection sont construits à partir de correspondances géométriques validées par estimation robuste d'homographie *a contrario* et utilisation des segments VLD.

Chalk (cf. figure 7.12)

Un objet est photographié par 4 appareils (Canon EOS20D, Canon Powershot G1, Sony A100 et Canon Powershot A540) avec des paramètres de balance différents (soleil, nuageux, fluorescent, personnalisé en fonction d'une mire colorée). Une série de 48 images présentant ainsi une large gamme de teintes est ainsi obtenue.

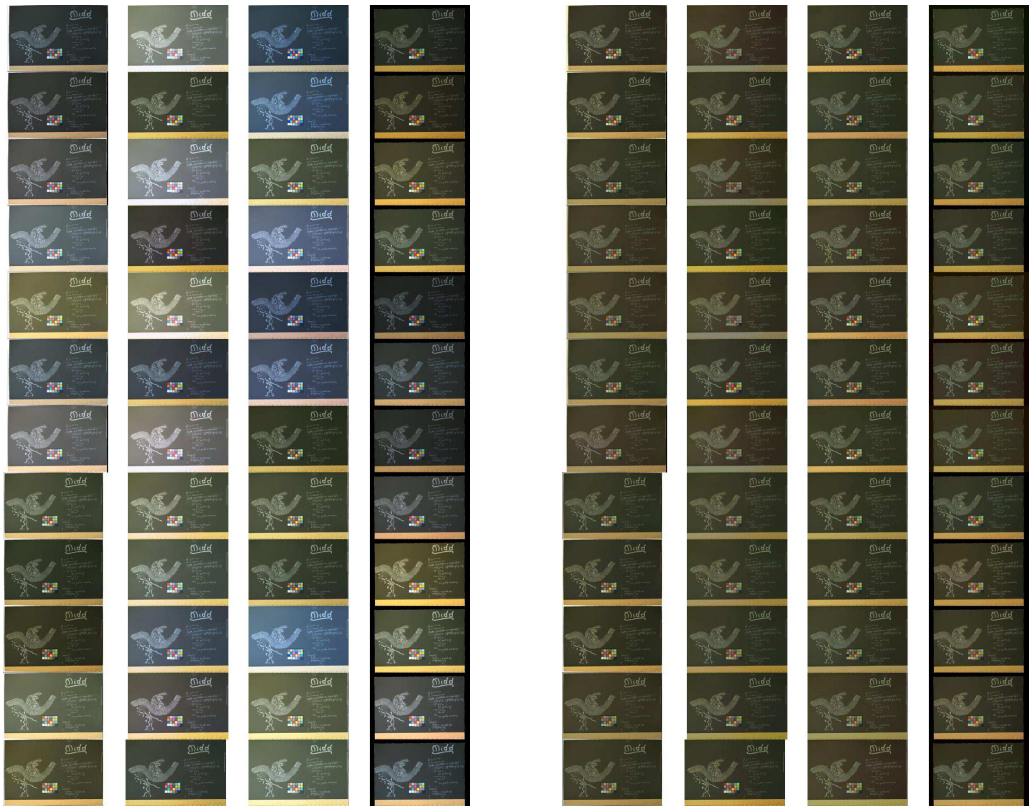
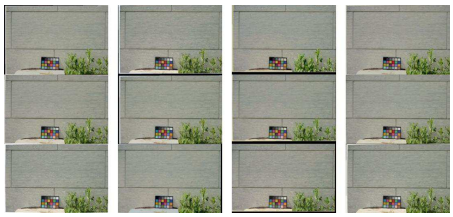


FIGURE 7.12 – Résultats de l'harmonisation colorée (droite) pour la scène Chalk pour les 5 caméras avec 4 temps d'exposition et 4 balances différentes (gauche).

BiWall (cf. figure 7.13)

Une scène est photographiée avec 1 appareil sous 4 balances différentes et 3 temps d'exposition. Une série de 12 images est obtenue.

(a) Images d'entrées de la scène Biwall.



(b) Le résultat de l'harmonisation colorée calculée de manière globale.

FIGURE 7.13 – Résultats de l'harmonisation colorée pour la scène Biwall pour la caméra Canon EOS20D avec 3 temps d'exposition et 4 balances différentes.

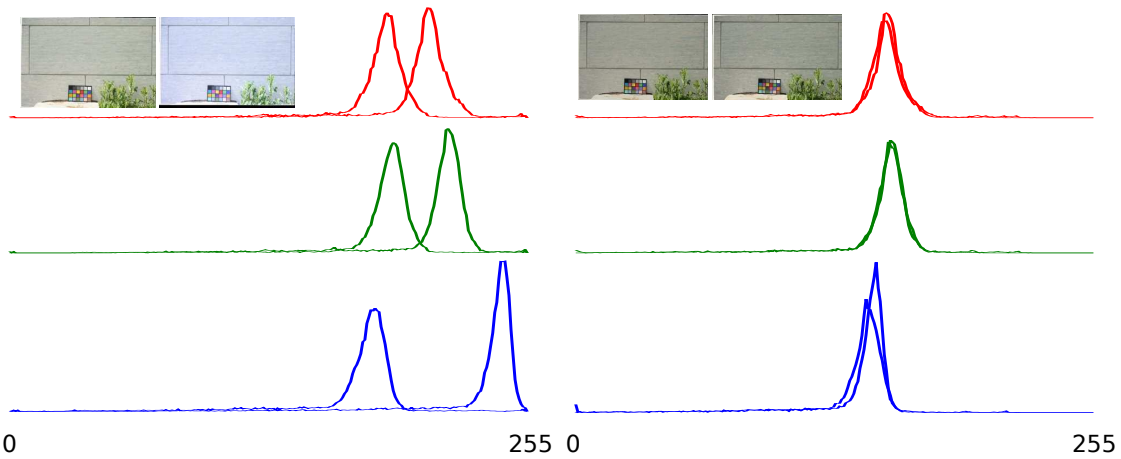


FIGURE 7.14 – Histogrammes avant (gauche) et après alignements (droite) pour une paire d'images de la figure 7.13.

On note que la méthode fonctionne comme escompté, les erreurs d'alignements des moyennes des histogrammes ont bien été réduites de manière significative (cf. figure 7.14). Les résultats numériques (cf. tableau 7.1) et visuels (cf. figure 7.12,7.13) le confirment.

Jeux de données	Canal couleur	Avant Harmonisation	Après Harmonisation	γ	#Images	#Arêtes
Chalk	Rouge	28.3	2.6	14.6	48	1128
	Vert	28.6	1.6	9.2		
	Bleu	33.5	2.5	7.5		
Biwall	Rouge	30.2	0.6	4	12	68
	Vert	29.1	0.6	3.9		
	Bleu	36.1	0.8	4.6		

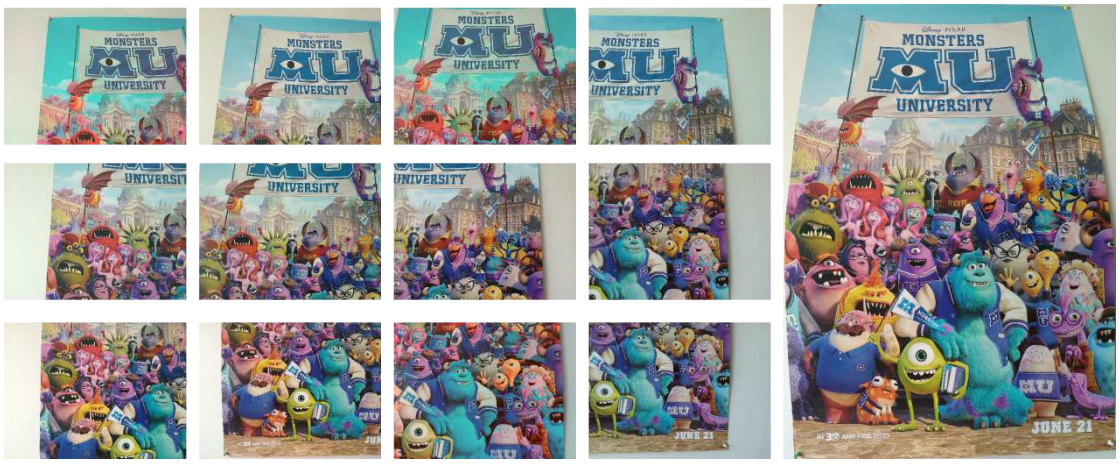
TABLE 7.1 – Résultat de l'harmonisation colorée (distances moyennes entre les moyennes des histogrammes pour l'ensemble des images, par canal couleur) et statistiques des graphes images considérés.

Examinons la réaction du système sur une scène un peu plus complexe (le cas d'une mosaïque).

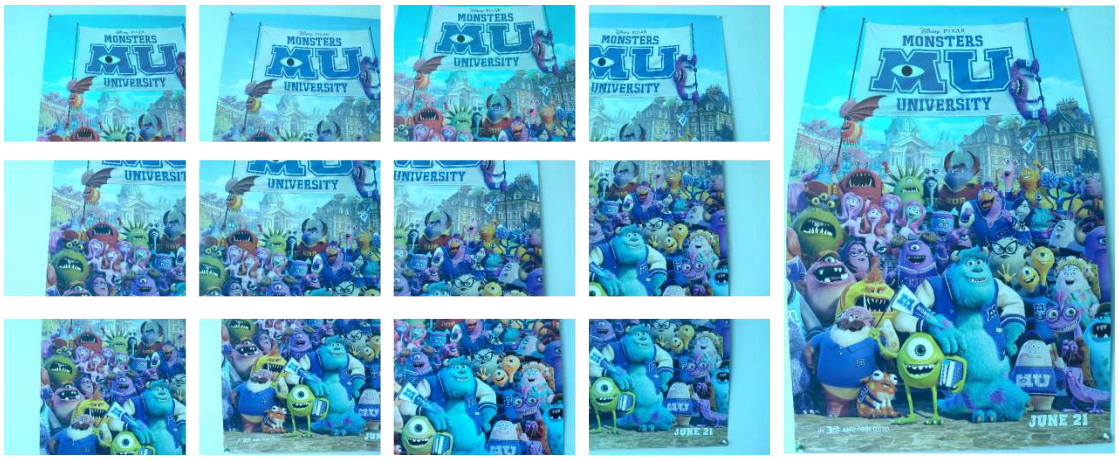
2. Vérification du fonctionnement sur une séquence d'images type panorama.

Nous allons ici vérifier le comportement de l'algorithme d'ajustement coloré dans le cas ou les images ne partagent qu'un contenu partiel : des images acquises pour la réalisation d'un panorama. Où 9 images sont acquises avec une balance lumière du jour et 3 images avec une balance incandescente sur un appareil photographique de téléphone mobile. Nous réalisons l'harmonisation globale de l'ensemble des images en choisissant une image issue de chaque configuration de balance (cf. figure 7.15).

(a) Images d'origine capturées avec une balance «lumière du jour» et «lumière incandescente».



(b) Images harmonisées par rapport à une image balance «lumière du jour» et mosaïque résultante.



(c) Images harmonisées par rapport à une image balance «lumière incandescente» et mosaïque résultante.

FIGURE 7.15 – Harmonisation colorée de 14 images sur un graphe comportant 80 arêtes.

On observe que les mosaïques d'images sont consistantes et que l'harmonisation visuelle donne des résultats convaincants, que ce soit pour l'harmonisation en balance «lumière du jour» ou bien «lumière incandescente». Les résultats numériques de l'harmonisation globale sont résumés sur le tableau 7.2. Un exemple visuel de l'alignement des histogrammes est présenté sur la figure 7.16, la dilatation déterminée est clairement visible.

Monstre UniversitéP12	Canal couleur	Avant Harmonisation	Après Harmonisation	γ	#Images	#Arêtes
lumière du jour	Rouge	33.4	6.4	24.0	12	80
	Vert	20.5	2.4	9.3		
	Bleu	28.4	3.0	11.6		
lumière incandescente	Rouge	33.4	5.9	20.6	12	80
	Vert	20.5	2.2	9.3		
	Bleu	28.4	2.9	12.9		

TABLE 7.2 – Résultat de l'harmonisation colorée (distances moyennes entre les histogrammes pour l'ensemble des images, par canal couleur) et statistique du graphe image considéré.

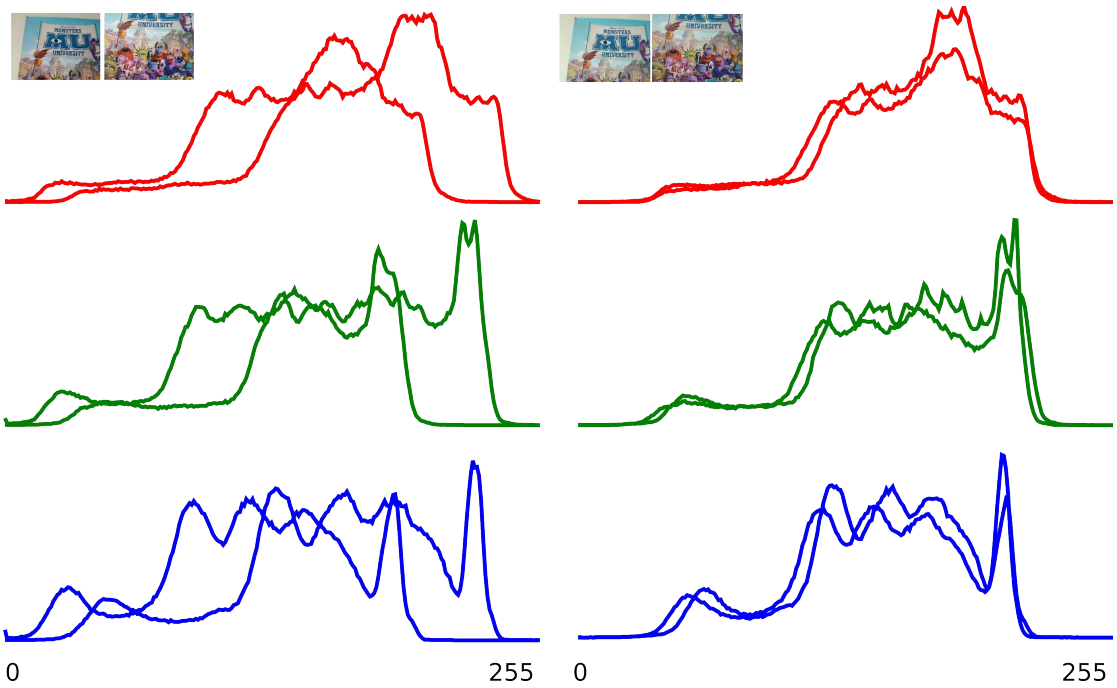


FIGURE 7.16 – Résultat de l'alignement des histogrammes pour une des 80 arêtes du graphe de l'expérience 7.15b. A gauche la situation avant alignement, à droite l'alignement réalisé. La dilatation des histogrammes calculée pour l'ajustement est nettement visible.

3. Application pour l'amélioration du rendu visuel d'un modèle 3D.

Nous proposons maintenant d'utiliser des images harmonisées pour améliorer la restitution visuelle de maillages. Des images avant et après harmonisation vont être utilisées et les restitutions comparées. La méthode de restitution visuelle d'ALLENE et al. (2008) est utilisée. ALLENE et al. (2008) identifient les images à projeter sur le

maillage sous la forme d'un problème de partition du surface. Pour chaque face du maillage une image est choisie de telle sorte qu'elle offre le plus de détail et le moins de discontinuité colorée sur son contour. Nous montrons que des coupures notables sont visibles lorsque les images présentent des différences de tons et que lorsque les images harmonisées sont utilisées les restitutions visuelles sont de meilleures qualités : les différences de tons sont moindres, voire absentes.

Notre méthode d'harmonisation colorée est testée sur deux jeux de données :

- MayaP50 (cf. figure 7.17),
- AntonyP29 (cf. figure 7.18).

La visibilité des points 3D reconstruits est utilisée pour créer les zones en cohérences entre des paires d'images. Chaque projection image utilisée ajoute une contribution aux masques de pixels communs (en utilisant un rayon de taille 10 pixels) (cf. figure 7.11).

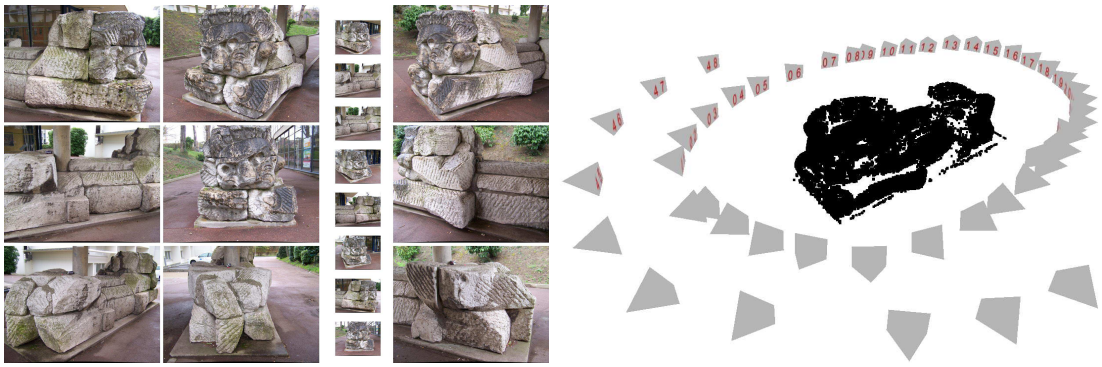


FIGURE 7.17 – MayaHeadP50 : images et configuration 3D de la scène (positionnement des caméras et les points 3D utilisés).

FIGURE 7.18 – AntonyP29 : images et configuration 3D de la scène (positionnement des caméras et les points 3D utilisés).

Résultats visuels.
Les figures 7.19, 7.20 et 7.21 présentent les restitutions visuelles identifiées avec les images originales et les images harmonisées. On note que les coupures franches ne sont plus visibles ou alors de très faible amplitude.

FIGURE 7.19 – MayaHeadP50 : Illustration de la restitution visuelle en utilisant les images originales (haut) ou bien en utilisant les images harmonisées (bas). Les différences de sélection d'images ne sont plus visibles.

FIGURE 7.20 – AntonyP29 : Gros plans sur les restitutions visuelles obtenues avec les images d'origine et les images harmonisées.

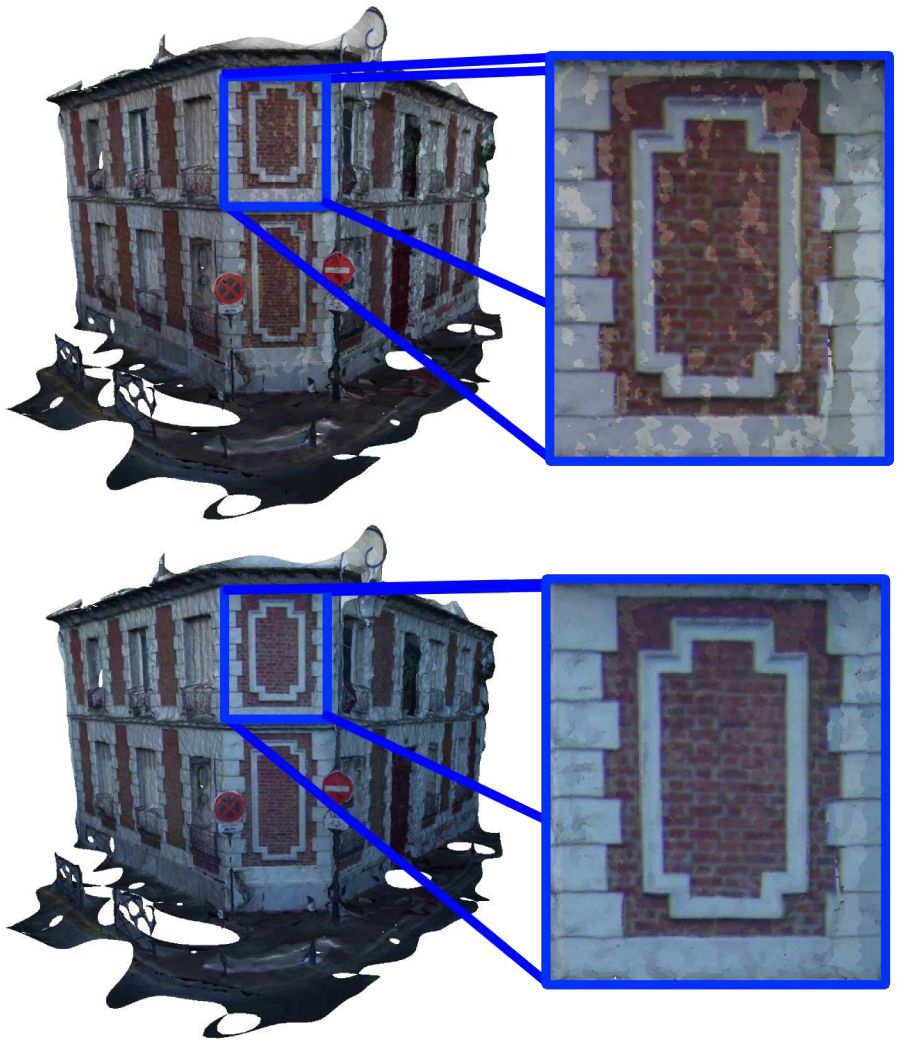


FIGURE 7.21 – AntonyP29 : Illustration de la restitution visuelle en utilisant les images originales (haut) ou bien en utilisant les images harmonisées (bas). Les différences de sélection d'images ne sont plus visibles ou de moindre amplitudes.

Résultats numériques. Les résultats visuels sont confirmés par les résultats numériques du tableau 7.3.

Jeux de données	Canal couleur	Avant Harmonisation	Après Harmonisation	γ	#Images	#Arêtes
MayaHeadP50	Rouge	13.8	2.6	16.2	50	311
	Vert	13.5	2.2	16.3		
	Bleu	13.2	2.8	17.1		
AntonyP29	Rouge	10.5	2.9	14.7	29	198
	Vert	13.7	4.3	13.3		
	Bleu	13.8	4.3	16.5		

TABLE 7.3 – Résultat de l'harmonisation colorée pour les scènes 3D MayaHeadP50 et AntonyP29 (distances moyennes entre les histogrammes pour l'ensemble des images, par canal couleur) et statistique du graphe image considéré.

7.4 Contributions et perspectives

Nous avons présenté dans ce chapitre les problématiques liées à l'harmonisation colorée pour des images capturées avec des points de vue différents. Les problématiques de sélection de données communes entre images et une méthode globale d'harmonisation d'histogramme basée sur une minimisation l_∞ ont été exposées.

Nos expériences utilisées dans les contextes suivants :

- des images en parfait alignement,
- des images sous contrainte homographique ne présentant que du contenu local,
- des images multiple-vues d'un objet,

démontrent expérimentalement que notre méthode aligne «aux mieux» les histogrammes pour les contraintes formulées et que les corrections calculées sont visuellement et quantitativement cohérentes.

Notre méthode permet de réaliser des ajustements avec plus de précision que la méthode de BROWN et LOWE (2007) tout en étant sans paramètres. L'utilisation d'histogrammes et non de valeurs moyennes permet un ajustement précis. L'utilisation d'un programme linéaire permet de forcer certains paramètres sans nécessiter l'utilisation de paramètres supplémentaires. Contrairement à REINHARD et al. (2001) qui aligne les moyennes et variances de deux distributions notre méthode réalise ce travail de manière plus complète et globale pour un ensemble de distributions en utilisant les quantiles, et ce grâce à une optimisation convexe.

Ces travaux ont été publiés à la conférence CVMP (MOULON et al. 2013c).

Perspectives : Le fait d'utiliser une formulation avec un programme linéaire du problème nous permet :

- d'ajuster à volonté le nombre d'images considérées comme référence, il suffit de rajouter des contraintes sur les gains et décalages des images concernées,
- de choisir si la compression ou dilatation d'histogrammes est utilisée, et ce, par image. Exemple : la compression des histogrammes peut être empêchée en définissant des gains supérieur ou égal à 1 : $\{g_i\}_i \geq 1, \quad \forall i \in \{0,...,n\}$.

Nous pensons que notre méthode permettrait de corriger de manière effective des effets de clignotement : *flickering*, ou des dérives locales d'ajustements d'exposition de fichiers vidéos. Du fait que plusieurs images puissent être configurées en tant que référence, un utilisateur pourrait facilement guider le processus pour obtenir le résultat souhaité. Nous pensons aussi que notre méthode pourrait être utilisée pour harmoniser les réponses colorées de différents appareils photographique ayant réalisé l'acquisition d'une scène.

Il serait intéressant de considérer non plus le cas d'alignement global d'une série d'histogrammes mais l'alignement d'une série de courbes. En estimant un gain et offset différent pour chaque quantile des déformations non linéaires pourraient être calculées. Il serait alors nécessaire d'imposer des contraintes de conservation de l'ordre des positions des quantiles, ce qui est toujours exprimable sous la forme d'un programme linéaire.

Limitations : Les limitations sont communes à toutes les méthodes exposées :

- des décalages d'histogrammes hors du domaine de définition de l'image peuvent être observés,

- des effets de saturation sont très souvent observés pour les zones de ciel car ces gammes de couleurs ne sont pas mises en correspondance et donc ne sont pas corrigées de manière optimale.

Chapitre 8

Conclusion et perspectives

Les travaux présentés dans ce manuscrit s'intéressent à l'analyse des problèmes critiques au cœur des méthodes de calibration externe et sont articulés autour de solutions pour améliorer leur performance (précision, robustesse, vitesse) et leur facilité d'utilisation (paramétrisation restreinte).

Précision, robustesse : Une attention toute particulière a été portée sur l'utilisation de méthodes permettant de vérifier la pertinence des solutions en cours de calcul, que ce soit avec l'utilisation :

- de la méthodologie *a contrario* pour évaluer statistiquement la pertinence d'un consensus face à un modèle paramétrique en cours d'évaluation,
- de méthodes d'optimisation convexe pour vérifier la faisabilité des problèmes en cours d'évaluation et garantir l'identification d'une solution optimale face à un jeu de contraintes.

Les résultats observés sont une amélioration notable des précisions d'estimations de pose et orientation des caméras pour les chaînes de calibration séquentielles et globales par rapport aux méthodes de l'état de l'art.

Vitesse : Afin de garantir les meilleurs temps de calcul possibles, nous avons proposé :

- un algorithme de suivi de points rapide et efficace,
- une chaîne de calibration externe globale parallélisable.

Paramétrisation restreinte : Nous avons montré que l'utilisation généralisée de l'estimation robuste de modèles paramétriques *a contrario* permet :

- de libérer l'utilisateur du réglage de nombreux seuils de détection,
- d'obtenir une chaîne de reconstruction qui s'adapte automatiquement aux données.

Dans un deuxième temps nous avons proposé d'améliorer la consistance colorée de collections d'images par optimisation convexe avec l'utilisation de la programmation linéaire. Cette contribution originale qui réalise l'alignement global de distributions colorées à travers un graphe a démontré sa performance dans le cas de l'amélioration de la restitution visuelle d'images panoramiques et de modèles 3D.

Les contributions de cette thèse sont les suivantes :

- un algorithme de suivi de points efficace et consistant,
- la généralisation de l'utilisation d'estimateurs robustes *a contrario* à de nouveaux modèles paramétriques :
  - matrices essentielles,
  - matrices de pose,
  - tenseurs tri-focaux,
  - adaptation pour utiliser des erreurs angulaires ;
- la vérification de l'impact, à large échelle, d'estimateurs robustes adaptatifs dans les méthodes séquentielles de calibration externe ;
- un apport pour la robustesse et le passage à l'échelle pour l'estimation globale de la position d'un réseau de caméras, via :
  - l'estimation rapide, précise et robuste de tenseurs tri-focaux réduits, par programmation linéaire et de la méthodologie *a contrario*,
  - la minimisation convexe de la fusion de translations relatives dans un repère global commun sous la norme l_∞ ;
- l'harmonisation colorée d'un ensemble d'images multiple-vues sous la norme l_∞.

Perspectives :

Suivi de points et construction de traces : Bien que notre algorithme de construction de traces présente l'avantage d'être extrêmement simple et d'identifier une solution au problème dans des temps quasi-linéaires, la possibilité de réaliser cette fusion de correspondances de manière parallèle et le problème de découpage des traces ambiguës restent à explorer.

Estimation robuste adaptative *a contrario* : Nos expérimentations démontrent que ne plus être dépendant des effets de l'efficacité relative liés à l'utilisation de seuils fixes permet de simplifier la chaîne algorithmique mais aussi d'améliorer la qualité des reconstructions. Cependant nous pensons que deux pistes restent à explorer :

1. L'estimation conjointe de la distorsion radiale et de modèles paramétriques (BRITO et al. 2013).
2. La suppression de seuils dans l'étape de mise en correspondance photométrique de points saillants (descripteurs) (RABIN et al. 2008).

Notre cadre de travail ne considérant que des images où la distorsion radiale était corrigée à l'avance nous pensons qu'estimer la distorsion radiale *a contrario* serait intéressant. Ceci est d'autant plus vrai que les méthodes développées utilisent pour l'instant des méthodes de types RANSAC et continuent à utiliser des seuils fixes alors que la distorsion fait elle aussi intervenir une précision δ inconnue.

Amélioration de la vitesse d'exécution de l'inférence : Nous avons constaté expérimentalement que l'étape d'inférence dans le graphe de rotations relatives est coûteuse sur des graphes de large dimension. Nous pensons que cette étape pourrait être accélérée et nous voulons suivre avec intérêt les travaux de CHATTERJEE et GOVINDU (2013) et pensons que de nombreuses méthodes liées aux SLAM, *Simultaneous Localization And Mapping*, pourraient être adaptées à ce problème avec succès (cf. LEE et al. (2013)).

Estimation de meilleures rotations relatives et globales : Des matrices essentielles sont pour l'instant utilisées pour générer des matrices de rotation sans prendre en compte la typologie de la scène observée, ce qui met en évidence un problème ouvert :

- les matrices de rotation devraient être estimées en fonction de la géométrie de la scène. Par exemple, une scène présentant de nombreuses correspondances planes devrait permettre d'estimer un matrice de rotation très précise par décomposition d'une homographie, et ce, sans risque d'établir de fausses correspondances qui auraient été valides sous géométrie épipolaire.
- les matrices de rotation pourraient être calculées et optimisées sans considérer les translations relatives (KNEIP et al. 2012).

Évolution de la chaîne de calibration globale : Nous pensons que notre chaîne de calibration globale peut être généralisée à d'autres modèles de caméras :

- Caméras sphériques :

  Nous avons démontré la faisabilité du calcul de la matrice essentielle *a contrario* pour des images panoramiques, cependant nous n'avons pas encore réalisé l'estimation de tenseur tri-focaux réduits pour des images panoramiques, phase indispensable pour garantir des optimisations stables. Nous sommes proches de pouvoir réaliser une chaîne de calibration externe globale à partir d'images panoramiques. Les problèmes de dérives souvent observés, (BAZIN et al. 2013 ; PAGANI et STRICKER 2011), seraient alors réduits pour des trajectoires rectilignes et supprimés pour le cas de cycles.
- Caméras couleur et profondeur (*RGB+Depth*) :

  Notre méthode supposant des mouvements relatifs de rotation et translation, elle peut être étendue pour gérer des reconstructions à partir de systèmes d'acquisitions à base de lumière structurée (type *Kinect*). Des correspondances denses entre captures proches pourraient alors fournir des rotations et translations relatives stables par utilisation de méthodes de type ICP (*Iterative Closest Point*). Tandis que les cycles seraient détectés par appariements d'images (GALVEZ-LOPEZ et TARDOS 2011). L'approche serait alors plus naturelle que les approches réalisant des corrections des effets de dérives (ZHOU et KOLTUN 2013).

De même nous pensons que l'utilisation de contraintes lors du calcul de translation globale serait une piste intéressante pour :

- Utiliser des données GPS de positionnement et obtenir une reconstruction directement à l'échelle monde métrique,
- Utiliser des contraintes inter-caméras pour forcer des distances entre images extraites de systèmes d'acquisition multi-caméras. Nous pensons notamment aux systèmes d'acquisition utilisés couramment en *mobile-mapping* ou plusieurs caméras sont fixées sur un véhicule.

Suppression des méthodes d'estimations basées sur le tirage stochastique de correspondances : Nous avons vu que l'ensemble des méthodes de calibration externes sont basées sur l'utilisation de modèles paramétriques qui sont identifiés par des méthodes de tirage aléatoire d'échantillons de n-uplets de correspondance. Le tirage d'un nombre important d'échantillon nous assure d'avoir évalué un grand nombre de solutions, cependant l'espace des solutions n'est pas évalué dans sa globalité. Il y a donc malgré tous nos efforts pour traiter le problème dans sa globalité un aspect local qui reste présent. Nous pensons que des méthodes globales pourraient être utilisées pour

l'ensemble des estimations de modèles paramétriques avec des méthodologies d'estimations basées sur des procédures de type séparation et évaluation, *Branch-And-Bound*. Des travaux préliminaires ont démontré la faisabilité (BAZIN et al. 2012), cependant un seuil fixe limitant la précision est toujours utilisé. Il serait alors intéressant de mêler les méthodes d'estimation d'optimalité *Branch-And-Bound* avec un critère de validation *a contrario* pour supprimer ce paramètre de précision a priori. Des méthodes d'estimations globales sans paramètres de précision a priori seraient ainsi utilisées du début à la fin de la chaîne de calibration.

Bibliographie

AANÆS, HENRIK, ANDERS LINDBJERG DAHL et KIM STEENSTRUP PEDERSEN (2012). Interesting interest points. Dans : *International Journal of Computer Vision*, **97** :1, p. 18–35 (cf. p. 101)

AGARWAL, SAMEER et KEIR MIERLE (2012). Ceres Solver : Tutorial & Reference. Dans : *Google Inc*, (cf. p. 48, 110)

AGARWAL, SAMEER, NOAH SNAVELY et STEVEN M SEITZ (2008). Fast algorithms for L_∞ problems in multiview geometry. Dans : *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, p. 1–8 (cf. p. 119)

AGARWAL, SAMEER, NOAH SNAVELY, IAN SIMON, STEVEN M SEITZ et RICHARD SZELISKI (2009). Building rome in a day. Dans : *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, p. 72–79 (cf. p. 29, 63)

ALCANTARILLA, P. F., J. NUEVO et A. BARTOLI (2013). Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. Dans : *British Machine Vision Conf. (BMVC)*. Bristol, UK (cf. p. 52)

ALCANTARILLA, PABLO FERNÁNDEZ, ADRIEN BARTOLI et ANDREW J DAVISON (2012). « KAZE features ». Dans : *Computer Vision–ECCV 2012*. Springer, p. 214–227 (cf. p. 52)

ALLENE, CÉDRIC, J-P PONS et RENAUD KERIVEN (2008). Seamless image-based texture atlases using multi-band blending. Dans : *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, p. 1–4 (cf. p. 161, 171)

ANDERSON, RICHARD J et HEATHER WOLL (1991). Wait-free parallel algorithms for the union-find problem. Dans : *Proceedings of the twenty-third annual ACM symposium on Theory of computing*. ACM, p. 370–380 (cf. p. 63)

ARIE-NACHIMSON, MICA, SHAHAR Z KOVALSKY, IRA KEMELMACHER-SHLIZERMAN, AMIT SINGER et RONEN BASRI (2012). Global motion estimation from point matches. Dans : *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*. IEEE, p. 81–88 (cf. p. 115, 118, 122, 143, 144)

ARIKAN, MURAT, MICHAEL SCHWÄRZLER, SIMON FLÖRY, MICHAEL WIMMER et STEFAN MAIERHOFER (jan. 2013). O-Snap : Optimization-Based Snapping for Modeling Architecture. Dans : *ACM Transactions on Graphics*, **32** : 6 :1–6 :15. URL : http://www.cg.tuwien.ac.at/research/publications/2013/arikan-2013-osn/ (cf. p. 25)

BAY, HERBERT, TINNE TUYTELAARS et LUC VAN GOOL (2006). « Surf : Speeded up robust features ». Dans : *Computer Vision–ECCV 2006*. Springer, p. 404–417 (cf. p. 52)

BAZIN, J, HONGDONG LI, INSO KWEON, CÉDRIC DEMONCEAUX, PASCAL VASSEUR et KATSUSHI IKEUCHI (2012). A Branch and Bound Approach to Correspondence and Grouping Problems. Dans : (cf. p. 182)

BAZIN, JEAN-CHARLES, OLIVIER SAURER, FRIEDRICH FRAUNDORFER et MARC POLLEFEYS (2013). Interactive Omnidirectional Indoor Tour. Dans : *Emerging Technologies for 3D Video : Creation, Coding, Transmission and Rendering*, p. 395–415 (cf. p. 181)

BEARDSLEY, PAUL, PHIL TORR et ANDREW ZISSERMAN (1996). « 3D model acquisition from extended image sequences ». Dans : *Computer Vision—ECCV'96*. Springer, p. 683–695 (cf. p. 29, 92, 94)

BHAT, PRAVIN et SEBASTIAN BURKE (2011). PhotoSpace : a vision based approach for digitizing props. Dans : *ACM SIGGRAPH 2011 Talks*. SIGGRAPH '11. Vancouver, British Columbia, Canada : ACM, 1 :1–1 :1. ISBN : 978-1-4503-0974-5. DOI : 10.1145/2037826.2037828. URL : http://doi.acm.org/10.1145/2037826.2037828 (cf. p. 19)

BRITO, JOSÉ HENRIQUE, ROLAND ANGST, KEVIN KÖSER et MARC POLLEFEYS (2013). Radial Distortion Self-Calibration. Dans : (cf. p. 180)

BROWN, DUANE C (1966). Decentering distortion of lenses. Dans : *Photometric Engineering*, **32** :3, p. 444–462 (cf. p. 40)

— (1976). The bundle adjustment—progress and prospects. Dans : *Int. Archives Photogrammetry*, **21** :3, p. 1–1 (cf. p. 17)

BROWN, M. et D. G. LOWE (2007). Automatic Panoramic Image Stitching using Invariant Features. Dans : *International Journal of Computer Vision*, **74** :1, p. 59–73 (cf. p. 161, 162, 164, 166, 176)

BROWN, M. et D. LOWE (2005a). Unsupervised 3D Object Recognition and Reconstruction in Unordered Datasets. Dans : *5th International Conference on 3D Imaging and Modelling (3DIM05)*. Ottawa, Canada (cf. p. 23, 29)

BROWN, MATTHEW et DAVID G LOWE (2005b). Unsupervised 3D object recognition and reconstruction in unordered datasets. Dans : *3-D Digital Imaging and Modeling, 2005. 3DIM 2005. Fifth International Conference on*. IEEE, p. 56–63 (cf. p. 94, 95)

BURT, PETER J et EDWARD H ADELSON (1983). A multiresolution spline with application to image mosaics. Dans : *ACM Transactions on Graphics (TOG)*, **2** :4, p. 217–236 (cf. p. 164)

BYRÖD, MARTIN, KLAS JOSEPHSON et KALLE ÅSTRÖM (2007). « Fast optimal three view triangulation ». Dans : *Computer Vision–ACCV 2007*. Springer, p. 549–559 (cf. p. 46)

CALONDER, MICHAEL, VINCENT LEPETIT, MUSTAFA OZUYSAL, TOMASZ TRZCINSKI, CHRISTOPH STRECHA et PASCAL FUA (2012). BRIEF : Computing a local binary descriptor very fast. Dans : *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **34** :7, p. 1281–1298 (cf. p. 53)

CHAKRABARTI, AYAN, DANIEL SCHARSTEIN et TODD ZICKLER (2009). An Empirical Camera Model for Internet Color Vision. Dans : *BMVC*. T. 1. 2. Citeseer, p. 4 (cf. p. 168)

CHATTERJEE, AVISHEK et VENU MADHAV GOVINDU (2013). « Efficient and Robust Large-Scale Rotation Averaging ». Dans : *Computer Vision (ICCV), 2013 IEEE International Conference on* (cf. p. 156, 180)

CHOI, JONGMOO et GÉRARD MEDIONI (2009). StaRSaC : Stable random sample consensus for parameter estimation. Dans : *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, p. 675–682 (cf. p. 70)

CHOI, SUNGLOK, TAEMIN KIM et WONPIL YU (2009). Performance Evaluation of RANSAC Family. Dans : *BMVC*, p. 1–12 (cf. p. 69)

CHUM, ONDREJ et JIRI MATAS (2005). Matching with PROSAC-progressive sample consensus. Dans : *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. T. 1. IEEE, p. 220–226 (cf. p. 69)

CHUM, ONDŘEJ, JIŘÍ MATAS et JOSEF KITTLER (2003). « Locally optimized RANSAC ». Dans : *Pattern Recognition*. Springer, p. 236–243 (cf. p. 70)

COLBERT, MARK, JEAN-YVES BOUGUET, JEFF BEIS, SPUDDE CHILDS, DANIEL FILIP, LUC VINCENT, JONGWOO LIM et SCOTT SATKIN (2012). Building indoor multi-

panorama experiences at scale. Dans : *ACM SIGGRAPH 2012 Posters*. ACM, p. 24 (cf. p. 81)

COURCHAY, JÉRÔME, ARNAK S DALALYAN, RENAUD KERIVEN et PETER STURM (2012). On Camera Calibration with Linear Programming and Loop Constraint Linearization. Dans : *International journal of computer vision*, **97** :1, p. 71–90 (cf. p. 121)

CRANDALL, DAVID, ANDREW OWENS, NOAH SNAVELY et DAN HUTTENLOCHER (2011). Discrete-continuous optimization for large-scale structure from motion. Dans : *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, p. 3001–3008 (cf. p. 29, 115)

DALALYAN, ARNAK et RENAUD KERIVEN (2012). Robust estimation for an inverse problem arising in multiview geometry. Dans : *JMIV*, p. 10–23 (cf. p. 120, 128)

DE LUCA, LIVIO (2006). Relevé et multi-représentations du patrimoine architectural ; définition d'une approche hybride de reconstruction 3D d'édifices. Dans : (cf. p. 14, 19)

— (2009). *La photomodélisation architecturale : relevé, modélisation et représentation d'édifices à partir de photographies*. Eyrolles. ISBN : 9782212125245. URL : http://books.google.fr/books?id=-dNoPgAACAAJ (cf. p. 13)

DEBEVEC, PAUL E. et JITENDRA MALIK (1997). Recovering high dynamic range radiance maps from photographs. Dans : *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. SIGGRAPH '97. New York, NY, USA : ACM Press/Addison-Wesley Publishing Co., p. 369–378. ISBN : 0-89791-896-7. DOI : 10.1145/258734.258884 (cf. p. 160)

DEBEVEC, PAUL E., CAMILLO J. TAYLOR et JITENDRA MALIK (1996). Modeling and rendering architecture from photographs : a hybrid geometry- and image-based approach. Dans : *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. SIGGRAPH '96. New York, NY, USA : ACM, p. 11–20. ISBN : 0-89791-746-4. DOI : 10.1145/237170.237191. URL : http://doi.acm.org/10.1145/237170.237191 (cf. p. 21, 22)

DELON, JULIE (2004). Midway image equalization. Dans : *Journal of Mathematical Imaging and Vision*, **21** :2, p. 119–134 (cf. p. 162)

— (2006). Movie and video scale-time equalization application to flicker reduction. Dans : *IEEE Transactions on Image Processing*, **15** :1, p. 241–248 (cf. p. 162)

DESOLNEUX, AGNÈS, LIONEL MOISAN et JEAN-MICHEL MOREL (2000). Meaningful alignments. Dans : *International Journal of Computer Vision*, **40** :1, p. 7–23 (cf. p. 72)

DESOLNEUX, AGNES, LIONEL MOISAN et JEAN-MICHEL MOREL (2007). *From Gestalt theory to image analysis : a probabilistic approach*. 1st. Springer. ISBN : 0387726357, 9780387726359 (cf. p. 72)

DUPAČ, JAN, JIŘÍ MATAS et FILIP NAISER (2012). Ultra-fast tracking based on zero-shift points. Dans : *Image and Vision Computing*, (cf. p. 56)

ENQVIST, OLOF, FREDRIK KAHL et CARL OLSSON (2011). Non-sequential structure from motion. Dans : *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, p. 264–271 (cf. p. 116, 123, 124, 131)

FAUGERAS, OLIVIER D. (1992). What can be seen in three dimensions with an uncalibrated stereo rig. Dans : *Proceedings of the Second European Conference on Computer Vision*. ECCV '92. London, UK, UK : Springer-Verlag, p. 563–578. ISBN : 3-540-55426-2. URL : http://dl.acm.org/citation.cfm?id=645305.648717 (cf. p. 43)

FISCHLER, MARTIN A et ROBERT C BOLLES (1981). Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography. Dans : *Communications of the ACM*, **24** :6, p. 381–395 (cf. p. 67)

FREDMAN, MICHAEL et MICHAEL SAKS (1989). The cell probe complexity of dynamic data structures. Dans : *Proceedings of the twenty-first annual ACM symposium on Theory of computing*. ACM, p. 345–354 (cf. p. 60)

FRITSCH, DIETER (2006). The Photogrammetric Week series–a centennial success story. Dans : (cf. p. 17)

FURUKAWA, YASUTAKA et JEAN PONCE (2010). Accurate, dense, and robust multiview stereopsis. Dans : *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **32** :8, p. 1362–1376 (cf. p. 33)

FURUKAWA, YASUTAKA, BRIAN CURLESS, STEVEN M SEITZ et RICHARD SZELISKI (2010). Towards internet-scale multi-view stereo. Dans : *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, p. 1434–1441 (cf. p. 33)

GALLER, BERNARD A et MICHAEL J FISHER (1964). An improved equivalence algorithm. Dans : *Communications of the ACM*, **7** :5, p. 301–303 (cf. p. 58)

GALVEZ-LOPEZ, DORIAN et JUAN D. TARDOS (2011). Real-time loop detection with bags of binary words. Dans : *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, p. 51 –58. DOI : 10.1109/IROS.2011.6094885 (cf. p. 181)

GAO, XIAO SHAN, XIAO-RONG HOU, JIANLIANG TANG et HANG-FEI CHENG (2003). Complete solution classification for the perspective-three-point problem. Dans : *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **25** :8, p. 930–943. ISSN : 0162-8828. DOI : 10.1109/TPAMI.2003.1217599 (cf. p. 47)

GARETH A. LLOYD, ROCHESTER NY et ROCHESTER NY STEVEN J. SASSON (déc. 1978). *Electronic still camera*. Patent. US 4131919. URL : http://www.patentlens.net/patentlens/patent/US_4131919/en/ (cf. p. 17)

GOVINDU, VENU MADHAV (2001). Combining two-view constraints for motion estimation. Dans : *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. T. 2. IEEE, p. II–218 (cf. p. 115, 117, 118, 122, 133–135)

— (2004). Lie-algebraic averaging for globally consistent motion estimation. Dans : *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. T. 1. IEEE, p. I–684 (cf. p. 115)

— (2006). « Robustness in motion averaging ». Dans : *Computer Vision–ACCV 2006*. Springer, p. 457–466 (cf. p. 116)

GUGAT, MARTIN (1996). A fast algorithm for a class of generalized fractional programs. Dans : *Management Science*, p. 1493–1499 (cf. p. 119)

GUILBERT, NICOLAS, FREDRIK KAHL, M OSKARSSON, K ÅSTRÖM, MARTIN JOHANSSON et ANDERS HEYDEN (2004). Constraint enforcement in structure and motion applied to closing an open sequence. Dans : *Proc. Asian Conf. on Computer Vision, Jeju Island, Korea* (cf. p. 110)

HACOHEN, YOAV, ELI SHECHTMAN, DAN B GOLDMAN et DANI LISCHINSKI (2011). Non-Rigid Dense Correspondence with Applications for Image Enhancement. Dans : *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2011)*, **30** :4, 70 :1–70 :9 (cf. p. 162, 164)

— (2013). Optimizing color consistency in photo collections. Dans : *ACM Transactions on Graphics (TOG)*, **32** :4, p. 38 (cf. p. 164)

HARALICK, ROBERT M. et LINDA G. SHAPIRO (1992). *Computer and Robot Vision*. 1st. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc. ISBN : 0201569434 (cf. p. 97)

HARRIS, CHRIS et MIKE STEPHENS (1988). A combined corner and edge detector. Dans : *Alvey vision conference*. T. 15. Manchester, UK, p. 50 (cf. p. 52)

HARRIS, CHRISTOPHER G et JM PIKE (1988). 3D positional integration from image sequences. Dans : *Image and Vision Computing*, **6** :2, p. 87–90 (cf. p. 92)

HARTLEY, RICHARD I. (juin 1997a). In Defense of the Eight-Point Algorithm. Dans : *IEEE Trans. Pattern Anal. Mach. Intell.*, **19** :6, p. 580–593. ISSN : 0162-8828. DOI : 10.1109/34.601246. URL : http://dx.doi.org/10.1109/34.601246 (cf. p. 43)

HARTLEY, RICHARD I (1997b). Lines and points in three views and the trifocal tensor. Dans : *International Journal of Computer Vision*, **22** :2, p. 125–140 (cf. p. 44)

HARTLEY, RICHARD et FREDERIK SCHAFFALITZKY (2004a). L_∞ minimization in geometric reconstruction problems. Dans : *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. T. 1. IEEE, p. I–504 (cf. p. 46)

— (2004b). L_∞ minimization in geometric reconstruction problems. Dans : *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. T. 1. IEEE, p. I–504 (cf. p. 119)

HARTLEY, RICHARD et ANDREW ZISSERMAN (2000). *Multiple view geometry in computer vision*. T. 2. Cambridge Univ Press (cf. p. 45–47, 83, 109)

HARTLEY, RICHARD, JOCHEN TRUMPF, YUCHAO DAI et HONGDONG LI (2013). Rotation Averaging. Dans : *IJCV*, p. 1–39 (cf. p. 115)

HAVLENA, MICHAL, AKIHIKO TORII et TOMÁŠ PAJDLA (2010). « Efficient structure from motion by graph optimization ». Dans : *Computer Vision–ECCV 2010*. Springer, p. 100–113 (cf. p. 110)

HIEP, VU HOANG, RENAUD KERIVEN, PATRICK LABATUT et J-P PONS (2009). Towards high-resolution large-scale multi-view stereo. Dans : *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, p. 1430–1437 (cf. p. 26)

HORN, BERTHOLD KP et BRIAN G SCHUNCK (1981). Determining optical flow. Dans : *Artificial intelligence*, **17** :1, p. 185–203 (cf. p. 56)

INDELMAN, VADIM, RICHARD ROBERTS, CHRIS BEALL et FRANK DELLAERT (2012). Incremental Light Bundle Adjustment. Dans : *BMVC*, p. 1–11 (cf. p. 110)

JEGOU, HERVE, MATTHIJS DOUZE et CORDELIA SCHMID (2011). Product quantization for nearest neighbor search. Dans : *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **33** :1, p. 117–128 (cf. p. 34)

KAHL, FREDRIK et RICHARD HARTLEY (2008). Multiple-View Geometry Under the {*L_infty*}-Norm. Dans : *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **30** :9, p. 1603–1617 (cf. p. 119)

KALAL, ZDENEK, KRYSTIAN MIKOLAJCZYK et JIRI MATAS (2012). Tracking-learning-detection. Dans : *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **34** :7, p. 1409–1422 (cf. p. 56)

KANATANI, KENICHI, YASUYUKI SUGAYA et HIROTAKA NIITSUMA (2008). Triangulation from two views revisited : Hartley-Sturm vs. optimal correction. Dans : *In practice*, **4** : p. 5 (cf. p. 46)

KIM, S, H LIN, ZHENG LU, SABINE SUSSTRUNK, STEPHEN LIN et M BROWN (2012). A new in-camera imaging model for color computer vision and its application. Dans : (cf. p. 160)

KLOPSCHITZ, MANFRED, CHRISTOPHER ZACH, ARNOLD IRSCHARA et DIETER SCHMALSTIEG (2008). Generalized detection and merging of loop closures for video sequences. Dans : *Proceedings of 3D Data Processing, Visualization, and Transmission* (cf. p. 110)

KNEIP, LAURENT, DAVIDE SCARAMUZZA et ROLAND SIEGWART (2011). A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. Dans : *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, p. 2969–2976 (cf. p. 47)

KNEIP, LAURENT, ROLAND SIEGWART et MARC POLLEFEYS (2012). « Finding the exact rotation between two images independently of the translation ». Dans : *Computer Vision–ECCV 2012*. Springer, p. 696–709 (cf. p. 181)

KRUPPA, E. (1913). *Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung*. Hölder. URL : http://books.google.fr/books?id=t8z5tgAACAAJ (cf. p. 17)

LEE, GIM HEE, FRIEDRICH FRAUNDORFER et MARC POLLEFEYS (2013). Robust Pose-Graph Loop-Closures with Expectation-Maximization. Dans : *Intelligent Robots and Systems (IROS)* (cf. p. 180)

LEPETIT, VINCENT, FRANCESC MORENO-NOGUER et PASCAL FUA (2009). Epnp : An accurate o (n) solution to the pnp problem. Dans : *International Journal of Computer Vision*, **81** :2, p. 155–166 (cf. p. 47)

LI, HONGDONG et RICHARD HARTLEY (2006). Five-point motion estimation made easy. Dans : *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. T. 1. IEEE, p. 630–633 (cf. p. 42)

LINDEBERG, TONY (1998). Feature detection with automatic scale selection. Dans : *International journal of computer vision*, **30** :2, p. 79–116 (cf. p. 52)

LINDSTROM, PETER (2010). Triangulation made easy. Dans : *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, p. 1554–1561 (cf. p. 46)

LIU, ZHE et RENAUD MARLET (2012). Virtual Line Descriptor and Semi-Local Graph Matching Method for Reliable Feature Correspondence. Dans : *BMVC*, p. 1–11 (cf. p. 167)

LONGUET HIGGINS, H.C. (1981). A Computer Algorithm for Reconstructing a Scene from Two Projections. Dans : *Nature*, **293** : (cf. p. 42)

LOURAKIS, MANOLIS et ANTONIS ARGYROS (2004). *The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm*. Rap. tech. Technical Report 340 Institute of Computer Science-FORTH, Heraklion Crete Greece (cf. p. 48, 94, 110)

LOWE, DAVID G (1999). Object recognition from local scale-invariant features. Dans : *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. T. 2. Ieee, p. 1150–1157 (cf. p. 23, 52, 53, 55, 94, 137)

LUCAS, BRUCE D. et TAKEO KANADE (1981). An Iterative Image Registration Technique with an Application to Stereo Vision. Dans : *IJCAI*, p. 674–679 (cf. p. 56)

LUONG, Q.T. (1992). *MATRICE FONDAMENTALE ET AUTOCALIBRATION EN VISION PAR ORDINATEUR*. URL : http://books.google.fr/books?id=jeerNwAACAAJ (cf. p. 43)

MARTINEC, DANIEL et TOMAS PAJDLA (2007). Robust rotation and translation estimation in multiview reconstruction. Dans : *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, p. 1–8 (cf. p. 29, 115, 120, 124)

MIKOLAJCZYK, KRYSTIAN et CORDELIA SCHMID (2001). Indexing based on scale invariant interest points. Dans : *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. T. 1. IEEE, p. 525–531 (cf. p. 52)

— (2005). A performance evaluation of local descriptors. Dans : *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **27** :10, p. 1615–1630 (cf. p. 53)

MOISAN, LIONEL (2003). *Modèles continus, numériques et statistiques pour l'analyse d'images*. Université Paris 11 (cf. p. 72)

MOISAN, LIONEL et BÉRENGER STIVAL (2004). A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. Dans : *International Journal of Computer Vision*, **57** :3, p. 201–218 (cf. p. 30, 69–71, 73, 77–79)

MOISAN, LIONEL, PIERRE MOULON et PASCAL MONASSE (2012). Automatic Homographic Registration of a Pair of Images, with A Contrario Elimination of Outliers.

Dans : *Image Processing On Line*, **2012** : DOI : 10.5201/ipol.2012.mmm-oh (cf. p. 41, 78, 79, 89)

MOULON, PIERRE et PASCAL MONASSE (2012). Unordered feature tracking made fast and easy. Dans : *Conference on Visual Media Production (CVMP12)* (cf. p. 63)

MOULON, PIERRE, PASCAL MONASSE et RENAUD MARLET (2013a). « Adaptive Structure from Motion with a contrario model estimation ». Dans : *Computer Vision–ACCV 2012*. Springer Berlin Heidelberg, p. 257–270 (cf. p. 79, 101, 109)

— (2013b). « Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion ». Dans : *Computer Vision (ICCV), 2013 IEEE International Conference on* (cf. p. 79, 156)

MOULON, PIERRE, BRUNO DUISIT et PASCAL MONASSE (2013c). Global Multiple-View Color Consistency. Dans : *Conference on Visual Media Production (CVMP13)* (cf. p. 176)

MOULON, PIERRE, PASCAL MONASSE et RENAUD MARLET (2013d). La bibliothèque openMVG : open source Multiple View Geometry. Dans : *Orasis, Congrès des jeunes chercheurs en vision par ordinateur* (cf. p. 63, 109)

MUJA, MARIUS et DAVID G LOWE (2009). Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. Dans : *VISAPP (1)*, p. 331–340 (cf. p. 54)

NI, KAI, HAILIN JIN et FRANK DELLAERT (2009). GroupSAC : Efficient consensus in the presence of groupings. Dans : *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, p. 2193–2200 (cf. p. 69)

NISTÉR, DAVID (2004). An Efficient Solution to the Five-Point Relative Pose Problem. Dans : *IEEE Trans. Pattern Anal. Mach. Intell.*, **26** :6, p. 756–777 (cf. p. 42)

NOURY, NICOLAS (oct. 2011). *Mise en correspondance A Contrario de points d'intérêt sous contraintes géométrique et photométrique*. Français. THESE. Université Henri Poincaré - Nancy I. URL : http://tel.archives-ouvertes.fr/tel-00640168 (cf. p. 71, 78, 79)

NOZAWA, KAZUKI, AKIHIKO TORII et MASATOSHI OKUTOMI (2013). « Stable two view reconstruction using the six-point algorithm ». Dans : *Computer Vision–ACCV 2012*. Springer, p. 122–135 (cf. p. 110)

OLSSON, CARL et OLOF ENQVIST (2011). « Stable structure from motion for unordered image collections ». Dans : *Image Analysis*. Springer, p. 524–535 (cf. p. 29, 116, 120, 122, 128, 130, 131, 139, 143, 144, 156)

PAGANI, ALAIN et DIDIER STRICKER (2011). Structure from Motion using full spherical panoramic cameras. Dans : *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, p. 375–382 (cf. p. 81, 86, 181)

PHOTOGRAPHIQUES, LES NOUVEAUTÉS (1910). *Le pigeon voyageur photographe*. URL : \url{http://gallica.bnf.fr/ark:/12148/bpt6k54904907.image.r=Neubronner.f70.vignettesnaviguer.langEN} (cf. p. 17)

POLLEFEYS, MARC, REINHARD KOCH, MAARTEN VERGAUWEN, ALBERT A DEKNUYDT et LUC J VAN GOOL (2000). Three-dimensional scene reconstruction from images. Dans : *Electronic Imaging*. International Society for Optics et Photonics, p. 215–226 (cf. p. 23, 29)

RABIN, JULIEN (déc. 2009). *Approches robustes pour la comparaison d'images et la reconnaissance d'objets*. Français. THESE. Télécom ParisTech. URL : http://tel.archives-ouvertes.fr/tel-00472442 (cf. p. 30, 55, 78, 79)

RABIN, JULIEN, JULIE DELON et YANN GOUSSEAU (2008). A contrario matching of sift-like descriptors. Dans : *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, p. 1–4 (cf. p. 180)

RAGURAM, RAHUL et J-M FRAHM (2011). RECON : Scale-adaptive robust estimation via Residual Consensus. Dans : *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, p. 1299–1306 (cf. p. 70)

REINHARD, ERIK, MICHAEL ADHIKHMIN, BRUCE GOOCH et PETER SHIRLEY (2001). Color transfer between images. Dans : *Computer Graphics and Applications, IEEE*, **21** :5, p. 34–41 (cf. p. 161, 176)

RODRIGUEZ, ANTONIO L, PEDRO E LOPEZ-DE TERUEL et ALBERTO RUIZ (2011). Reduced epipolar cost for accelerated incremental sfm. Dans : *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, p. 3097–3104 (cf. p. 110, 118)

ROUSSEEUW, PETER J (1984). Least median of squares regression. Dans : *Journal of the American statistical association*, **79** :388, p. 871–880 (cf. p. 69)

SCHAFFALITZKY, FREDERIK et ANDREW ZISSERMAN (2002). Multi-view Matching for Unordered Image Sets, or "How Do I Organize My Holiday Snaps ?" Dans : *ECCV (1)*, p. 414–431 (cf. p. 94)

SEO, YONGDUEK et RICHARD I. HARTLEY (2007). A Fast Method to Minimize L_∞ Error Norm for Geometric Vision Problems. Dans : *ICCV* (cf. p. 128)

SHARP, GREGORY C, SANG WOOK LEE et DAVID K WEHE (2001). Toward multiview registration in frame space. Dans : *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*. T. 4. IEEE, p. 3542–3547 (cf. p. 115)

SHUM, HEUNG-YEUNG, MEI HAN et RICHARD SZELISKI (1998). Interactive construction of 3D models from panoramic mosaics. Dans : *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*. IEEE, p. 427–433 (cf. p. 81)

SIM, KRISTY et RICHARD HARTLEY (2006). Recovering camera motion using linfty minimization. Dans : *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. T. 1. IEEE, p. 1230–1237 (cf. p. 117, 118, 122, 128, 134, 135, 142)

SINHA, SUDIPTA N, DREW STEEDLY, RICHARD SZELISKI, MANEESH AGRAWALA et MARC POLLEFEYS (2008). Interactive 3D architectural modeling from unordered photo collections. Dans : *ACM Transactions on Graphics (TOG)*. T. 27. 5. ACM, p. 159 (cf. p. 25)

SLAMA, CHESTER C, CHARLES THEURER, SOREN W HENRIKSEN et al. (1980). *Manual of photogrammetry*. Ed. 4. American Society of photogrammetry (cf. p. 17)

SNAVELY, NOAH, STEVEN M. SEITZ et RICHARD SZELISKI (2006). Photo tourism : Exploring photo collections in 3D. Dans : *SIGGRAPH Conference Proceedings*. New York, NY, USA : ACM Press, p. 835–846. ISBN : 1-59593-364-6 (cf. p. 23, 24, 29, 33, 57, 60, 62, 95, 97, 101, 102, 110, 143)

SNAVELY, NOAH, STEVEN M SEITZ et RICHARD SZELISKI (2008). Skeletal graphs for efficient structure from motion. Dans : *CVPR*. T. 1, p. 2 (cf. p. 110)

STEWART, CHARLES V. (1995). MINPRAN : A new robust estimator for computer vision. Dans : *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **17** :10, p. 925–938 (cf. p. 70)

STRECHA, CHRISTOPH, WOLFGANG VON HANSEN, LUC VAN GOOL, PASCAL FUA et ULRICH THOENNESSEN (2008). On benchmarking camera calibration and multi-view stereo for high resolution imagery. Dans : *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, p. 1–8 (cf. p. 27, 29, 101, 142)

STRECHA, CHRISTOPH, TIMO PYLVANAINEN et PASCAL FUA (2010). Dynamic and scalable large scale image reconstruction. Dans : *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, p. 406–413 (cf. p. 29)

STRECHA, CHRISTOPH, ALEXANDER M BRONSTEIN, MICHAEL M BRONSTEIN et PASCAL FUA (2012). LDAHash : Improved matching with smaller descriptors. Dans :

Pattern Analysis and Machine Intelligence, IEEE Transactions on, **34** :1, p. 66–78 (cf. p. 53)

SVARM, LINUS, ZHAYIDA SIMAYIJIANG, OLOF ENQVIST et CARL OLSSON (2012). Point track creation in unordered image collections using Gomory-Hu trees. Dans : *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, p. 2116–2119 (cf. p. 63)

TARJAN, ROBERT ENDRE (1975). Efficiency of a good but not linear set union algorithm. Dans : *Journal of the ACM (JACM)*, **22** :2, p. 215–225 (cf. p. 59)

TOLA, ENGIN, CHRISTOPH STRECHA et PASCAL FUA (2012). Efficient large-scale multi-view stereo for ultra high-resolution image sets. Dans : *Machine Vision and Applications*, **23** :5, p. 903–920 (cf. p. 34)

TOLDO, ROBERTO et ANDREA FUSIELLO (2008). « Robust multiple structures estimation with j-linkage ». Dans : *Computer Vision–ECCV 2008*. Springer, p. 537–547 (cf. p. 34)

TOMASI, CARLO et TAKEO KANADE (1991). Detection and Tracking of Point Features. Dans : *Order A Journal On The Theory Of Ordered Sets And Its Applications*, **37** :7597, p. 165–168. URL : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.5899&rep=rep1&type=pdf (cf. p. 56)

TORII, AKIHIKO, ATSUSHI IMIYA et NAOYA OHNISHI (2005). Two-and three-view geometry for spherical cameras. Dans : *Proceedings of the sixth workshop on omnidirectional vision, camera networks and non-classical cameras*. Citeseer (cf. p. 81)

TORR, P. H. S. et D. W. MURRAY (1997). The development and comparison of robust methods for estimating the fundamental matrix. Dans : *International Journal of Computer Vision*, **24** : p. 271–300 (cf. p. 43)

TORR, PHILIP HS et ANDREW ZISSERMAN (2000). MLESAC : A new robust estimator with application to estimating image geometry. Dans : *Computer Vision and Image Understanding*, **78** :1, p. 138–156 (cf. p. 69)

TRIGGS, BILL, PHILIP F MCLAUCHLAN, RICHARD I HARTLEY et ANDREW W FITZGIBBON (2000). « Bundle adjustment—a modern synthesis ». Dans : *Vision algorithms : theory and practice*. Springer, p. 298–372 (cf. p. 17, 48)

WOCHENBLATT (1867). *Die Photogrammetrie*. Wochenblatt des Architektenverein zu Berlin. Beelitz. URL : http://books.google.de/books?id=mHc_AAAAcAAJ (cf. p. 17)

WU, CHANGCHANG (2007). SiftGPU : A GPU implementation of scale invariant feature transform (SIFT). Dans : *URL http ://cs. unc. edu/~ ccwu/siftgpu*, (cf. p. 101)

— (2013). Towards Linear-time Incremental Structure from Motion. Dans : *3DV*, (cf. p. 29, 101, 102, 110, 111, 143)

WU, CHANGCHANG, SAMEER AGARWAL, BRIAN CURLESS et STEVEN M SEITZ (2011a). Multicore bundle adjustment. Dans : *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, p. 3057–3064 (cf. p. 29, 48, 101, 110)

WU, LUN, ARVIND GANESH, BOXIN SHI, YASUYUKI MATSUSHITA, YONGTIAN WANG et YI MA (2011b). « Robust photometric stereo via low-rank matrix completion and recovery ». Dans : *Computer Vision–ACCV 2010*. Springer, p. 703–717 (cf. p. 16)

XU, WEI et JANE MULLIGAN (2010). Performance evaluation of color correction approaches for automatic multi-view image and video stitching. Dans : *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, p. 263–270 (cf. p. 161)

ZACH, CHRISTOPHER (2010-2011). *ETH-V3D Structure-and-Motion software. ETH Zurich* (cf. p. 60, 62)

ZACH, CHRISTOPHER et MARC POLLEFEYS (2010). « Practical methods for convex multi-view reconstruction ». Dans : *Computer Vision–ECCV 2010*. Springer, p. 354–367 (cf. p. 120)

ZACH, CHRISTOPHER, MANFRED KLOPSCHITZ et MANFRED POLLEFEYS (2010). Disambiguating visual relations using loop constraints. Dans : *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, p. 1426–1433 (cf. p. 117, 123, 124, 138)

ZHOU, QIAN-YI et VLADLEN KOLTUN (2013). Dense Scene Reconstruction with Points of Interest. Dans : *Siggraph 2013*, (cf. p. 181)