Mise en correspondance a contrario de points d'intérêt sous contraintes géométrique et photométrique

Nicolas Noury

Nancy-Université
Université Henri Poincaré

Département de formation doctorale en informatique **École doctorale IAEM Lorraine**
UFR Sciences et Technologies

Mise en correspondance A Contrario de points d'intérêt sous contraintes géométrique et photométrique

THÈSE

présentée et soutenue publiquement le 13 Octobre 2011

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1
(spécialité informatique)

par

Nicolas NOURY

Composition du jury

Rapporteurs :	Adrien BARTOLI	Professeur, Université d'Auvergne, Clermont-Ferrand
	Charles KERVRANN	Directeur de Recherche INRIA, Rennes
Examinateurs :	Yann GOUSSEAU	Maître de Conférences, ENST, Paris
	Dominique MÉRY	Professeur, UHP, Nancy
	Frédéric SUR	Maître de Conférences, INPL, Nancy
	Marie-Odile BERGER	Chargée de Recherche INRIA, Nancy

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Mis en page avec la classe thloria.

Remerciements

Je tiens en premier lieu à remercier Marie-Odile Berger, ainsi que Frédéric Sur, qui m'ont guidé pendant ces années de travail. D'abord, pour m'avoir initié à la recherche en vision par ordinateur pendant mon stage de master, et proposé, puis convaincu, de faire une thèse. Leur disponibilité et leur soutien m'ont permis de mener à bien ce projet.

Je remercie également les rapporteurs de ce mémoire, Charles Kervrann et Adrien Bartoli ; ainsi que Yann Gousseau qui a accepté d'être examinateur. Enfin, Dominique Méry pour l'honneur qu'il m'a fait de présider le jury de thèse.

Ensuite, un grand merci aux collègues et amis du Loria, plus particulièrement de l'équipe Magrit : Erwan, Brigitte, Flavio, Diego, Sébastien, Nicolas P., Nicolas F., Frédéric Sp., Evren, Patricia, Cédric, Ting, Blaise, Pierre-Frédéric, Ahmed, Christel.

Un énorme merci particulier à mon grand frère de thèse, Michael, et à Lise, qui en plus de surmonter les aléas des voyages en train, savent monter des guets-apens inoubliables.

Merci bien entendus aux nancéiens, et ceux qui le sont de moins et moins : Aude et Adrien, et leurs biens jolis t-shirts californiens. Céline, Gilles le dessinateur voileux marathonien et Zoé pour leurs barbecues toujours sous le soleil. Landry, mon photographe forgeron préféré à qui je dois un nombre incalculable de sorties photos repoussées. Xevel, l'homme fuseau horaire qui parlait à l'oreille des Xachicomas. Jlg, qui s'assure que la montagne est toujours là à ma place. Aurélie, Alex et Apolline, envoyés spéciaux en Bavière.

Merci aussi à tous ceux qui n'ont pas besoins d'être cités car présent depuis très très longtemps : à Benoît, le frère que j'aurai rêvé avoir. Aux condorcéens, petits et grands : Alice, Béné, Florence, Laure, Benjamin, Fabien, Julien, et Yann, pour tous les moments partagés passés et à venir. Merci aux globe-trotteurs, toujours plus nombreux, pour leurs chambres d'amis que je compte bien utiliser.

Merci aussi à tous ceux que j'ai oublié.

Merci pour tout à mes parents.

Enfin, merci à Élise.

Table des matières

Introduction

Les progrès en vision par ordinateur, branche de l'intelligence artificielle, permettent aux machines d'extraire de l'information à partir d'images. Diverses approches sont apparues, permettant des applications aussi variées que la reconnaissance d'objets ou de personnes, la construction de modèles virtuels, le positionnement de robots mobiles... Les travaux de cette thèse sont relatifs au problème de l'*analyse de la structure et du mouvement*, et plus particulièrement à la mise en correspondance de points d'une image à l'autre. Il s'agit d'estimer la forme d'objets 3D et la position de la caméra à partir de photos ou de vidéos. Leur acquisition est accessible au plus grand nombre via les téléphones portables et les ordinateurs personnels. Les applications sont multiples : réalité augmentée par l'incrustation d'objets virtuels dans une scène réelle, post-production vidéo, aide au diagnostic médical, modélisation architecturale...

Des outils, tel que Bundler [Sna, SSS07] et Boujou [vic00] permettent l'estimation de trajectoires de cameras. D'autres, Autopano [Kol04] et Autostitch [Inc09], rendent possibles la réalisation de panoramas. Ces logiciels utilisent des points mis en correspondance, ou appariements, entre deux ou plusieurs images. Comment procéder pour associer deux points est l'enjeu principal de notre travail. Dans des cas difficiles, comme en présence de motifs répétés ou de forts changements de points de vue, peu ou pas d'appariements corrects sont trouvés. L'amélioration de leur fiabilité est une condition importante pour parfaire l'estimation du mouvement. Parmi les nombreuses étapes nécessaires à la résolution de ce problème, nous nous sommes intéressés dans cette thèse à la mise en correspondance de points de deux images distinctes d'une même scène, sous contraintes géométrique et photométrique. La contrainte géométrique est exprimée soit à l'aide de la géométrie projective reliant deux vues, appelée géométrie épipolaire, soit à l'aide d'une homographie planaire. La contrainte photométrique impose la ressemblance entre les voisinages de certains points que l'on va privilégier.

Le mouvement de la caméra entre deux positions est estimé en prenant en compte les contraintes photométriques et géométriques précédentes, et en observant le déplacement des points considérés dans les images en tenant compte du phénomène de parallaxe. Le procédé utilisé habituellement repose sur les étapes suivantes.

Tout d'abord, l'extraction de points d'intérêt sélectionne des points facilement repérables dans une image. Il est souhaitable de ne considérer qu'un nombre limité de points d'une image autant pour des raisons de combinatoire que pour ne conserver que les plus pertinents. Leurs positions et caractéristiques sont les résultats d'un algorithme détecteur. L'information contenue dans les niveaux de gris des pixels de l'image, qu'on appelle photométrie, est pour chaque point représentée par un descripteur encodant celle de son voisinage. Les travaux de Schmid [Sch96], puis les détecteurs et descripteurs SIFT [Low04] et SURF [BTG06], par exemple, ont permis un saut qualitatif permettant d'envisager des utilisations de ceux-ci dans des cas difficiles de changements d'illumination et / ou de point de vue importants.

Ensuite la recherche d'appariements, ou mise en correspondance de points d'intérêt d'une image à l'autre d'une même scène observée avec un point de vue différent, est effectuée à l'aide des descripteurs photométriques.

Puis, on détermine les paramètres de la matrice de mouvement [Zha98, HZ00, FLP01, TM97, TFZ99], soit dans le cas de la géométrie épipolaire, où sont calculés les paramètres de la matrice fondamentale encodant le mouvement de translation et de rotation séparant les deux vues à partir des coordonnées des points d'intérêt mis en correspondance, soit en restreignant la mise en correspondance à deux plans, la relation à estimer étant alors une homographie.

Il est en même temps nécessaire de réaliser un filtrage des correspondances. Les appariements extraits par les méthodes citées ci-dessus ne sont pas tous exacts. L'utilisation de procédés capables de traiter la présence de données aberrantes, tel RANSAC [FB81], permet de filtrer les données afin de ne prendre en compte dans l'évaluation du mouvement que celles jugées fiables, et ainsi de rendre robuste l'estimation statistique subséquente. RANSAC nécessite la fixation de deux seuils qui ont une forte influence sur le résultat, l'un pour la précision des appariements retenus, et l'autre pour la taille minimale de l'ensemble retenu.

Enfin, à partir de ces données filtrées, les informations concernant le positionnement de la caméra sont obtenues.

Enjeux

Les méthodes de type RANSAC nécessitent la fixation de seuils par l'utilisateur et n'opèrent pour le moment que sur les coordonnées des points, c'est-à-dire des informations géométriques. Nous avons retenu la méthodologie *A Contrario* [DMM00, DMM08, CLM+08] pour réaliser la mise en correspondance robuste de points d'intérêt [MS04b]. Son fonctionnement résout la principale difficulté de la recherche dans l'approche RANSAC, définir automatiquement des seuils pour la taille des ensembles retenus et la précision de ceux-ci, tout en minimisant l'erreur d'estimation.

La formalisation du processus de détection est la suivante. On suppose disposer d'un modèle a minima de détection des points d'intérêt. Par exemple, on choisit une loi uniforme expliquant leur distribution dans les images. On évalue ensuite l'espérance d'une détection à tort de l'événement observé pour ce modèle, en prenant en compte la combinatoire du problème et en normalisant l'erreur de mesure. Si l'espérance est très faible, il est très peu probable que cette détection soit fausse, et *a contrario*, il existe une meilleure explication que l'hypothèse naïve sur les données (le hasard). Les avantages sont qu'elle permet une amélioration de la sélection robuste des correspondances et de fixer automatiquement les seuils de l'approche RANSAC.

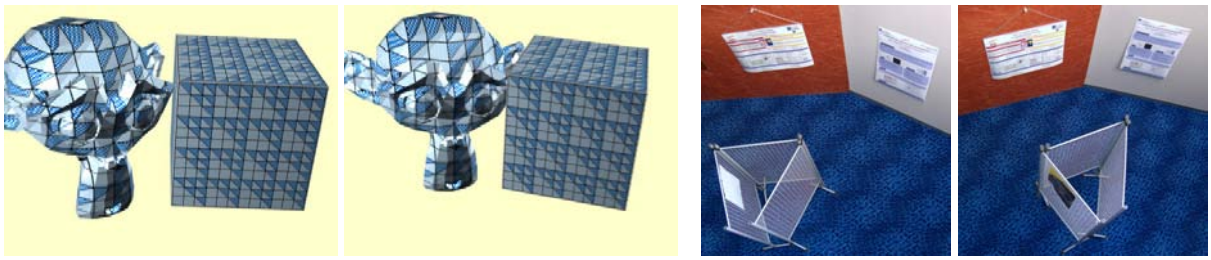

FIGURE 1 – *Couples d'images difficiles à mettre en correspondance.*

Cependant, utiliser cette approche ne modifie pas le paradigme usuel de recherche d'appariements en deux étapes : la mise en correspondance de points d'intérêt par un critère photométrique,

suivi par un filtrage robuste à l'aide d'une contrainte géométrique choisie. Quand la scène contient des motifs répétés, comme pour les images de la figure 1, les candidats à l'appariement retenus à l'étape de comparaison des descripteurs photométriques ne sont pas cohérents, et ceux-ci ne permettent pas de valider une hypothèse de mouvement. La figure 2 en est un bon exemple : pour chaque point, on choisit comme correspondant le point avec le descripteur le plus ressemblant, et seulement s'il se détache bien du second plus proche voisin. Sur la moquette, de nombreux motifs sont identiques et mal mis en correspondances. Vu que la plupart des candidats sont faux, il est alors quasi impossible de filtrer un ensemble cohérent avec le mouvement de la caméra.

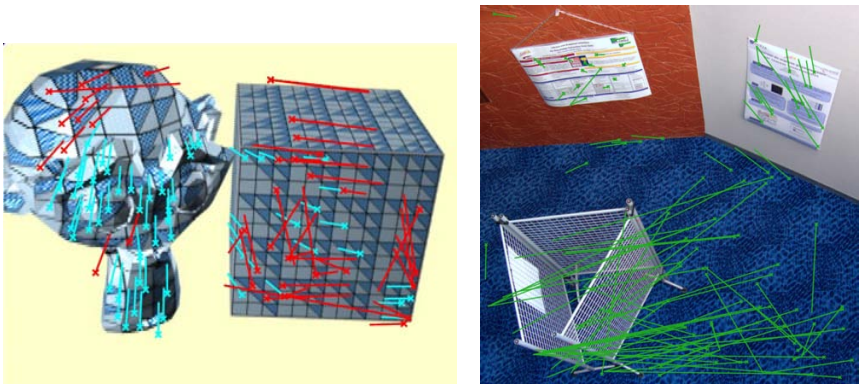

FIGURE 2 – *Approche standard pour la mise en correspondance* pour les couples d'images visible en 1. Elle produit sur le cube ou la moquette des correspondances de points d'intérêt pour la plupart incorrectes. La transformation géométrique recherchée entre les images est une homographie. Les points d'intérêt SIFT sont signalés par une croix, le segment représente le mouvement apparent avec le point d'intérêt apparié dans l'autre image. A gauche, en bleu, les correspondances correctes (*inliers*) après un filtrage robuste de type RANSAC, en rouge, celles non retenues, considérées aberrantes (*outliers*). A droite, en vert, les correspondances avant filtrage robuste, celui-ci ne produisant rien de pertinent dans ce cas.

Plutôt que séparer la recherche en deux étapes, en élaborant une liste de couples 1-1 de candidats à l'appariement utilisant uniquement l'information photométrique, qui est ensuite filtrée, nous proposons de considérer des candidats de 1 point dans une image vers N dans l'autre. Notre schéma statistique détermine ainsi les appariements en une seule étape, en prenant en compte simultanément les contraintes photométriques et géométriques, ce qui permet ainsi de lever l'ambiguïté quand plusieurs zones différentes ont la même apparence. Lever cette limitation en évitant cette sélection a priori des appariements est l'objet de notre première contribution. On peut voir le résultat sur la figure 4 et noter qu'avec ces ensembles de candidats, on augmente le nombre d'appariements finalement extraits.

Une deuxième difficulté de la mise en correspondance est la baisse de ressemblance des descripteurs quand les changements de point de vue sont plus forts. Les solutions existantes de la littérature, qui utilisent des points d'intérêt invariants aux transformations affines, n'ont qu'une invariance limitée aux changements de points de vue [MP07]. Les techniques de simulation de points de vue, et notamment ASIFT [MY09b], permettent d'obtenir une meilleure résistance à ces forts changements d'apparence, mais ne se comportent pas bien en présence de motifs répétés, comme montré en figure 3. Lever cette limitation est le sujet de la seconde contribution. Nous y utilisons l'approche en une étape de notre contribution précédente sous contrainte homographique, ce qui permet d'ob-

tenir des appariements fiables pour de nombreuses zones planaires différentes, et de retrouver le mouvement global de la caméra dans la scène.

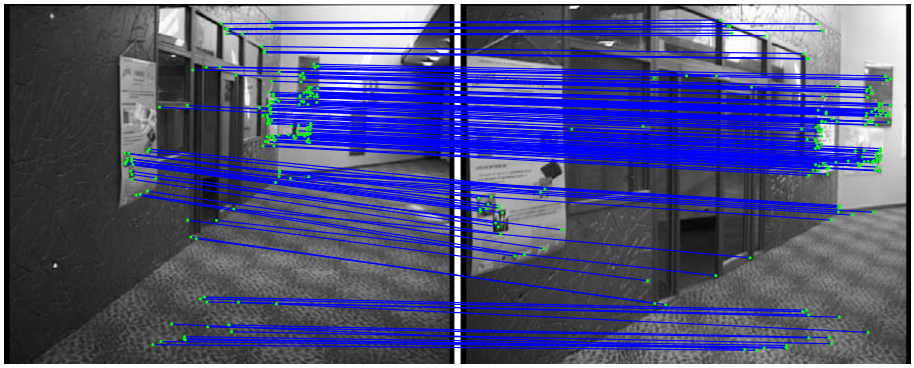

FIGURE 3 – *Changement de point de vue plus important et motifs répétés*. ASIFT [MY09b]. La transformation géométrique recherchée est une matrice fondamentale. Des points de la moquette au premier plan ne sont pas mis en correspondance correctement : les motifs répétés sont à proximité des droites épipolaires associées.

Les contributions de cette thèse

Le point de départ de ce travail est l'approche *A Contrario* RANSAC de Moisan et Stival [MS04b] qui permet la sélection robuste de points mis en correspondance dans le cas de la géométrie épipolaire. La première partie de mon travail a consisté en l'évaluation [NSB07b] de cette méthode de mise en correspondance robuste entre deux vues. Celle-ci s'inscrit parmi la riche littérature des améliorations de la méthode de sélection robuste RANSAC, et est, à notre connaissance, la seule qui permet de s'abstraire totalement de seuils et de paramètres ad-hoc.

Nous présentons ensuite **notre première contribution, un modèle a contrario qui réalise la mise en correspondance à l'aide des critères photométrique et géométrique, ainsi que le filtrage robuste en une seule étape** [NSB07a, NSB10b, NSB10a]. Cette méthode permet de lever l'ambiguïté existant habituellement, en examinant des candidats à l'appariement autres que ceux plus proches voisins pour la ressemblance photométrique, et de mettre en correspondance des scènes contenant des motifs répétés, comme visible en figure 4. Nous montrons son efficacité pour plusieurs modèles géométriques différents d'explication du mouvement, l'homographie et la géométrie épipolaire. Nous étudions aussi le rôle de la distance photométrique utilisée. En effet, la ressemblance photométrique entre voisinages de points est évaluée en comparant des descripteurs à base d'histogrammes, étape souvent réalisée dans la littérature à l'aide de la distance euclidienne. Celle-ci ne produit pas toujours la distance la plus faible pour les appariements les plus ressemblants, et peut être remplacée par des distances plus performantes, comme les distances L_1, du χ^2, ou du terrassier.

Notre second apport [NSB10c] consiste en l'amélioration d'une méthode de mise en correspondance de points invariante aux fortes déformations affines, ASIFT [MY09b], qui génère les points d'intérêt et les descripteurs associés par simulation de point de vue, et permet la mise en correspondance pour de forts changements de point de vue. Notre extension **remplace le schéma de mise en correspondance et améliore la sélection des paires d'images simulées pour fournir des correspondances plus nombreuses et plus densément réparties. Elle permet en outre de**

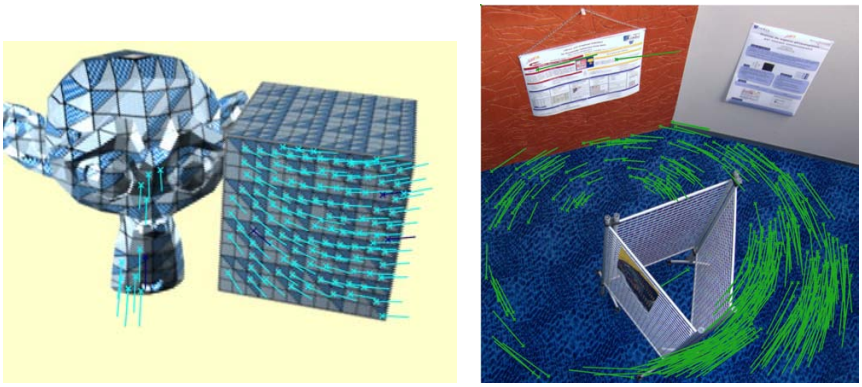

FIGURE 4 – *Notre première contribution* permet d'extraire des appariements corrects en dépit d'une forte ambiguïté perceptuelle, grâce à la prise en compte simultanée des contraintes photométriques et géométriques. Ici, contrairement à la figure 2 des appariements cohérents avec une homographie planaire sont retrouvés entre les images. A gauche, sur la face avant du cube et sur la partie du singe alignée avec elle se trouvant à la même profondeur. A droite, sur la moquette.

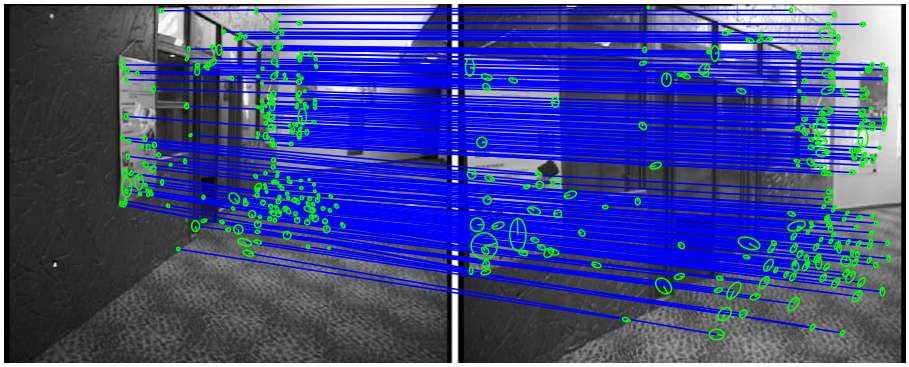

FIGURE 5 – *Notre seconde contribution :* I-ASIFT. Cette fois, comparée à la figure 3, les points sont correctement mis en correspondance sur la moquette.

mettre en correspondance des scènes contenant de nombreux motifs répétés, par exemple des scènes d'intérieur comme en figure 5 ou d'extérieur en milieu urbain, dans lesquelles de telles zones peuvent être observées sous des angles très différents.

Plan

L'organisation du manuscrit est la suivante : après la présente introduction, le chapitre 1, en préambule, décrit les enjeux de la prise de décision en vision par ordinateur via l'utilisation des tests statistiques, et présente le cadre *a contrario*. Nous le décrivons d'abord pour la détection d'alignements, puis pour la sélection robuste d'appariements de points d'intérêt visibles dans deux vues différentes à l'aide de la géométrie épipolaire.

Le chapitre 2 décrit ensuite l'état de l'art des méthodes d'appariements de points d'intérêt, avec en particulier quelques points importants sur leur détection et l'extraction de descripteurs photométriques associés, les distances utilisées pour comparer ces descripteurs afin de réaliser la mise

en correspondance, et enfin les méthodes de filtrage robuste de type RANSAC. Suit une validation de la méthode de filtrage robuste *a contrario* de Moisan et Stival dans le chapitre 3, qui montre un net progrès en robustesse par rapport aux approches plus classiques, mais qui reste néanmoins impuissante dans des cas plus difficiles, en particulier dans des scènes contenant des motifs répétés et ambigus, ou présentant de forts changements de points de vue.

Avec le chapitre 4, après un état de l'art sur les méthodes de mise en correspondance, nous présentons notre première contribution. Nous décrivons le modèle *a contrario* utilisé, ainsi que les modélisations des contraintes géométriques et photométriques. Nous détaillons ensuite les choix techniques réalisés, et terminons avec la présentation de l'algorithme. Les résultats associés sont le sujet du chapitre 5 : ils montrent des résultats d'appariements dans des scènes contenants des motifs répétés et ambigus, souvent visibles dans des environnements réels urbains ou en intérieur.

Notre seconde contribution est traitée au chapitre 6. Nous rappelons les différentes approches existantes visant à obtenir des points d'intérêt invariants aux transformations affines. Nous présentons ensuite l'approche ASIFT et expliquons comment la modifier pour bénéficier à la fois de bonnes performances vis à vis des forts changements de point de vue et de la présence de motifs répétés dans les images.

Enfin, une conclusion revient sur les idées importantes de notre travail et les perspectives qu'il ouvre.

Chapitre 1

Préambule
La méthodologie *a contrario*

Nous présentons dans ce préambule le cadre de la décision a contrario *: celui-ci, présenté pour la première fois par Desolneux, Moisan et Morel pour la détection d'alignements significatifs dans les images [DMM00], est utilisé à de nombreuses reprises dans cette thèse. L'organisation de ce chapitre est la suivante. D'abord est posé le cadre des tests d'hypothèses pour la prise de décision en vision par ordinateur en section 1.1. Puis, nous présentons la méthodologie* a contrario *en section 1.2 à l'aide de son application à la détection d'alignement. Enfin, en section 1.3, nous décrivons la méthode de filtrage robuste* a contrario *des appariements de points d'intérêt, présentée par Moisan et Stival dans [MS04b], puis son adaptation à la détection d'occurrences multiples par Rabin et al. [RDGM10].*

1.1 La prise de décision en vision par ordinateur

Les méthodes d'analyse d'images, telle la détection de primitives comme des droites, ou la reconnaissance d'objet font l'objet d'un intérêt toujours renouvelé depuis les premiers balbutiements du domaine il y a maintenant une quarantaine d'années. Lorsque l'on dispose d'un ensemble d'observations, c'est-à-dire des mesures réalisées dans une image, et que l'on cherche à valider la cohérence de celles-ci avec un modèle, plusieurs approches d'analyse et de décision ont été proposées. Duda et Hart [DHS00] présentent de nombreuses méthodes de reconnaissance de formes : la théorie de la décision bayésienne, l'estimation au maximum de vraisemblance, ainsi que celle au maximum a posteriori, ainsi que de nombreuses méthodes d'apprentissage.

L'analyse d'images ou de séquences vidéo utilisée dans cette thèse repose sur le schéma habituel suivant :

1. Des traitements bas-niveau (calcul d'invariants dans les images, statistiques sur les données d'intensité pixels, recherche d'éléments structurants tels des contours, extraction de points d'intérêt) réduisent la taille des données initiales et produisent des *descripteurs*, dont les propriétés dépendent de l'objectif recherché : invariance, répétabilité, représentation locale ou globale...
2. Ces *descripteurs* peuvent servir à calculer les paramètres d'un modèle, comme les rotations et translations pour un déplacement rigide dans les cas de l'estimation du mouvement, ou encore à prendre une décision de classification lorsque l'on cherche à réaliser de la reconnaissance d'objets.

Dans *Computer Vision : A Modern Approach* [FP02], ces étapes sont décrites comme une approche en cascade, où chaque étape filtre les informations transmises à l'étape suivante : à chaque fois, une décision est prise, souvent via une comparaison avec un seuil. Justifier les choix de seuil est difficile, et les informations non retenues ne sont plus disponibles pour les étapes suivantes.

Forsyth et Ponce plaident pour l'utilisation d'approches probabilistes qui permettent de mesurer la confiance dans les données manipulées, et non de les rejeter une fois pour toute lorsqu'on ne les considère pas suffisamment fiables. Cependant, le volume de données conservées lorsque l'on n'utilise pas les approches en cascade reste pour eux un point bloquant.

La prise de décision peut être formalisée dans le cadre des tests statistiques d'hypothèses.

1.1.1 Tests d'hypothèses

Les approches classiques pour résoudre les problèmes de prise de décision sont décrites par la théorie des tests statistiques. Le test d'hypothèse est le plus souvent binaire : une hypothèse $\mathscr{H}_0$, appelée hypothèse nulle, est l'hypothèse que l'on souhaite confirmer ou rejeter. Lorsqu'elle est rejetée, c'est l'hypothèse alternative, $\mathscr{H}_1$ qui est considérée valide. Les tests statistiques permettent de choisir une des deux hypothèses et d'évaluer le risque d'erreur associé à cette décision. La table de vérité se présente alors par le tableau suivant (d'après [Sap90]) :

Réalité \ Décision	$\mathscr{H}_0$	$\mathscr{H}_1$
$\mathscr{H}_0$	vrai positif	faux négatif (risque de première espèce)
$\mathscr{H}_1$	faux positif (risque de seconde espèce)	vrai négatif (puissance)

Lorsque l'on connaît les vraisemblances, O étant l'observation, $\Pr(O|\mathscr{H}_0)$ et $\Pr(O|\mathscr{H}_1)$, plusieurs méthodes existent afin de déterminer quelle hypothèse choisir.

- L'estimation au maximum de vraisemblance : pour $\Pr(O|\mathscr{H}_0) > \Pr(O|\mathscr{H}_1)$, on choisit $\mathscr{H}_0$.
- La méthode de Neyman et Pearson permet de fixer un seuil S maximisant la puissance pour une valeur fixée du risque de première espèce. Si

$$\frac{\Pr(O|\mathscr{H}_0)}{\Pr(O|\mathscr{H}_1)} \geqslant S,$$

  alors on accepte $\mathscr{H}_0$.
- La théorie de l'inférence bayésienne permet de réaliser un compromis entre la probabilité d'un faux positif et d'une non-détection. Le test de Bayes s'écrit, lorsque l'on connaît les priors $\Pr(\mathscr{H}_0)$ et $\Pr(\mathscr{H}_1)$: si

$$\Pr(\mathscr{H}_0)\Pr(O|\mathscr{H}_0) \geqslant \Pr(\mathscr{H}_1)\Pr(O|\mathscr{H}_1),$$

  alors on accepte $\mathscr{H}_0$.

1.1.2 Limites

En pratique, on ne connaît ni $\Pr(O|\mathscr{H}_0)$, ni les probabilités *a priori* $\Pr(\mathscr{H}_0)$ et $\Pr(\mathscr{H}_1)$. Par exemple, si l'on cherche à détecter des segments de droites dans une image, on peut difficilement présumer de leur nombre. Celui-ci est ainsi très différent entre une scène urbaine et un paysage naturel. Définir des lois *a priori* dépend fortement du type d'images considérées et semble alors arbitraire. De plus, il est difficile de préjuger de la robustesse des traitements bas-niveaux, ceux-ci

étant souvent perturbés par la quantité de bruit présente dans les images, par la taille relative des structures observées par rapport à la résolution des images obtenues, et toutes les déformations dues à la forme du capteur et au dispositif physique de formation de l'image.

L'approche la plus simple pour prendre une décision est de réaliser une comparaison avec un seuil. Par exemple, lorsque l'on détecte des alignements, pour une droite donnée, un seuil limitant la différence d'orientation entre le segment et la droite qu'on accepte pour les considérer alignés. Plutôt que fixer un seuil par rapport à une distance, ce qui est difficile en général car souvent dépendant du contexte, il est possible d'affiner le critère de prise de décision en prenant en compte l'incertitude d'estimation des paramètres du modèle calculé. Ainsi, Nicholas Ayache [Aya89] définit le problème de l'appariement de primitives bruitées pour une relation géométrique donnée. Il s'agit de la capacité à distinguer les appariements plausibles des appariements aberrants. La distance géométrique classique est remplacée par une distance de Mahalanobis généralisée et la comparaison à un seuil est réalisée avec un test du χ^2. Cette distance est calculée en utilisant la covariance d'estimation des coordonnées du point en utilisant l'incertitude de mesure et les propagations linéaires de celle-ci.

Par ailleurs, il est nécessaire de limiter le nombre de paramètres des modèles recherchés ou les seuils de détections. Fixer un seuil est une étape délicate. On souhaite disposer d'une approche :

- ne nécessitant pas de définir de loi *a priori* dépendant du contexte,
- sans seuil ou paramètre,
- robuste à la présence de données aberrantes.

1.2 Le principe de détection *a contrario*

L'approche *a contrario* rentre dans le cadre des tests d'hypothèses, et elle résout les difficultés de définition de loi à priori : elle propose un modèle de fond minimaliste, qui explique le phénomène observé comme dû au hasard. Ainsi, dans [Moi03], L. Moisan écrit : "Il est beaucoup plus simple de définir un modèle à réfuter (typiquement un modèle uniforme) qu'un modèle précis des objets que l'on souhaite détecter."

Valider un test d'hypothèse nécessite de disposer pour une observation d'une mesure d'adéquation à un modèle. Pour l'approche *a contrario*, l'adéquation est forte avec un modèle de fond lorsque :

- l'observation possède une bonne mesure parce que le modèle explique sa présence (vrai positif),
- l'observation suit le modèle par hasard : il n'y alors pas de lien de causalité entre l'observation de l'évènement et l'hypothèse supposée (faux positif). Il s'agit d'une fausse alarme.

Définition 1. ***Evènement ε-significatif*** *(Définition 1 dans [DMM00]) – Un évènement est ε-significatif si l'espérance du nombre d'occurrences de cet évènement est inférieure à ε.*

Lorsque ε est inférieur à 1, on parle d'évènements significatifs. Calculer le *nombre de fausses alarmes* permet de savoir si l'observation est due au hasard, ou s'il s'agit d'un *évènement significatif*, comme nommé par Desolneux, Moisan et Morel [DMM00].

1.2.1 Une application : la détection d'alignements

Nous développons maintenant un exemple pour décrire le fonctionnement de la modélisation *a contrario*.

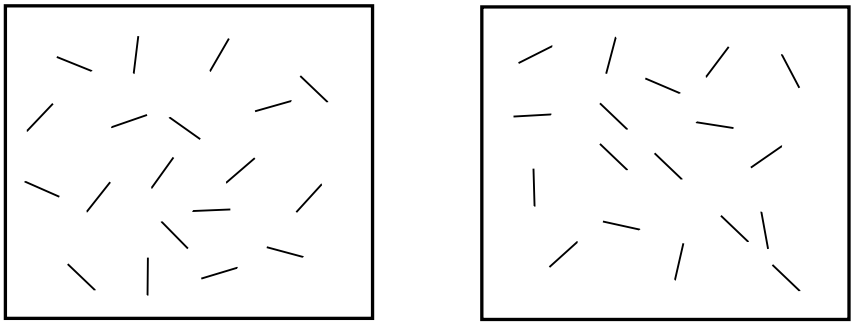

FIGURE 1.1 – A gauche, les alignements sont répartis aléatoirement en suivant une loi uniforme, ce qui correspond à l'hypothèse du modèle de fond. A droite, ce n'est pas le cas. L'alignement visible n'a que très peu de chance d'arriver par hasard. Le principe imagé ici, est formalisé par l'approche *a contrario*.

La théorie de la perception essaie d'expliquer les processus de cognition du cerveau humain. Elle décrit en particulier une méthode qui explique notre capacité à discriminer un objet quand celui-ci est suffisamment différent du reste de l'environnement le contenant. Desolneux *et al.* [DMM00] définissent ainsi un Gestalt comme la perception d'une structure obtenue par l'interprétation de l'image constituée d'un ensemble de petites zones d'intensités lumineuses différentes, sans signification les unes par rapport aux autres lorsque observées de façon isolée. Seul l'aspect global, via le groupage, permet d'obtenir une signification. En particulier, l'être humain est doué lorsqu'il s'agit de détecter des bordures ou des alignements : les neurones du cortex visuel de type simple ont une réponse qui est sélective et dépendante de l'orientation [FP02]. On peut s'en convaincre par notre capacité à reconnaître un objet ou un visage via une esquisse réalisée en ligne claire [FF87], et ce bien qu'une majorité de l'information supposée utile à l'identification est manquante, tels les textures, les ombres, le contexte. Ce principe est la principale source d'inspiration de la modélisation *a contrario*, où la détection se fait via une déviation du modèle de fond statistiquement significative. Cette dernière propriété est appelée principe de Helmholtz.

Il est possible de détecter des alignements en suivant ce principe, tel qu'illustré par la figure 1.1. Les déviations importantes d'une variable aléatoire uniforme, ici l'orientation des alignements, sont facilement perceptibles. Afin d'illustrer le principe de détection *a contrario*, nous présentons son utilisation pour la détection d'alignements en reprenant les grandes lignes de l'article de Desolneux *et al.*.

1.2.1.1 Principe

Un modèle *a contrario* est une approche statistique utilisant :

- un modèle de fond associé à l'hypothèse nulle $\mathscr{H}_0$,
- une mesure réalisée sur les structures que l'on cherche à détecter.

Le modèle de fond décrit les données observées lorsque celles-ci sont considérées indépendantes les unes des autres. Il s'agit le plus souvent d'une hypothèse a minima sur la forme des données lorsqu'aucune structure significative n'est visible. Par exemple, si on s'intéresse à la position de points dans une image, une répartition au hasard suivant une loi uniforme est un modèle de fond possible.

1.2.1.2 Modèle de fond

L'existence d'un modèle de fond est supposée : dans les images, les directions des alignements sont réparties au hasard et suivent une loi uniforme.

Soit u une image discrète de taille $N \times N$. Pour chaque point, le gradient d'intensité lumineuse est calculé sur un voisinage du point (une zone de 2 pixels de côté), ce qui permet d'obtenir la direction de la ligne de niveau correspondante par rotation de $\pi/2$. Ce calcul du gradient repose sur un filtre (par exemple le filtre de Sobel), auquel aucune étape de floutage préalable n'est appliquée afin de préserver l'hypothèse d'indépendance sur les directions.

Les directions des lignes de niveaux en deux points X et Y sont les mêmes à la précision $2\pi/n$ lorsque

$$|\theta_X - \theta_Y| \equiv \Delta_\theta(2\pi) \text{ et } \Delta_\theta \leqslant \frac{2\pi}{n}, \tag{1.1}$$

avec θ_X et θ_Y les directions extraites pour les points X et Y. On définit $p = \frac{1}{n}$, qui s'interprète comme la probabilité que deux points indépendants ont une direction proche à la précision $2\pi/n$ près, sous hypothèse d'uniformité.

L'hypothèse $\mathscr{H}_0$ est la suivante : la direction en tous points de l'image suit une variable aléatoire uniforme sur $[0, 2\pi[$. On suppose que toute déviation significative de cette hypothèse de répartition aléatoire conduit à la détection d'une structure.

Soit A un segment de l'image constitué de l points indépendants. Deux points sont indépendants lorsque leur distance en suivant le segment considéré est suffisamment grande (plus grande que 2). Nous comptons le nombre de points de A dont la direction est alignée avec la direction du segment A à la précision p près. Nous souhaitons ainsi connaître le nombre minimal k de points correctement alignés sur une longueur l d'un segment afin que cette observation soit un évènement significatif, c'est à dire une déviation du modèle de fond.

Soient $x_1, x_2, ..., x_l$ les l points indépendants de A, et X_i la variable aléatoire valant 1 lorsque la direction pour le point x_i est alignée avec celle de A à $2\pi/n$ près. Avec $p = \Pr(X_i = 1)$, et en utilisant l'indépendance des X_i, la loi de probabilité du nombre de points en lesquels la perpendiculaire au gradient est aligné avec le segment à la précision $2\pi/n$ près, est une loi binomiale

$$\Pr\left(\sum_{i=1}^{l} X_i = k\right) = \binom{l}{k} p^k (1-p)^{l-k}. \tag{1.2}$$

1.2.1.3 Significativité

Pour accepter ou réfuter le modèle de fond que l'on vient de définir, il est nécessaire pour cela de définir une mesure de similarité que l'on va pouvoir comparer à la variable aléatoire correspondant à la distribution des mesures pour le modèle combinatoire.

Un segment orienté est défini par la paire de points de l'image constituant chacune de ses extrémités. Le nombre total de segments possibles, pour une image de taille N par N, est, avec N^2 choix pour le premier point, et, l'ayant choisi, $N^2 - 1$ pour l'autre extrémité,

$$N^2(N^2 - 1) \simeq N^4. \tag{1.3}$$

Définition 2. ***Segment ε-significatif*** *(Définition 2 dans [DMM00]) – Un segment A de longueur l est ε-significatif dans une image de taille $N \times N$ s'il contient au moins $k(l)$ points ayants leurs directions alignées à la précision $\frac{2\pi}{n}$ avec celle du segment tel que :*

$$k(l) = \min\left\{k \in \mathbb{N}, N^4 \Pr\left(\sum_{i=1}^{l} X_i \geqslant k\right) \leqslant \varepsilon\right\}. \tag{1.4}$$

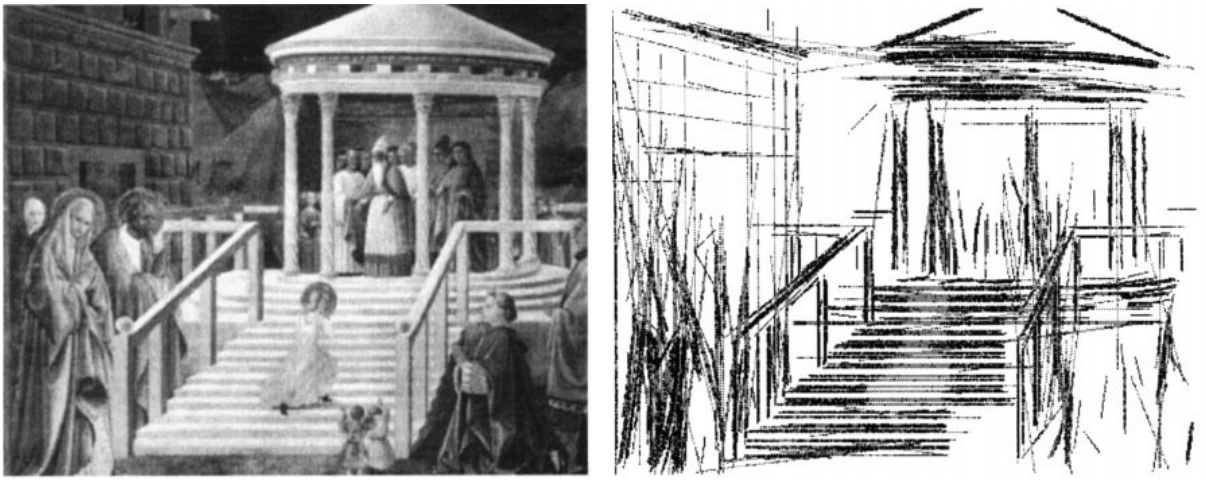

FIGURE 1.2 – A gauche, *La présentation de la vierge au temple* d'Uccello. A droite, segments les plus significatifs détectés. Extrait de [DMM00]

Définition 3. ***Nombre de fausses alarmes*** *(Définition 3 dans [DMM00]) – Soit A un segment de longueur l avec au moins k de ses points orientés dans la direction de A. Le nombre de fausses alarmes est défini par :*

$$\mathrm{NFA}(k,l) = N^4 \sum_{i=k}^{l} \binom{l}{i} p^i (1-p)^{l-i}. \tag{1.5}$$

[DMM00] ont démontré que le nombre de fausses alarmes est un majorant de l'espérance de l'évènement au moins k points de A sont alignés avec A sous hypothèse nulle $\mathscr{H}_0$. Il permet d'adapter automatiquement la longueur des segments détectés en fonction de la taille N de l'image.

La fiabilité de détection est mesurée par une majoration de l'espérance du nombre de fausses alarmes, plutôt que par la probabilité d'une fausse alarme. Cela est intéressant pour de nombreux problèmes car cela permet de faire un compromis entre un ensemble de petite taille associé à une mesure précise, et un ensemble plus grand satisfaisant lui aussi l'objectif recherché mais moins précis. Un tel comportement nécessite habituellement de régler des seuils pour chaque cas. L'avantage d'utiliser une espérance par rapport à une probabilité est qu'elle permet de comparer des groupes d'observations de cardinalités différentes. Ici l'adaptation se fait automatiquement, le nombre de fausse alarme devant être naturellement inférieur à 1 pour pouvoir déclarer l'évènement significatif.

1.2.1.4 Maximalité

Lorsqu'un alignement assez long est détecté car significatif, il est possible qu'un ou plusieurs de ses sous-segments soient aussi significatifs. De manière à ne pas multiplier les détections de structures intéressantes mais partielles, Desolneux et al. ont défini la notion de segment maximal, que l'on ne détaille pas ici.

1.2.2 Conclusion

Pour définir un modèle *a contrario*, il est nécessaire de se donner un modèle de fond, que l'on cherchera à réfuter, afin de choisir une meilleure explication. Pour cela, il est nécessaire d'établir une mesure de similarité entre les données observées : dans le cas de la détection d'alignement, il s'agit de la différence d'orientation. On définit aussi la significativité, qui permet d'estimer si une

mesure de similarité forte est due au hasard, et enfin une majoration de l'espérance de l'évènement associé, encore appelée nombre de fausses alarmes.

Maintenant que les principes de la méthodologie *a contrario* ont été décrits, nous allons présenter son application à la mise en correspondance de points d'intérêt sous contrainte géométrique.

1.3 Mise en correspondance *a contrario*

Dans cette section, nous rappelons rapidement les principes géométriques de la mise en correspondance de points entre images, puis nous présentons le modèle *a contrario* de Moisan et Stival [MS04b]. La géométrie qui relie deux images, appelée *géométrie épipolaire*, se représente par une quantité matricielle nommée *matrice fondamentale*. Elle est décrite au paragraphe suivant et son estimation est présentée en annexe A.

1.3.1 La géométrie épipolaire

La caméra suit le modèle de sténopé. Les coordonnées d'un point 3D $\mathbf{M} = [x, y, z]^T$ dans un repère du monde et les coordonnées de sa projection $\mathbf{m} = [u, v]^T$ dans le plan image suivent la relation suivante :

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{P} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \tag{1.6}$$

avec s un facteur d'échelle quelconque non nul, et $\mathbf{P}$ une matrice 3×4, appelée matrice de projection perspective. La matrice $\mathbf{P}$ se décompose ainsi :

$$\mathbf{P} = \mathbf{K}[\mathbf{R}\mathbf{t}] \tag{1.7}$$

avec $\mathbf{K}$ une matrice triangulaire 3×3 encodant les paramètres intrinsèques de la caméra (taille des pixels, focale de la caméra, angle entre les axes du plan image), $(\mathbf{R}, \mathbf{t})$ étant la translation et rotation appliquées pour réaliser le changement de repère entre repère du monde et repère de la caméra.

La géométrie épipolaire est définie pour toute paire de caméras distinctes. Dans la figure 1.3, C et C' sont les centres optiques de chaque caméra. Les projections du point $\mathbf{M}$ dans chaque vue $\mathscr{I}$ et $\mathscr{I}'$, sont notées $\mathbf{m}$ et $\mathbf{m}'$. La *contrainte épipolaire* s'exprime de la façon suivante : pour tout point $\mathbf{m}$ de $\mathscr{I}$, son correspondant dans $\mathscr{I}'$ se situe sur une droite appelée *droite épipolaire* $\mathbf{l}'_{\mathbf{m}}$, définie comme l'intersection du plan Π contenant $\mathbf{m}$, C et C', avec le second plan image $\mathscr{I}'$. On note en outre que toutes les droites épipolaires concourent en un point $\mathbf{e}$ (respectivement $\mathbf{e}'$) appelé *épipôle*, qui est image de C' dans $\mathscr{I}$ (et respectivement de C dans $\mathscr{I}'$).

La *contrainte épipolaire* s'exprime alors de la façon suivante :

$$\mathbf{m}'^{\mathbf{T}}\mathbf{F}\mathbf{m} = \mathbf{0} \text{ avec } \mathbf{F} = \mathbf{K}'^{-\mathbf{T}}[\mathbf{t}]_{\times}\mathbf{R}\mathbf{K}^{-\mathbf{1}}, \tag{1.8}$$

avec $[\mathbf{t}]_{\times}$ la matrice antisymétrique telle que

$$[\mathbf{t}]_{\times} = \begin{pmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{pmatrix} \text{ si } \mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix}$$

$\mathbf{F}$ s'appelle matrice fondamentale, est de taille 3×3, et est définie à un facteur multiplicatif près.

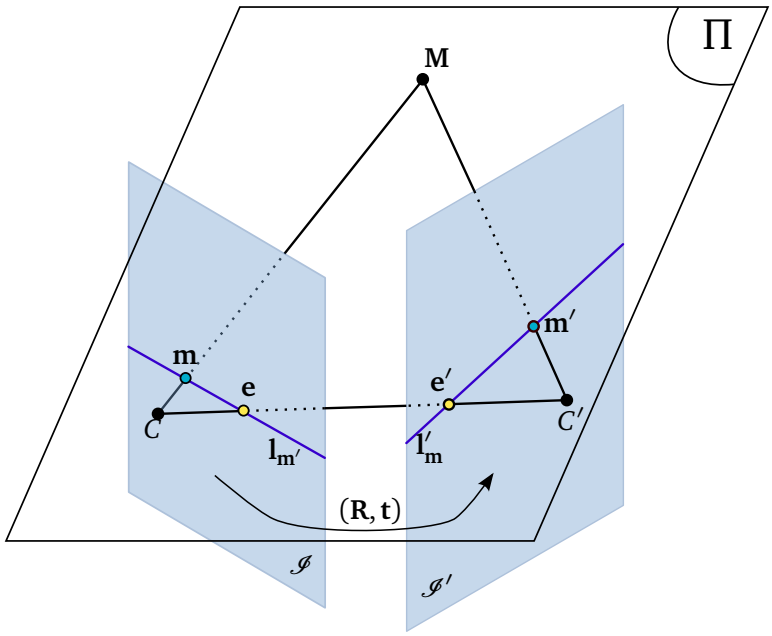

FIGURE 1.3 – La géométrie épipolaire. Notations de la section 1.3.1.

1.3.2 Les difficultés d'estimation de la matrice fondamentale

La matrice fondamentale **F** peut être déterminée à l'aide de la méthode de résolution aux moindres carrés. Habituellement, on utilise la Direct Linear Transform pour résoudre l'équation (1.8) à l'aide d'au moins 8 correspondances de points. Il est aussi possible de la résoudre à l'aide de 7 correspondances car elle n'a que 7 degrés de liberté, étant définie à un facteur d'échelle près et de rang 2. Ces deux méthodes d'estimation de la matrice fondamentale sont décrites en annexe A. Ces approches ne sont pas robustes à la présence d'appariements aberrants, ce qui se produit souvent en pratique. Les points mis en correspondance que l'on extrait via les algorithmes bas-niveau ne fournissent pas que des appariements corrects ; nombre d'entre eux sont aberrants, c'est-à-dire ne correspondent pas au même point $3D$ dans la scène considérée.

À partir d'un ensemble d'observations, l'algorithme Ransac, présenté en section 2.4.2, permet de séparer les observations réalisations d'un modèle paramétrique (dont on cherche à déterminer les paramètres) des observations indépendantes de ce modèle, dues à une mise en correspondance erronée. Cet algorithme comporte deux seuils, l'un sur la distance au modèle dist($\mathbf{m}', \mathbf{Fm}$) (en dessous duquel on décide que les observations considérées sont réalisations du modèle), l'autre sur la taille de l'ensemble dit de consensus (au dessus duquel on considère avoir trouvé un ensemble suffisamment grand pour estimer les paramètres du modèle). Ces deux seuils sont délicats à régler en pratique. De plus on ne dispose pas de critère pour juger *a priori* de la qualité de l'ensemble de consensus : cette qualité dépend à la fois de la taille de l'ensemble de consensus et de l'adéquation des observations retenues par rapport au modèle.

Dans le cadre de l'estimation de la matrice fondamentale entre deux vues, L. Moisan et B. Stival [MS04b] proposent une approche *a contrario* qui ne nécessite pas de fixer de seuil arbitraire et fournit une mesure de la pertinence de la solution trouvée. L'approche est basée sur un modèle *a contrario* : un ensemble d'observations est d'autant plus significatif qu'il a peu de chances d'être produit « par hasard », le hasard correspondant ici à une hypothèse de distribution uniforme des points dans l'image. Stewart [Ste95], a proposé une formalisation analogue qui repose sur une probabilité,

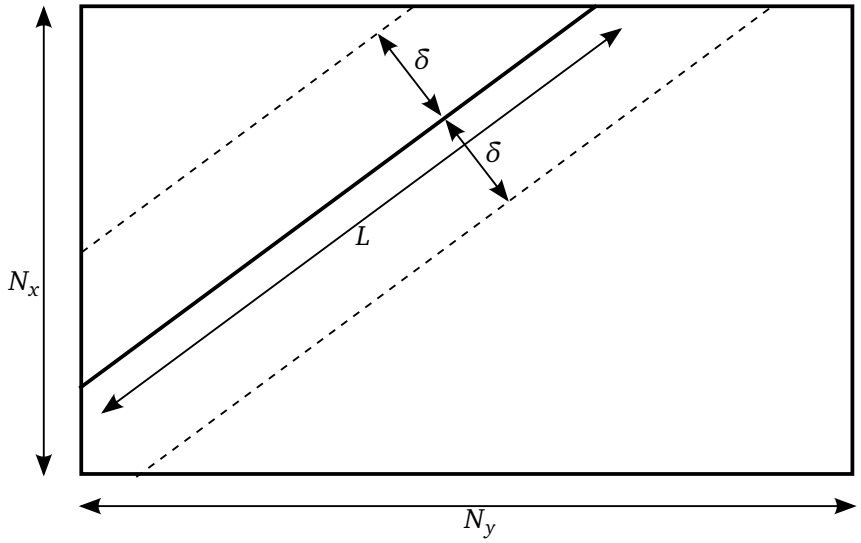

FIGURE 1.4 – Modèle de Moisan et Stival. La démonstration du théorème 1.

et non sur l'espérance. Comme indiqué précédemment, cette différence est majeure : l'espérance permet de comparer des ensembles de tailles différentes. Combiner les mesures de plusieurs paramètres, tels qu'ici la taille de l'ensemble considéré et la précision mesurée, en un seul est essentiel pour appliquer un critère de décision universel, et non dépendant du modèle ou de la qualité des données disponibles.

1.3.3 Modèle

1.3.3.1 α-rigidité

On cherche à évaluer la matrice fondamentale $\mathbf{F}$ entre deux images $\mathscr{I}_1$ et $\mathscr{I}_2$ à partir d'une collection S d'appariements (des couples $(\mathbf{m}, \mathbf{m}')$ de points d'intérêt en correspondance, le premier dans $\mathscr{I}_1$, le second dans $\mathscr{I}_2$). Comme un appariement $(\mathbf{m}, \mathbf{m}')$ est correct au sens de la géométrie épipolaire si $\mathbf{m}'$ est sur la droite $\mathbf{Fm}$, on dispose d'un critère naturel mesurant la rigidité des données S par rapport au modèle $\mathbf{F}$:

Définition 4. *(Définition 1 dans [MS04b]) Soit* $\mathbf{F}$ *une matrice* 3×3 *de rang 2. La* $\mathbf{F}$*-rigidité de l'ensemble* S *d'appariements est*

$$\alpha_F(S) = \frac{2D}{A} \max_{(\mathbf{m},\mathbf{m}')\in S} \operatorname{dist}(\mathbf{m}', \mathbf{Fm}) \tag{1.9}$$

où A *est l'aire du domaine de la seconde image,* D *son diamètre, et* dist *la distance euclidienne d'un point à une droite (voir figure 1.4).*

$\alpha_{\mathbf{F}}(S)$ est la probabilité pour que les points $\mathbf{m}'$ soient positionnés à une distance inférieure à $\max_{(\mathbf{m},\mathbf{m}')\in S} \operatorname{dist}(\mathbf{m}', \mathbf{Fm})$ des droites $\mathbf{Fm}$, lorsqu'ils sont générés par un processus uniforme dans l'image $\mathscr{I}_2$.

On définit alors :

Définition 5. *(Définition 2 dans [MS04b]) Un ensemble d'appariements* S *(de taille* $n \geqslant 7$*) est* α*-rigide s'il existe une matrice fondamentale* $\mathbf{F}$ *estimée à partir d'un sous-ensemble de 7 appariements tel que* $\alpha_{\mathbf{F}}(S) \leqslant \alpha$.

Donc plus α est petit, moins un ensemble α-rigide peut être le « produit du hasard ». L'explication alternative (cet ensemble respecte la contrainte épipolaire) sera alors retenue.

1.3.3.2 Ensembles d'appariements significatifs

Définition 6. *(Définition 3 dans [MS04b]) Soit $\mathbb{E}$ un ensemble fini de variables aléatoires prenant leurs valeurs dans un ensemble ordonné. L'événement "$E \leqslant t$" est ε-significatif si le nombre de fausses alarmes associé, défini par $|\mathbb{E}|P[E \leqslant t]$, est inférieur à ε.*

La ε-significativité permet de réduire l'ensemble de paramètres (le nombre d'appariement n, la taille k de l'ensemble de consensus S, le seuil de rigidité α) en un seul paramètre intuitif, le nombre attendu de fausses alarmes.

Théorème 1. *(Proposition 1 dans [MS04b]) Un ensemble S d'appariements de taille $n (n \geqslant 8)$ est ε-significatif dès qu'il est α-rigide avec*

$$\varepsilon_1(\alpha, n) = 3\binom{n}{7}\alpha^{n-7} \leqslant \varepsilon.$$

Démonstration. La quantité ε_1 représente la significativité d'un ensemble d'appariements. Il s'agit de tirer 7 appariements parmi n, chaque sous-ensemble génère au maximum 3 matrices $\mathbf{F}$. On nomme ce sous-ensemble T. La taille de l'ensemble $\mathbb{E}$ est dans ce cas de $3\binom{n}{7}$. Ensuite, parmi les appariements, on est en présence de deux cas :

- l'appariement $(\mathbf{m_i}, \mathbf{m_i'}) \in T$, alors la distance de $\mathbf{m_i'}$ à la ligne épipolaire associée $\mathbf{l} = \mathbf{F}\mathbf{m_i}$ est nulle par construction.
- l'appariement appartient à $S - T$, alors $\mathbf{m_i'}$ et $\mathbf{l}$ sont indépendants.

Comme visible dans la figure 1.4, l'intersection entre $\mathbf{l}$ et le plan image $\mathscr{I}'$ a pour longueur maximale la longueur de la diagonale de l'image D. L'aire de la bande de largeur δ autour de $\mathbf{l}$ est alors inférieure à $2\delta D$.

En supposant que les points détectés sont uniformément distribués dans l'image $\mathscr{I}'$ de taille $N_x \times N_y$, la probabilité de $\text{dist}(\mathbf{m_i'}, \mathbf{l}) \leqslant \delta$, qu'un point soit à une distance inférieure à δ d'un segment de droite de longueur L est $2\delta L/(N_x N_y)$. Si D est la longueur de la diagonale de $\mathscr{I}$ et A son aire, alors $A = N_x N_y$ et $L \leqslant D$. On en conclut que cette probabilité est majorée par $\alpha = 2D\delta/A$, aire de la zone mise en valeur sur la figure 1.4.

Pour chacun des $n - 7$ points, la probabilité est donc inférieure à α. Les points étant tirés indépendamment les uns des autres, la probabilité de $\alpha_F(S) \leqslant \alpha$ est donc inférieure à α^{n-7}. □

En présence de données aberrantes – Néanmoins, l'ensemble des appariements entre les deux vues contient des faux appariements (car les algorithmes de détection des appariements entre points d'intérêt en fournissent toujours). On cherche donc des sous-ensembles α-rigides de l'ensemble des appariements, qui sont formés d'appariements cohérents avec un mouvement entre les deux vues.

Considérons un ensemble S de n appariements. Il y a $\binom{n}{k}$ sous-ensembles de S à k éléments, $(n-7)$ choix pour k (qui varie entre 8 et n), $\binom{k}{7}$ choix pour les 7 couples de points permettant d'estimer la matrice fondamentale, et chacun de ces 7-uplets fournit (au plus) 3 matrices fondamentales. Il y a donc $3(n-7)\binom{n}{k}\binom{k}{7}$ matrices fondamentales à tester lorsque l'on examine tous les sous-ensembles de S.

D'autre part, la probabilité qu'un ensemble de k appariements soit α-rigide « par hasard » (i.e. sous hypothèse d'indépendance entre les points d'intérêts qui n'interviennent pas dans le calcul de la matrice fondamentale) est inférieure à α^{k-7}.

On en déduit que l'espérance du nombre des sous-ensembles α-rigides de taille k dans S (sous hypothèse d'indépendance des points d'intérêt et d'uniformité de leur distribution) est inférieure à

$$\varepsilon_2(\alpha, n, k) := 3(n-7)\binom{n}{k}\binom{k}{7}\alpha^{k-7}. \tag{1.10}$$

Ceci motive la définition suivante.

Définition 7. *(Proposition 2 dans [MS04b]) Un ensemble $S' \subset S$ de k appariements parmi n est ε-significatif s'il est α-rigide et $\varepsilon_2(\alpha, n, k) \leqslant \varepsilon$.*

La quantité $\varepsilon_2(\alpha, n, k)$ majore le nombre de sous-ensembles comme S' attendus sous l'hypothèse d'indépendance des points d'intérêt. Si ε_2 est faible (en particulier inférieur à 1) alors l'hypothèse d'indépendance est à rejeter. Les appariements de S' respectent alors la contrainte épipolaire.

ε_2 réalise un compromis entre la précision α des appariements selon la contrainte épipolaire et le nombre k de points dans l'ensemble considéré.

1.3.3.3 Algorithme Ransac a contrario

Comme rechercher le sous-ensemble le plus significatif parmi les $3(n-7)\binom{n}{k}\binom{k}{7}$ cas est impossible, une heuristique de type RANSAC est proposée dans [MS04b] et décrite ici dans l'algorithme 1. Dans cette heuristique, on minimise la quantité ε_2. Nous reviendrons sur l'approche RANSAC dans l'état de l'art du chapitre suivant, en section 2.4.2. Il est appelé Rsa dans [MS04b] pour Random Sampling Algorithm. Dans la suite de la thèse, les résultats présentés pour cette méthode seront appelés AC-RANSAC.

Algorithme 1 : Rsa (AC-RANSAC) – Déterminer l'ensemble d'appariements le plus significatif

Entrées : Ensemble S de n appariements avec $n \geqslant 7$
Sorties : Ensemble $\overline{S}$ le plus significatif de k appariements et sa significativité $\overline{\varepsilon}$
début
  $\overline{\varepsilon} \leftarrow +\infty$
  $N \leftarrow 10000$
  répéter
    échantillonner aléatoirement un ensemble T de 7 appariements parmi S
    estimer les matrices fondamentales associées (au plus 3)
    pour chaque *matrice fondamentale* **F** *associée à T* **faire**
      le calcul de l'ensemble le plus significatif U associé à **F**
      le classement des appariements par α croissant
      (on construit ainsi des groupes imbriqués de taille k et de rigidité α croissante)
      la recherche du groupe avec un $\varepsilon_2(\alpha, n, k)$ minimal
      si $\varepsilon_2(U) < \overline{\varepsilon}$ **alors**
        $\overline{\varepsilon} \leftarrow \varepsilon_2(U)$
        $\overline{S} \leftarrow U$
  jusqu'à *ce que le nombre d'essais soit supérieur à N*
fin

1.3.4 Prise en compte d'occurrences multiples

MAC-RANSAC [RDGM10] est une méthode *a contrario* permettant la détection de transformations multiples entre images. Cette approche étend RANSAC afin de détecter de façon séquentielle des groupes de points d'intérêt compatibles avec des mouvements différents.

L'objectif recherché est d'éviter la fusion de transformations proches (par exemple, trouver une seule homographie pour deux correspondances de plans aux orientations proches), et la sur-segmentation en plusieurs groupes d'ensembles de points pourtant cohérents (au contraire ici, trouver deux homographies pour des morceaux différents d'un même plan). Pour cela, un test de fusion de groupes est ajouté à l'approche de Moisan et Stival [MS04b] lorsqu'un groupe significatif est trouvé. Cela permet d'expliquer le mouvement d'objets déplacés dans une scène, ou encore de trouver automatiquement les différents plans visibles dans la scène, à partir du moment où ceux-ci contiennent suffisamment de points d'intérêt pour être détectables.

Pour pouvoir utiliser AC-RANSAC de façon séquentielle, Rabin et al. réalisent les modifications suivantes :

- Filtrage des correspondances obtenues par les détecteurs bas niveau pour retirer les candidats dits redondants ou autosimilaires : cela permet d'éviter les amas de points dans certaines zones.
- Après la détection d'un groupe d'appariements significatifs avec AC-RANSAC, détection d'une éventuelle fusion de plusieurs groupes correspondant à des transformations proches. Cette détection est réalisée via un critère de découpage appliqué aux groupes les plus significatifs trouvés. Si à partir d'un groupe S_0, il est possible de trouver deux sous-groupes disjoints S_1 et S_2 tels que $\text{NFA}(S_1)\,\text{NFA}(S_2) < \text{NFA}(S_0)$, S_0 est découpé. Ce critère est appliqué récursivement tant qu'il est possible de découper les groupes.

1.4 Conclusion

Nous avons présenté dans ce chapitre quelques fondements de la prise de décision en vision par ordinateur, en particulier des tests d'hypothèses. C'est dans ce cadre que s'inscrivent les méthodes *a contrario*, décrites ici d'abord via l'exemple de la détection d'alignements de Desolneux et al. [DMM00]. Les propriétés importantes de cette approche qui ont retenu notre attention sont la fixation automatique des seuils via l'estimation du nombre de fausses alarmes, ainsi que la robustesse en présence de taux élevés de données aberrantes. Une fois le cadre *a contrario* posé, nous avons rappelé quelques notions de géométrie épipolaire [Zha98, HZ00], et présenté le modèle *A Contrario* de mise en correspondance de points d'intérêt, AC-RANSAC, de Moisan et Stival [MS04b], qui sert de base aux travaux de cette thèse.

Chapitre 2

La mise en correspondance de points d'intérêt

Ce chapitre s'organise de la manière suivante. Après avoir présenté le schéma habituel d'estimation du mouvement via la mise en correspondance de points d'intérêt en section 2.1, nous décrivons les points d'intérêt ainsi que leurs descripteurs en section 2.2. Ces descripteurs permettent une mise en correspondance sur des critères purement photométriques*, comme on le verra en section 2.3. Celle-ci nécessite une étape supplémentaire de filtrage robuste des correspondances sur des critères géométriques, ce qui est l'objet de la section 2.4.*

Dans ce chapitre, nous présentons un état de l'art sur l'appariement, i.e. la mise en correspondance de points d'intérêt entre images. Les points d'intérêt recherchés sont ceux, parmi les ensembles de points extraits dans chaque vue, qui correspondent aux projetés pour chaque point de vue d'un même point de la scène en 3 dimensions.

L'appariement de tels points est utilisé par de nombreuses applications en vision par ordinateur. Citons par exemple :

- la reconnaissance d'objets [SZ03b],
- la création de panoramas [BL07] (voir l'exemple en figure 2.1),
- la stabilisation numérique en post-production [CFR99],
- l'analyse de la structure et du mouvement (*structure from motion* - SFM) et la reconstruction de scène 3D [PGV+04].

Les forts changements de point de vue, ainsi que la présence de motifs répétés dans une scène compliquent singulièrement la mise en correspondance comme on le verra en section 2.3.2.

2.1 Le schéma habituel d'estimation du mouvement

Afin de déterminer l'orientation relative de deux caméras à partir d'appariements de points entre ces vues (comme présenté par Longuet-Higgins [LH81]), une partie importante de la littérature en vision par ordinateur repose sur la mise en correspondance de points d'intérêt entre plusieurs vues. La *Mise en Correspondance* vise à identifier les points d'intérêt détectés parmi plusieurs images qui sont les projetés du même point en 3 dimensions. Cette action est souvent réalisée en prenant en compte des descripteurs locaux, c'est-à-dire un encodage des valeurs en niveaux de gris du voisinage d'un point d'intérêt.

Nous allons maintenant décrire le schéma de mise en correspondance entre deux images distinctes d'une même scène 3D statique prise par une caméra en mouvement. Une méthode habituelle

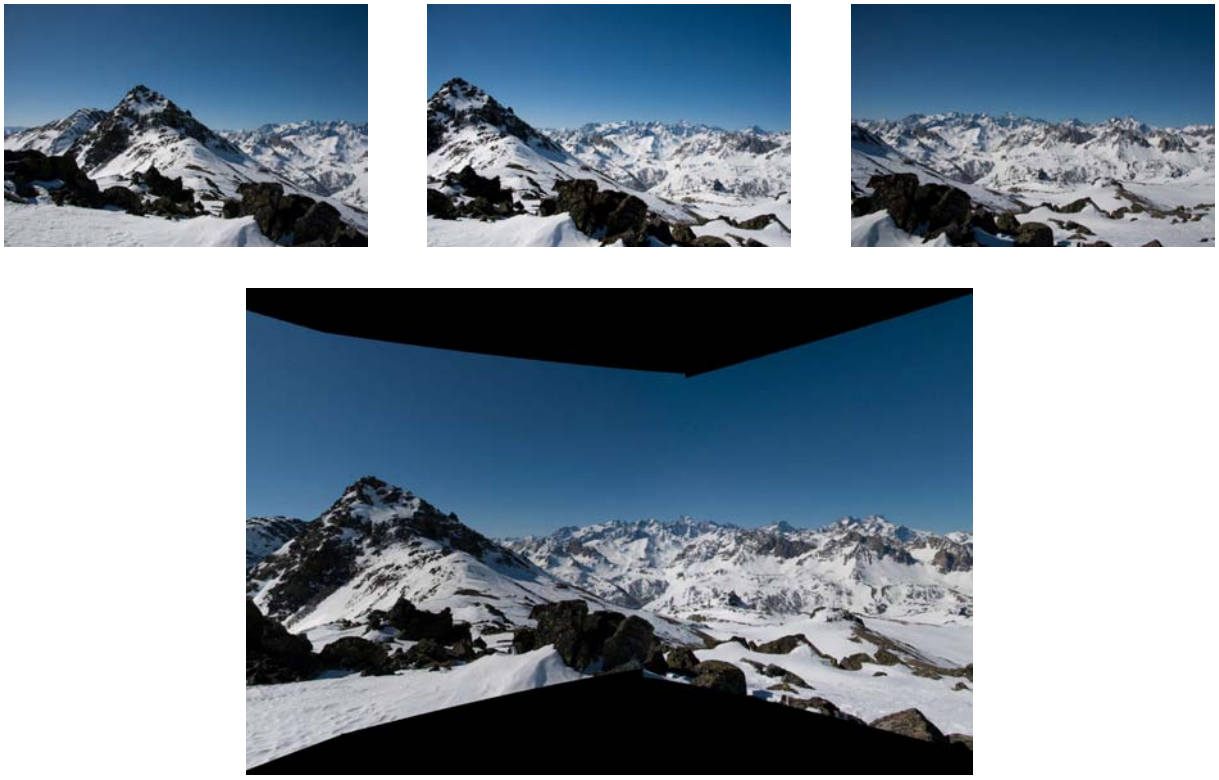

FIGURE 2.1 – Panorama réalisé à l'aide de mise en correspondance de points d'intérêt

est la suivante :

1. Dans chaque vue, extraire des points d'intérêt ainsi que leurs descripteurs associés représentant la photométrie au voisinage, dans une fenêtre de taille limitée centrée sur le point.
2. Les mettre en correspondance à l'aide d'une mesure de (dis-)similarité entre les descripteurs. Historiquement, une première façon de procéder est de comparer des fenêtres carrées centrées sur chaque point par corrélation.

   Définition 8. ***Corrélation*** *entre deux fenêtres d'images A et B de même taille de n pixels de côté. La mesure s'écrit :*

   $$\sum_{i,j=1}^{n} A_{i,j} B_{i,j} \tag{2.1}$$

   Cependant, une telle mesure ne prend pas en compte les déformations liées aux changements de points de vue. De plus, la corrélation ainsi définie n'est pas invariante par changement de contraste entre les images. Il existe des versions plus élaborées centrées et normalisées, avec des performances légèrement supérieures, mais elles ne suffisent à décrire le voisinage d'un point de façon invariante à tout type de transformation. La version centrée et normalisée s'écrit, avec A et B d'intensités moyennes $\bar{A}$ et $\bar{B}$:

   $$\frac{\sum_{i,j=1}^{n}(A_{i,j}-\bar{A})(B_{i,j}-\bar{B})}{\sqrt{\sum_{i,j=1}^{n}(A_{i,j}-\bar{A})^2 \sum_{i,j=1}^{n}(B_{i,j}-\bar{B})^2}} \tag{2.2}$$

   Par exemple, sans information a priori, il est impossible de savoir si les deux points sont observés à la même *échelle*, nécessaire pour définir la taille du voisinage que l'on considère.

Ces observations mettent en évidence la nécessité de disposer de représentations invariantes du voisinage d'un point. On reviendra sur ce point en section 2.2.

3. Filtrer les appariements en sélectionnant les plus cohérents avec une transformation géométrique imposée par un mouvement de caméra plausible.
4. Estimer le mouvement de la caméra entre les deux vues. Puis rendre l'ensemble des appariements retenus plus dense via une phase de relaxation de la mise en correspondance réalisée, en utilisant l'estimation que l'on vient d'établir. Cette étape est souvent appelée "mise en correspondance guidée".
5. Pour terminer, estimer le mouvement de la caméra à partir de l'ensemble final de correspondances obtenues.

Combiner toutes ces étapes est un travail difficile. Ceci nécessite en effet de fixer de nombreux paramètres, et un mauvais choix pour l'un d'eux risque de corrompre le résultat du procédé dans son ensemble.

Pour l'estimation de mouvement, on souhaite disposer de points aussi précis que possible, qui correspondent aux projetés de points 3D dans le plan image, et ce malgré les changements d'aspect du voisinage des points. Idéalement, connaître ce changement d'aspect nécessite d'avoir une connaissance complète de la scène : position de la caméra, forme de l'objet observé, matière le composant, positions des sources de lumière. A défaut de connaître l'information nécessaire à la création d'une représentation invariante dans le cas général, on se contente d'invariance partielle. Pour une zone planaire, la transformation reliant deux vues d'une telle zone est l'homographie. Une telle approche est utilisée dans [MDR04]. L'approximation à l'ordre 1 en est la transformation affine [MTS+06]. En simplifiant encore, la similitude (zoom et rotation, utilisée par l'approche SIFT de Lowe [Low04]) est une transformation pour laquelle la recherche d'invariant est plus aisée : elle nécessite de définir une échelle et une orientation.

Nous allons maintenant détailler les points cruciaux des étapes de mise en correspondance dans les prochaines sections, afin de mettre en évidence les points qui ont retenu notre attention dans cette thèse.

2.2 Points d'intérêt et descripteurs locaux

Dans cette section, nous présentons l'extraction de points d'intérêt, la première étape du schéma d'estimation de mouvement : d'abord, la partie détection, afin d'exhiber des points faciles à localiser d'une image à l'autre ; ensuite la partie descripteur, vecteur de caractéristiques invariantes utilisées pour la mise en correspondance.

Les points d'intérêt sont des endroits saillants, faciles à repérer. Depuis Moravec et ses premiers travaux sur la corrélation de zones d'intérêt [Mor77], la littérature traitant la détection de primitives bas-niveau s'est enrichie de nombreuses méthodes reposant sur différents outils : détection de contour via des filtres, calculs de gradient, de courbure. A partir de cette première vague de travaux, le détecteur de Harris-Stephens [HS88] lève certaines limitation de l'approche de Moravec en modifiant la façon dont l'auto-corrélation est mesurée, permettant la prise en compte de rotations quelconques, et non plus de pas $\pi/2$. I_x et I_y étant les dérivées partielles dans chaque direction de $I(x,y)$, la fonction image donnant l'intensité en niveau de gris pour le pixel courant, on écrit la matrice de Harris $A(x,y)$ telle que :

$$A(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}} \otimes \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \tag{2.3}$$

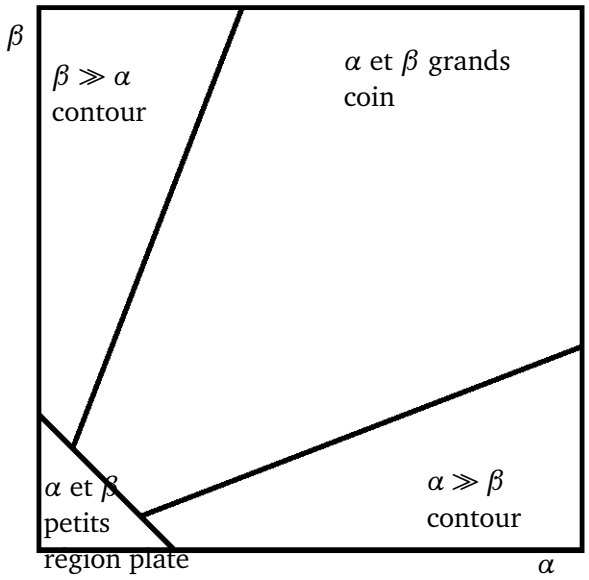

FIGURE 2.2 – Ordre de grandeur des valeurs propres α et β de la matrice de Harris et caractère de coin.

où $\otimes$ désigne le produit de convolution. La matrice de Harris permet d'estimer la courbure principale au voisinage du point considéré : pour qu'un point soit considéré comme point d'intérêt, il faut que les deux valeurs propres de cette matrice soient grandes, comme indiqué dans la figure 2.2. Plutôt que de calculer les valeurs propres, ce qui est coûteux, la fonction suivante est évaluée :

$$\alpha\beta - \kappa\,(\alpha+\beta)^2 = \det(A) - \kappa\ \mathrm{trace}^2(A) \tag{2.4}$$

avec κ un paramètre qui permet de définir le caractère plus ou moins marqué des coins recherchés, fixé habituellement à 0.04.

Tuytelaars et Mikolajcszyk [TM08] définissent les points d'intérêt comme un motif dans l'image qui diffère de son voisinage immédiat, pour des propriétés comme l'intensité, la couleur et la texture. Une telle définition couvre les coins, les jonctions en T, en L, les motifs de texture, et certaines parties de contours. Leurs caractéristiques attendues varient suivant l'utilisation que l'on fait des points d'intérêt : en reconnaissance d'objets, on souhaite par exemple reconnaître une classe d'objet (un vélo), même si son instance spécifique (un vtt) n'est pas présent dans la base d'apprentissage. Au contraire, lorsqu'on vise un positionnement précis, il est nécessaire de distinguer les objets très similaires, voire identiques, mais positionnés différemment. Pour leurs différentes utilisations, on leur associe un descripteur regroupant des mesures prises sur une région locale centrée sur le point détecté.

Malgré l'utilisation de stratégies d'observation de zones d'intérêt de tailles variables, la robustesse aux changements d'échelle reste faible [Mor80]. A la suite des travaux de Koenderink sur l'espace échelle gaussien [Koe84], Lindeberg [Lin98] propose de définir des points d'intérêt munis d'une échelle caractéristique de détection. Afin de détecter une échelle de façon fiable, il utilise celle pour laquelle on obtient les extrema du Laplacien.

Ces méthodes, détectant position et échelle de concert, si elles permettent une meilleure détection de points, n'aident que peu leur appariement, qui reste limité par les méthodes de corrélation. A partir des travaux fondateurs de Schmid et Mohr [SM97], qui associent aux points d'intérêt des descripteurs photométriques locaux, invariants aux rotations et changements d'échelles, une grande

variété de méthodes ont été proposées afin d'associer à chaque point d'intérêt un descripteur photométrique local.

2.2.1 Critères d'évaluation

Afin d'évaluer la performance de ces points, on définit la *répétabilité* d'un point détecté dans une scène comme la capacité à le repérer, confronté à différentes transformations 2D d'une image. Ces transformations sont par exemple : zoom, rotation, échantillonnage colorimétrique de l'image, flou...

La *répétabilité* est forte lorsque, pour deux conditions différentes d'observation d'une même scène, un grand pourcentage des points détectés dans la première image est aussi visible dans la seconde. Elle est calculée pour une précision δ fixant la distance maximale entre les coordonnées des points détectés dans la première image et projetés dans la seconde suivant la transformation expliquant leur déplacement, et celles mesurées dans la seconde, entre une image de référence et sa variante soumise à une transformation 2D quelconque.

Les articles d'évaluation de Mikolajczyk proposent une analyse de la répétabilité des points d'intérêt [MS04a, MTS+06].

Nous distinguons la *répétabilité* à des transformations 2D de la capacité à mettre en correspondance des points dans le cas d'un changement de point de vue dû à un mouvement de la caméra comprenant une translation du centre optique (pas de rotation pure autour de l'axe optique) : dans ce cas, seule la connaissance complète de la scène nous permettrait de le mesurer. Moreels et Perona [MP07] présentent une évaluation des détecteurs prenant en compte cette limitation, via l'utilisation d'une table robotisée tournante et d'un système de validation à trois vues. A partir de la géométrie calculée entre chacune des paires de vues, qui donne une contrainte point / droite, il est possible d'en déduire une contrainte point / point via l'intersection de deux droites épipolaires dans la troisième vue, ce qui permet un calcul automatisé de la répétabilité pour toute forme d'objet. Cela leur permet d'établir que les combinaisons de détecteurs et descripteurs Hessian-affine et SIFT sont les plus robustes vis à vis des changements de points de vue, et que pour les changements d'éclairage et de focale des caméras les Harris-affine avec SIFT, et les Hessian-affine avec des shape context donnent les meilleurs résultats. Ils montrent aussi qu'aucune des combinaisons testées ne donne des bons résultats pour des changements de points de vue de plus de 25–30°.

Obtenir une répétabilité forte à une certaine classe de transformations ne nous assure cependant pas du caractère *distinctif* de l'information apportée par le point. Ce caractère distinctif, qui signifie empiriquement que le descripteur du point contient suffisamment d'information pour mettre en correspondance sans ambiguïté, varie en fonction du contexte. Lorsqu'une scène contient de nombreux motifs répétés, l'ambiguïté reste entière si l'on utilise uniquement l'information photométrique.

2.2.2 Caractéristiques souhaitables d'un détecteur

Nous indiquons les qualités recherchées pour un point d'intérêt utilisé dans le cadre de l'estimation de mouvement. On souhaite disposer de points d'intérêt munis de descripteurs, ayant les qualités suivantes :

- une précision de localisation élevée ;
- une bonne répétabilité vis à vis des transformations de l'image ;
- un caractère local pour une robustesse aux occultations ;
- une description invariante :
  - à certains changements de points de vue (similitudes, transformations affines, ou homographies) ;

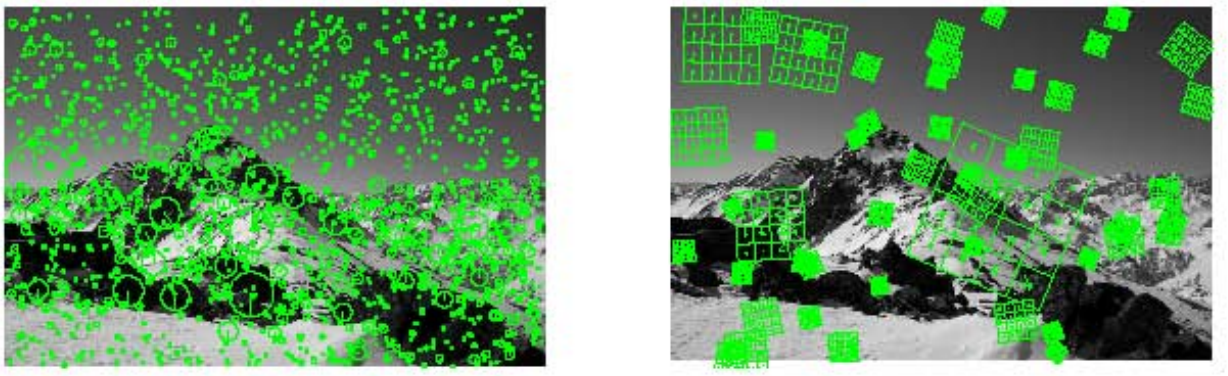

FIGURE 2.3 – A gauche – Points d'intérêt détectés. Les rayons des cercles sont proportionnels aux échelles calculées, et l'orientation trouvée est matérialisée par le rayon. A droite – Descripteurs pour une sélection de points. Les grilles affichées contiennent les statistiques d'orientation du gradient dans un voisinage du point dont la taille est proportionnelle à l'échelle détectée. (via la bibliothèque VLFeat [VF08])

– aux différences de contraste (à défaut de la prise en compte des variations d'éclairage).

Un des algorithmes les plus populaires, qui satisfait raisonnablement les caractéristiques ci-dessus, est probablement SIFT de David Lowe [Low04]. Nous utilisons dans nos expériences cette approche, et allons donc le présenter plus en détails. La figure 2.3 présente un exemple de détections.

2.2.3 Le détecteur Sift

Le détecteur SIFT, proposé par [Low04] extrait des points remarquables d'une image. L'extraction par ce procédé est stable dans le sens où une autre image raisonnablement éloignée d'une même scène permettra d'obtenir des points correspondants pour la plupart aux mêmes coordonnées 3D. De tels points sont invariants aux modifications d'échelle et aux rotations, et partiellement invariants aux changements de luminosité et de point de vue 3D.

Grandes étapes Les étapes principales de ce procédé sont les suivantes :

– **Détection d'extrema dans l'espace-échelle** gaussien, via le calcul des extrema des différences de gaussienne convoluées avec l'image. L'espace-échelle (voir la figure 2.4 est défini comme la fonction $L(x, y, \sigma)$, produit de convolution entre la gaussienne $G(x, y, \sigma)$ et l'image $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) \otimes I(x, y), \tag{2.5}$$

avec $\otimes$ le produit de convolution et

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}. \tag{2.6}$$

La fonction différence de gaussienne convoluée avec l'image s'écrit alors :

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{2.7}$$

Le point d'intérêt est alors défini comme extrema des différences de gaussiennes.

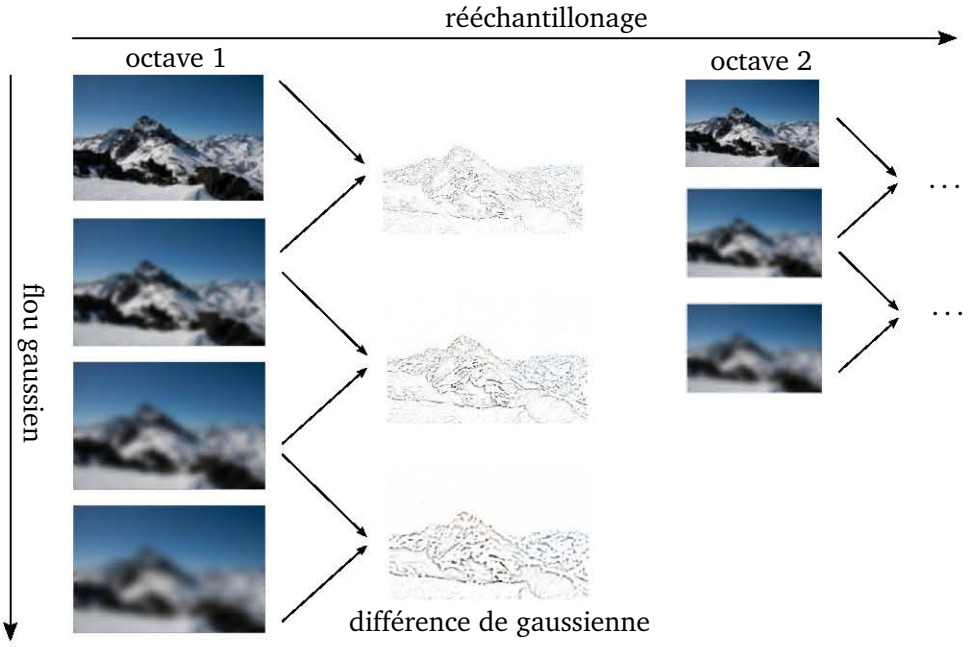

FIGURE 2.4 – Espace échelle gaussien. On recherche les extrema des différences de gaussienne afin de déterminer la position des points d'intérêt. Une octave correspond au doublement de l'échelle du flou gaussien σ.

- **Localisation de points clés :** pour chaque candidat, une optimisation sous-pixelique des coordonnées x, y et de l'échelle s est réalisée : si la position obtenue est stable (les coordonnées trouvées après l'optimisation décrivent un point suffisamment proche du point détecté initialement), il est déclaré point clé. Les points d'intérêt candidats correspondant à une zone de faible contraste, ou ceux dont la réponse est analogue à un contour sont éliminés.
- **Affectation d'une orientation :** chaque point clé se voit attribuer une orientation θ définie par la direction dominante du gradient au voisinage.

2.2.4 Les descripteurs de points clés

Une fois qu'un point d'intérêt a été obtenu et est muni d'une position, d'une échelle et d'une orientation, il est possible de calculer un descripteur.

Calcul du descripteur pour SIFT En utilisant les gradients pré-calculés pour tous les niveaux de lissage à l'étape d'extraction, on construit un descripteur orienté par la direction du gradient. Une fonction de pondération gaussienne est appliquée à l'amplitude suivant l'éloignement de l'échantillon considéré sur un disque centré au point clé. On ré-échantillonne ensuite par interpolation les orientations des gradients suivant 8 directions principales, en groupant des carrés de 4 pixels de côté. On obtient ainsi pour une grille de 4×4 histogrammes avec 8 classes – les orientations – un descripteur contenant 128 éléments. La figure 2.5 représente un descripteur sur une grille de taille 2×2, i.e. contenant 4 histogrammes.

Les voisinages des points d'intérêt, ayant des propriétés d'invariance aux changements d'échelle et d'orientation sont définis à l'aide de l'échelle donnée par les extrema des différences de gaussiennes, et de la direction principale du gradient à l'intérieur d'un cercle dont le rayon est proportionnel à cette échelle. Le descripteur du point est constitué d'histogrammes qui regroupent les

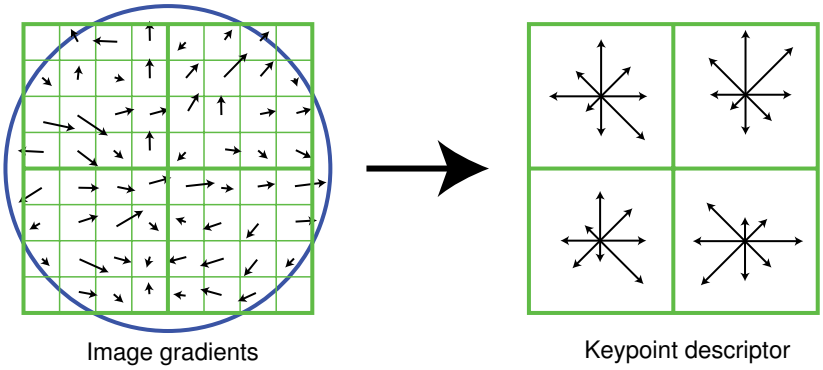

FIGURE 2.5 – A gauche, détails des valeurs et directions des gradients ; à droite, descripteur SIFT correspondant. Le cercle bleu matérialise la pondération gaussienne suivant l'éloignement du centre de la zone clé. (Extrait de [Low04]).

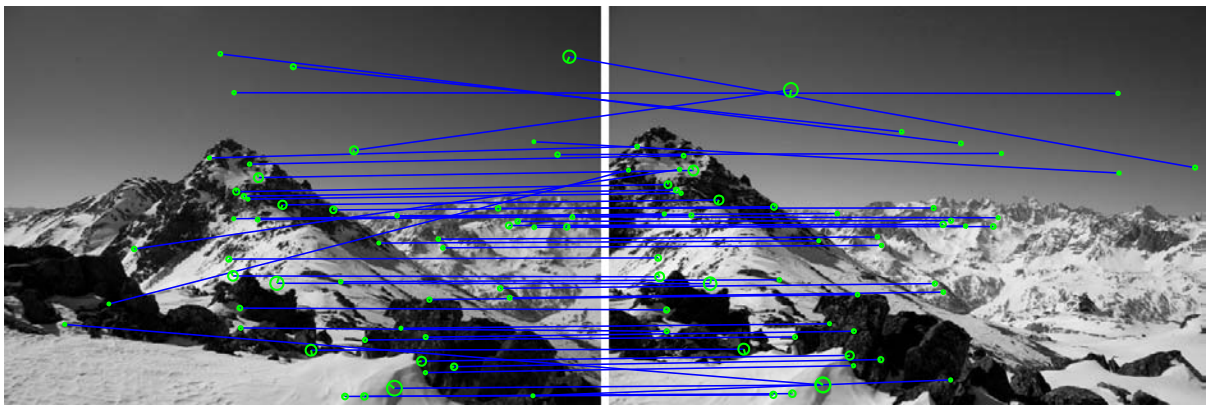

FIGURE 2.6 – Points d'intérêt issus de la figure 2.3, mis en correspondance par le critère du rapport de distances. Notez les quelques correspondances aberrantes, pour lesquelles les points d'intérêt appariés ne correspondent pas aux même points physiques.

statistiques sur les directions du gradient dans ce voisinage.

Remarquons que la direction du gradient est invariante aux changements de contraste, au contraire de la norme du gradient. Le vecteur construit est la concaténation d'histogrammes des directions du gradient. Les votes contribuant à chaque classe sont pondérés par la norme du gradient. Afin de réduire les effets des changements d'éclairage, une normalisation du vecteur ainsi formé est finalement réalisée.

2.3 Mise en correspondance sous contrainte photométrique

La mise en correspondance est certainement un des points les plus problématiques de la méthode. Pour réaliser l'appariement de points d'intérêt, il est nécessaire de se munir d'une mesure de similarité photométrique entre les descripteurs, qui établit leur ressemblance ou dissemblance. Définir cette métrique est difficile. Une approche naïve est alors de définir comme appariements ceux dont la distance suivant cette mesure de similarité est en dessous d'un certain seuil. Cependant, mettre un simple seuil sur la distance euclidienne entre les descripteurs afin de définir des appariements ne fonctionne pas. Cette approche, dite du *seuil global*, n'est pas concluante selon [Low04].

Une méthode habituelle [Low04], montrée comme plus robuste, pour définir un ensemble d'appariements est de garder le plus proche voisin, au sens de la distance entre descripteurs, pourvu que le rapport r de distances entre le plus proche et le second plus proche voisin soit inférieur à un seuil ρ fixé (inférieur à 1, typiquement entre 0.6 et 0.8). Cette méthode est utilisée dans le logiciel Bundler [Sna].

Définition 9. *Méthode du* **rapport de distances entre plus proches voisins**. *(*NNT *pour Nearest Neighbourgh Threshold). $\widetilde{\mathsf{dist}}$ est la distance entre descripteurs D. x, y_1 et y_2 sont des points extraits d'images, munis de descripteurs $D(x)$, $D(y_1)$ et $D(y_2)$. $D(y_1)$ est le plus proche voisin de $D(x)$ pour la distance $\widetilde{\mathsf{dist}}$, et $D(y_2)$ est son second plus proche voisin.*

$$x \text{ est en correspondance avec } y_1 \text{ si et seulement si } r < \rho \text{ avec } r = \frac{\widetilde{\mathsf{dist}}(D(x), D(y_1))}{\widetilde{\mathsf{dist}}(D(x), D(y_2))} \tag{2.8}$$

Le plus proche voisin est d'autant plus fiable que le rapport des distances est faible. Cette approche, qui utilise la distance euclidienne, donne des résultats convenables mais encore pollués par des correspondances aberrantes (la figure 2.6 montre le résultat de cette étape). Une telle approche a pour vertu de supprimer les correspondances pour des points munis de descripteurs peu discriminants, qui ont une distance faible avec de nombreux descripteurs. Supprimer ces points mal caractérisés permet d'éliminer des appariements qui risquent forts d'être faux. Néanmoins, si deux zones contiennent des motifs identiques et sont munies de points d'intérêt, ceux-ci seront aussi éliminés par ce critère, ce qui est alors dommageable pour la suite du processus d'estimation de mouvement.

2.3.1 Distance entre descripteurs à base d'histogrammes

Comme la plupart des descripteurs sont composés d'histogrammes d'orientations du gradient, quelques chercheurs proposent de remplacer la distance euclidienne par une distance plus adaptée à la comparaison d'histogrammes. On peut mentionner (par complexité croissante), la distance du χ^2, la distance de diffusion de Ling et Okada [LO06], la distance du terrassier (Earth Mover's Distance – EMD – voir les travaux fondateurs de Rubner et al. [RTG00], Rabin et al. [RDG08b, RDG08a, RDG09a], ou encore Ling and Okada [LO07]).

2.3.1.1 Distances euclidienne et de Manhattan

L'approche naïve lorsqu'il s'agit de mettre en correspondance des descripteurs est d'utiliser la distance euclidienne, ou celle de Manhattan. On a alors, lorsque $D(x)_j$ est la valeur de la j-ème classe parmi N de l'histogramme $D(x)$:

$$\widetilde{\mathsf{dist}}(D(x), D(y)) = \left(\sum_{j=1}^{N} |D(x)_j - D(y)_j|^p \right)^{1/p} \tag{2.9}$$

avec $p = 1$ (distance de Manhattan) or $p = 2$ (euclidienne).

2.3.1.2 Distance du χ^2

Certains articles [LO06, LO07] démontrent que comparer des histogrammes classe à classe à l'aide de la distance du χ^2 améliore la mise en correspondance par rapport à la distance euclidienne,

et ceci sans nécessiter beaucoup de calculs supplémentaires.

$$\widetilde{\mathrm{dist}}(D(x),D(y)) = \chi^2(D(x),D(y)) = \sum_{j=1}^{N} \frac{\left(D(x)_j - D(y)_j\right)^2}{D(x)_j + D(y)_j} \tag{2.10}$$

2.3.1.3 Distance CEMD de Rabin et al.

Une autre approche, étudiée dans [LO07, RDG09a, RTG00], est d'utiliser la distance du terrassier(*Earth Mover's Distance* - EMD), qui est particulièrement bien adaptée à la comparaison d'histogrammes.

En effet, la plupart des descripteurs photométriques (et SIFT en particulier) sont constitués d'histogrammes de la direction du gradient, distribuée sur l'intervalle circulaire $[0, 2\pi[$. Par conséquent, il est approprié d'utiliser une métrique qui prend en compte ces histogrammes circulaires, comme la distance proposée par Rabin et al. [RDG09a]. La suite de cette section est un résumé de [RDG09a] auquel nous renvoyons le lecteur pour plus de détails. Nous utilisons ici les mêmes notations que dans [RDG09a].

La signification intuitive de EMD est qu'elle correspond au coût minimum des transports accomplis par un terrassier pour fabriquer l'histogramme y à partir de l'histogramme x, étant connu le coût $c_{i,j}$ de déplacement d'une unité de matière d'une classe i de l'histogramme x à une classe j de l'histogramme y.

Soient $x = (x_1, \ldots, x_N)$ et $y = (y_1, \ldots, y_N)$ deux histogrammes (circulaires) constitués de N classes. La distance circulaire EMD (noté CEMD) entre eux est alors définie comme la solution du problème de programmation linéaire suivant :

$$\mathrm{CEMD}(x,y) = \min_{(\alpha_{i,j})\in\mathcal{M}} \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_{i,j} c(i,j), \tag{2.11}$$

avec

$$\mathcal{M} = \left\{ (\alpha_{i,j}) \text{t.q. } \alpha_{i,j} \geq 0, \sum_j \alpha_{i,j} = x_i, \sum_i \alpha_{i,j} = y_j \right\} \tag{2.12}$$

et $c_{i,j}$ est le coût du transport entre les classes x_i and y_j. Ceci correspond à la définition générale de la distance du terrassier. Pour des histogrammes circulaires, les auteurs de [RDG09a] ont choisi une fonction de coût circulaire naturelle L^1 :

$$c_{i,j} = \frac{1}{N} \min(|i-j|, N - |i-j|). \tag{2.13}$$

Rabin et al. prouvent dans [RDG09b] que le calcul de CEMD est réalisable et facile en pratique. Ils considèrent l'histogramme cumulé X^k de x commençant à la k-ème classe pour chaque $i \in \{1, \ldots, N\}$, $X_i^k = \sum_{j=k}^{k+i} x_j$, avec x qui est indexé de façon circulaire. L'histogramme cumulé Y^k est défini pour y de la même façon. En supposant x et y normalisés, (i.e. $\sum_i x_i = 1$), Rabin et al. montrent :

$$\mathrm{CEMD}(x,y) = \min_{k=1\ldots N} ||X^k - Y^k||_1. \tag{2.14}$$

Cependant, lorsque que l'on considère des descripteurs photométriques, les histogrammes d'orientation x et y **ne** sont **pas** normalisés (pour SIFT, le descripteur est normalisé *dans son ensemble*). Néanmoins, en accord avec [RDG09a], nous conservons cette méthode pour calculer CEMD.

2.3.1.4 De l'histogramme au descripteur

Nous avons maintenant défini les différentes distances que nous utilisons pour comparer les descripteurs à base d'histogramme. Dans le cas du descripteur SIFT, le descripteur est la concaténation de 16 histogrammes. Les distances finales que nous avons étudiées sont soit la somme des distances issues de chaque histogramme, soit le maximum de ces distances.

Un travail de comparaison de ces différentes distances entre descripteurs a été réalisé [NSB10a], que nous présenterons au chapitre 5.

2.3.2 L'appariement en présence d'aliasing perceptuel

La présence de motifs répétés dans une image est analogue au concept d'aliasing perceptuel – *Perceptual aliasing* – introduit par Whitehead et Ballard en 1991 [WB91]. Il désigne une situation telle que :

> "*a state in the world, depending upon the configuration of the sensorymotor subsystem, may map to several internal states ; [and] conversely, a single internal state may represent multiple world states.*"

Lors de l'appariement de points d'intérêt, les descripteurs invariants permettent de résoudre le premier aspect du problème de l'aliasing perceptuel tant que le changement de point de vue n'est pas trop important : un point 3D – *a state in the world* – a des descripteurs analogues pour plusieurs points de vues différents – *several internal states*. Un descripteur invariant comme celui de SIFT est en effet supposé produire une représentation unique du point 2D correspondant, pour les changements d'échelle et les rotations. Cependant, le second aspect du problème reste entier : les motifs répétés sont aussi encodés par des descripteurs analogues – *a single internal state*, bien qu'ils ne représentent pas les même points 3D – *multiple world states*. De même, des points munis de descripteurs faiblement discriminants, vont être mis en correspondance avec de nombreux points. Ainsi, presque tous les algorithmes de mise en correspondance de points d'intérêt échouent en présence de motifs répétés, ou sur des textures à faible contraste, à moins qu'ils ne soient pris en compte via une application dédiée (détection de motifs autosimilaires dans [SZ00], super-résolution via les motifs autosimilaires dans [GBI09]). Ce problème est de première importance : les motifs répétés sont très fréquents dans les environnements façonnés par l'homme. Prenons l'exemple de deux photos d'une façade d'immeuble prises selon différents points de vue : mettre en correspondance correctement les fenêtres est tout simplement impossible si l'on prend en compte uniquement les descripteurs invariants des points. Il est nécessaire d'utiliser de l'information supplémentaire prenant en compte la géométrie. C'est le sujet de notre première contribution que nous présenterons au chapitre 4.

Néanmoins, le problème de l'aliasing perceptuel n'est alors pas complètement résolu. Comme indiqué précédemment, pour des forts changements de points de vue, l'invariance n'est plus satisfaite et un point 3D ne sera plus représenté par des descripteurs suffisamment proches dans différentes vues. Récemment, l'appariement a été rendu possible pour des changements de points de vue très forts (jusqu'à 80 degrés d'angle d'incidence entre certaines zones mises en correspondance) par les approches de simulation de points de vue, permettant l'invariance affine, comme celle présentée par Morel et Yu dans ASIFT [MY09b]. Notre seconde contribution est d'adapter cette approche pour permettre de prendre en compte les ambiguïtés, et ainsi de profiter du gain en invariance pour réussir la mise en correspondance. Nous la présentons au chapitre 6.

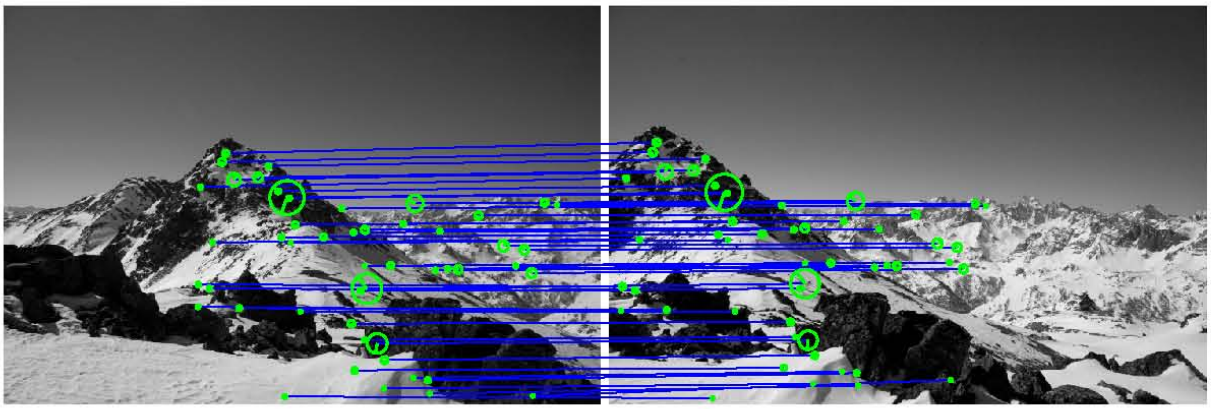

FIGURE 2.7 – Points d'intérêt issus de la figure 2.3 mis en correspondance en figure 2.6, et filtrés en suivant une contrainte homographique – pour un panorama, on tourne la caméra autour de son centre optique, ce mouvement global de rotation est expliqué par une homographie – par la méthode *A Contrario* RANSAC.

2.4 Le filtrage robuste à l'aide de la contrainte géométrique

Une fois qu'un ensemble d'appariements a été extrait des images, nous cherchons à sélectionner un sous-ensemble constitué d'appariements qui sont compatibles avec un modèle géométrique sous-jacent. Supposons que le mouvement de la caméra ne se limite pas à une rotation autour de son centre optique, et que les points 3D observés ne sont pas tous coplanaires. Lorsque l'on suit le modèle de sténopé, la *géométrie épipolaire* est caractérisée par la *matrice fondamentale* (ou la matrice essentielle si les paramètres intrinsèques de la caméra sont connus) [FLP01, HZ00]. Comme les appariements sont pollués par des couples de points d'intérêt mal mis en correspondance – des appariements qui se ressemblent, mais ne correspondent pas au même point 3D – appelés appariements aberrants par la suite (*outliers* en anglais), des méthodes de statistiques robustes, telles que les LMedS or M-estimateurs [TM97, Zha98] sont employées. Une méthode habituellement retenue pour réaliser le filtrage robuste est RANSAC [FB81], ainsi que de nombreuses variantes qui en découlent (MSAC, MLESAC [TZ00], MAPSAC [Tor02] et d'autres encore [CM05] sans vouloir être exhaustif).

Le résultat de cette étape, cette fois avec une contrainte homographique, est visible en figure 2.7.

2.4.1 Limites de l'estimation aux moindres carrés

L'évaluation de quantités à partir de mesures polluées par des données aberrantes, ou outliers, n'est pas aisée au moyen d'algorithmes tels que les moindres carrés : ce type de méthode est y en effet très sensible. Ceux-ci perturbent fortement la solution trouvée.

Un exemple frappant de l'absence de robustesse de l'algorithme des moindres carrés est illustré par la figure 2.8 : la minimisation étant effectuée suivant l'erreur quadratique, les points aberrants prennent le pas sur les points corrects du fait de leur poids important. Six des sept points constituent des données valides, cependant l'algorithme des moindres carrés prend en compte le point aberrant et renvoie une solution attirée par celui-ci.

Ce problème est fréquent en vision par ordinateur, où les données obtenues en analysant les images ne sont pas fiables au sens qu'elles contiennent souvent une proportion importante d'outliers. Afin de résoudre ces difficultés, les estimateurs robustes ont donc été introduits par les statisticiens.

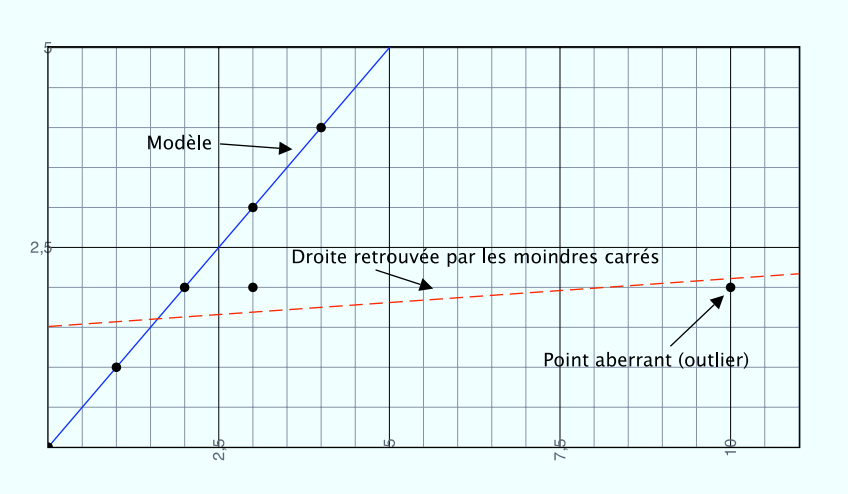

FIGURE 2.8 – Résolution aux moindres carrés avec des données aberrantes : recherche d'un modèle affine f minimisant $\sum_{i=1}^{n} |y_i - f(x_i)|^2$ où les (x_i, y_i) sont les données ; la solution trouvée est beaucoup plus attirée par le point éloigné que par les quatre points suivant le modèle. D'après [FB81]

Les approches populaires en vision sont les M-estimateurs [Hub81], ainsi que les moindres carrés médians (*Least Median of Squares* – LMS) [RL87]. Les M-estimateurs sont des fonctions de résidus, continues et symétriques avec un minimum en 0, utilisées afin d'écrêter les résidus dont l'influence croît linéairement pour une estimation aux moindres carrés classique. Ils sont utilisés pour minimiser de façon robuste la somme d'une fonction de résidus. Les moindres carrés médians, ou LMS, consistent eux à minimiser la médiane des résidus au carré, ce qui permet d'éliminer la deuxième moitié des résidus triés par ordre croissant.

Les performances de ces estimateurs sont mesurées via deux quantités [RL87] : *le point de rupture*, le plus petit pourcentage de données aberrantes pouvant faire échouer l'estimation, et *l'efficacité relative* qui traduit la précision de l'estimateur. En pratique, le point de rupture est de l'ordre de 20% pour les M-estimateurs et de 50% pour les LMS. La fraction de données aberrantes doit rester inférieure au point de rupture, et 50% n'est souvent pas suffisant en vision par ordinateur.

Une solution existe pour obtenir un point de rupture plus élevé : les approches dérivées de RANSAC [FB81], cherchent à extraire les données correctes, dites *inliers*, de celles aberrantes, *outliers*, à partir d'un ensemble de données fournies. Contrairement aux approches précédentes, une fois classifiées comme outliers, les données aberrantes n'interviennent plus dans le résidu de l'estimation. On obtient alors un point de rupture beaucoup plus élevé (entre 70 et 90% suivant les variantes).

La transformée de Hough [Hou59] est une autre solution. Elle a l'avantage d'être robuste, et de permettre la détection de multiples occurrences de structures suivant le même modèle. Cependant, la discrétisation de l'espace des paramètres qu'elle implique rend son application difficile à de nombreux problèmes : paramétrer les cellules de l'espace de vote pour qu'elles soient équiprobables est complexe. La procédure de vote employée devient de plus très coûteuse quand le nombre de paramètres augmente. Enfin, reconnaître les points d'accumulation afin d'établir les valeurs des paramètres du modèle est aussi un problème en soi.

2.4.2 RANSAC

RANSAC [FB81] (voir l'algorithme 2) est un paradigme reposant sur un principe simple : on recherche le plus grand nombre de données compatibles avec un modèle paramétrique, c'est à dire

dont l'erreur $d_{\mathbf{F}}$ est inférieure à un seuil δ fixé a priori. Par ailleurs, lorsqu'un ensemble de données est pollué, si la proportion de données aberrantes n'est pas trop importante, moins on utilise de données, moins on a de chance d'être confronté à des outliers. Plutôt que d'utiliser toutes les données disponibles pour calculer les paramètres du modèle, on utilise alors le nombre de données minimales pour que le système à résoudre soit suffisamment contraint, en fonction du nombre de degrés de liberté du modèle. Enfin, on ne réalise pas une recherche exhaustive, souvent impossible de par son coût calculatoire, mais par échantillonnage.

RANSAC est une procédure itérative qui repose sur deux étapes : a) choisir aléatoirement un échantillon minimal d'appariements afin d'estimer les paramètres du modèle géométrique, et b) construire un sous-ensemble d'appariements qui est compatible avec cette géométrie. Cet ensemble est appelé *ensemble de consensus*. La compatibilité est mesurée par la taille de l'ensemble de consensus (pour la version originale de RANSAC) ou à l'aide d'une mesure d'adéquation au modèle plus sophistiquée (MSAC, MLESAC). La recherche est arrêtée quand un ensemble de consensus de taille suffisante est trouvée, ou lorsqu'un nombre suffisant de tirages a été effectué pour garantir la fiabilité de la solution.

L'approche RANSAC est très populaire en vision par ordinateur pour déterminer les paramètres d'une transformation géométrique fixée, dont la nature est connue a priori. La mise en correspondance de plans, via l'homographie, et celle de deux vues d'une même scène, via la matrice fondamentale, ont donné une littérature riche (voir section 2.4.3) cherchant à parfaire l'utilisation de cette approche. L'algorithme 2 est une version simple pour l'estimation de la matrice fondamentale.

Une difficulté de RANSAC est d'évaluer le nombre suffisant de tirages N pour assurer qu'à une probabilité p, au moins 1 échantillon de taille s n'est pas pollué. Si le taux d'inliers w est connu, alors la probabilité de choisir un échantillon non pollué est w^s, on en déduit pour N tirages la probabilité de choisir un échantillon pollué $(1-w^s)^N$, probabilité de tirer N fois un s-échantillon contenant au moins 1 outlier. On a alors $(1-w^s)^N \leqslant 1-p$, et en fixant p :

$$N \geqslant \frac{\log(1-p)}{\log(1-w^s)} \tag{2.15}$$

La réussite de RANSAC dépend ainsi de ces deux paramètres, le seuil δ d'appartenance à un modèle, ainsi que le nombre de tirages N. Idéalement, on souhaiterait filtrer des ensembles contenant beaucoup de données en examinant le plus d'hypothèses possibles, mais il est d'abord nécessaire d'améliorer la robustesse pour de forts taux de données aberrantes et l'efficacité de cette approche.

2.4.3 Évolutions de RANSAC

Afin d'améliorer RANSAC, et de résoudre les difficultés de choix des deux seuils précédemment décrits, les perfectionnements de RANSAC s'articulent autours de plusieurs axes :

- utiliser de meilleures fonctions d'évaluation du consensus : comment déterminer le seuil sur la distance au modèle ou la taille minimale d'un ensemble pour considérer les paramètres du modèle comme corrects vis à vis des données ? Cela peut être réalisé en modélisant la distribution des outliers (MINPRAN [Ste95]), en estimant le nombre de fausses alarmes avec les méthodes *A Contrario* : AC-RANSAC [MS04b], MAC-RANSAC [RDGM10]), ou encore en distinguant les inliers selon leur qualité (en privilégiant les inliers les plus précis par rapport au résidu géométrique MLESAC [TZ00]).
- obtenir une solution plus rapidement avec des techniques d'échantillonnage différentes :
  - uniforme, ou biaisé suivant une loi aléatoire fixée en utilisant une connaissance a priori [TM02, CM05] (issue de la phase d'extraction) sur la qualité des données pour guider la recherche. Il est aussi possible de biaiser l'échantillonnage en ajoutant des informations de contexte :

Algorithme 2 : RANSAC – Déterminer la matrice fondamentale entre deux images à l'aide de son ensemble de consensus

Entrées : Ensemble S de n appariements avec $n \geqslant 7$
Sorties : Ensemble de consensus $\overline{S}$
début
  répéter
    échantillonner aléatoirement (tirage uniforme) un ensemble T de 7 appariements parmi S
    estimer les matrices fondamentales associées (au plus 3)
    pour chaque *matrice fondamentale* $\mathbf{F}$ *associée à* T **faire**
      le calcul de la distance dist($\mathbf{m}'$, $\mathbf{Fm}$) pour chaque appariement candidat
      déterminer l'ensemble d'inliers cohérents avec $\mathbf{F}$ en réalisant le test $d_{\mathbf{F}} < \delta$
      $\overline{S} \leftarrow$ le groupe avec le plus grand nombre d'inliers, dit *ensemble de consensus*
  jusqu'à *ce que le nombre d'essais soit supérieur à N ou que* $\overline{S}$ *soit suffisamment grand*
fin

par exemple, en groupant les points d'intérêt à l'aide d'une segmentation, ou d'une mesure du flot optique [NJD09].

– avec des optimisations d'ordre combinatoire, en ne construisant pas l'ensemble de consensus lorsque ça n'en vaut pas la peine. Le test $T_{d,d}$ [CM02] vérifie l'hypothèse courante avant d'évaluer son consensus sur d données choisies au hasard, et l'accepte uniquement si toutes les données d sont inliers pour ce modèle. Nistér propose dans [Nis05] d'utiliser un test $T_{c,d}$, où seules c réalisations parmis d sont nécessaires à la réussite du test préemptif. Il évite ainsi de rejeter de nombreux échantillons corrects, améliorant ainsi l'idée de [CM02]. Le test Bail-Out [Cap05] évalue à la volée le taux d'inliers d'une hypothèse afin d'estimer la probabilité que l'échantillon courant permette d'améliorer la solution trouvée, i.e. de remplacer la précédente meilleure solution. Le test SPRT [CM08] permet d'estimer d'une façon analogue la vraisemblance de l'hypothèse courante.
– en contraignant la recherche en un temps fixé, avec Preemptive-RANSAC [Nis05], et ARRSAC [RFP08].
– en mettant en œuvre différentes stratégies d'optimisation de l'estimation : recherche locale Lo-RANSAC [CMK03], raffinements itératifs [CM05, SM06, GS06], ou encore ajouts de nouvelles correspondances guidées par la géométrie, une fois une solution initiale trouvée : Tordoff et Murray dans [TM02] et dans [TM05]. Dans ce dernier article, ils proposent une stratégie alternative de filtrage robuste permettant de choisir parmi plusieurs hypothèses d'appariements lorsqu'il existe une ambiguïté pendant la construction de l'ensemble de consensus.

On note que dans toutes ces méthodes, les ensembles d'appariements candidats sont toujours fournis par l'étape de mise en correspondance, puis filtrés. L'ajout de la contrainte géométrique pour la sélection des points ne permet pas de récupérer ceux qui sont rejetés par erreur lors de la mise en correspondance à l'aide de la photométrie, ce qui est fatal pour une scène contenant des motifs répétés.

Nous allons maintenant présenter les extensions de RANSAC les plus proches de nos travaux.

2.4.3.1 Affiner l'évaluation du consensus

Nous présentons ici une amélioration de RANSAC considérée indispensable par Torr et Zisserman [TZ00], et qui est d'ailleurs reprise par la plupart des améliorations abordées précédemment. Ils proposent de palier à un des problèmes de RANSAC qui considère tous les appariements inliers équitablement quelle que soit leur erreur par rapport au modèle. Une telle méthode revient à fixer la fonction de coût suivante :

$$\rho(e^2) = \begin{cases} 0 & e^2 < \delta^2 \\ 1 & e^2 \geq \delta^2 \end{cases} \tag{2.16}$$

dans le but de minimiser $\sum \rho(e^2)$.

Sans entraîner de calculs supplémentaires, on peut remplacer la fonction de coût ρ par ρ_2 telle que :

$$\rho_2(e^2) = \begin{cases} e^2 & e^2 < \delta^2 \\ T^2 & e^2 \geq \delta^2 \end{cases} \tag{2.17}$$

Cette approche est appelée MSAC (*m-estimator sample consensus*).

Cherchant à affiner le critère de décision utilisé, MSAC est une première étape pour réaliser une estimation de l'erreur au maximum de vraisemblance, plutôt que d'analyser la qualité d'un ensemble d'inliers trouvés uniquement en fonction de sa taille.

2.4.3.2 Modifier la technique d'échantillonnage

Chum et Matas proposent dans [CM05] une modification du paradigme RANSAC. Plutôt que de traiter tous les appariements de la même manière, ils proposent dans leur approche *Progressive Sample Consensus* (PROSAC) d'exploiter les informations sur la ressemblance entre points mis en correspondance, suivant ainsi l'idée proposée par Tordoff et Murray dans [TM02]. Ils sélectionnent ainsi les échantillons sur un ensemble de correspondances les mieux appariées, ensemble qu'ils élargissent petit à petit. Leur approche est intéressante au sens qu'elle combine une recherche qu'on espère plus rapide du consensus, avec une fonction d'évaluation du consensus prenant en compte la probabilité de fausses alarmes, comme dans l'approche *A Contrario* de Moisan et Stival [MS04b]. Le choix de cet ensemble est un compromis entre l'hypothèse trop optimiste d'une bonne qualité du tri effectué par la phase d'appariements, et la trop pessimiste approche RANSAC qui traite toutes les correspondances de la même façon. Au lieu d'utiliser la probabilité qu'une correspondance est correcte, ils utilisent l'hypothèse a minima d'un lien entre probabilité de ressemblance et la distance entre descripteurs. De telles séquences de correspondances sont appelées *not-worse-than-random* (pas pires qu'aléatoires).

Deux conditions à satisfaire sont proposées pour l'arrêt de la recherche :

– *pas pire qu'aléatoire* : éviter de choisir une solution qui est cohérente "par hasard". On cherche à minimiser la probabilité $P_n(i)$ qu'un groupe de correspondances soit considéré comme correct alors qu'il ne l'est pas :

$$P_n(i) = \beta^{i-m}(1-\beta)^{n-i+m}\binom{n-m}{i-m}, \tag{2.18}$$

où m est la taille de l'échantillon minimal, n le nombre de correspondances à trier, i la taille de l'ensemble examiné et β est la probabilité qu'un modèle calculé à partir d'un échantillon contenant un outlier soit considéré comme correct alors qu'il ne l'est pas. Cette probabilité est supposée connue a priori. On en déduit un critère sur le nombre minimal d'inliers I_{n^*} que doit

comporter la solution :

$$I_{n^*} \geq min \left\{ j : \sum_{i=j}^{n} P_n(i) < \Psi \right\}, \tag{2.19}$$

où Ψ est un seuil (habituellement 5%). Ce critère est à rapprocher de l'évaluation du nombre de fausses alarmes par Moisan et Stival. Cependant, plutôt que d'évaluer ce nombre de façon purement combinatoire, il est ici nécessaire de fixer β a priori : cela revient à ajouter un paramètre à la méthode.

– *maximalité* : imposer que la probabilité qu'il existe un ensemble contenant plus d'inliers que celui trouvé après k tirages est inférieure à η_0 (autre seuil habituellement fixé à 5%). Il s'agit ici de l'estimation habituelle du nombre de tirage à effectuer lorsque l'on suppose connaître le taux d'outliers, telle qu'estimée avec l'équation (2.15) pour RANSAC (section 2.4.2).

En ce qui concerne l'algorithme, les échantillons sont choisis dans un sous-ensemble correspondant aux données de meilleures qualités. La taille de cet ensemble est ensuite augmentée peu à peu. Si m est la taille de l'échantillon, $m-1$ appariements sont situés dans le sous ensemble et le dernier appariement nécessaire à la construction de l'appariement est ajouté au sous-ensemble à l'étape suivante.

La limitation de cette approche est que l'évaluation de β reste un problème entier, et les deux tests d'arrêt nécessitent chacun un seuil.

2.4.3.3 Enrichir l'ensemble de consensus

Tordoff et Murray [TM02, TM05] discutent l'opportunité d'utiliser une approche à hypothèses multiples lors du filtrage robuste. Plutôt que de filtrer des couples de points d'intérêt (appariements 1-1) de chaque image, ils proposent qu'un point de la première image puisse avoir plusieurs candidats à la mise en correspondance dans la seconde (appariements 1-N). Ils différencient cette technique de filtrage d'une autre méthode, appelée *rematching*, qui consiste en un enrichissement à posteriori de la solution retenue par l'étape de filtrage robuste.

Pour eux, recourir à des hypothèses multiples se justifie lorsque le nombre de correspondances aberrantes est trop élevé et empêche de trouver le mouvement correct : dans ces conditions, le *rematching* échoue. Ils considèrent que ce cas arrive peu souvent et préfèrent le *rematching* pour son coût algorithmique moins élevé.

Ils proposent deux modifications pour inclure les appariements ambigus. La première, quand un appariement est sélectionné dans l'échantillon minimal, tous les autres appariements candidats faisant intervenir le même point d'intérêt dans la seconde image sont retirés. La seconde, lorsque l'on construit le consensus, pour chaque groupe de candidats 1-N à la mise en correspondance, tous les couples possibles sont testés et le plus probable, celui dont la distance à la droite épipolaire est la plus faible, est sélectionné.

Pour notre part, nous considérons la gestion des hypothèses multiples comme indispensable pour lever les ambiguïtés inhérentes à la présence d'aliasing perceptuel.

2.5 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art sur la mise en correspondance de points d'intérêt, en mettant l'accent sur la mise en correspondance entre descripteurs ainsi que sur les méthodes de filtrage robuste. En particulier, nous avons décrit l'approche SIFT que nous utilisons

ensuite à de nombreuses reprises. Les améliorations des méthodes RANSAC, permettent de traiter plus d'appariements potentiels tout en obtenant de meilleurs résultats. Nous en utilisons certaines (la gestion des hypothèses multiples, les tests préemptifs, une évolution du modèle *A Contrario* de Moisan et Stival) pour la mise en correspondance en présence de motifs répétés au chapitre 4, ainsi que combinées avec des points d'intérêt aux descripteurs plus invariants, comme dans ASIFT, au chapitre 6. Nous allons d'abord réaliser une validation de l'approche *a contrario* pour l'estimation robuste de correspondances et de la matrice fondamentale dans le chapitre 3.

Chapitre 3

A Contrario RANSAC - Évaluation

Ce chapitre présente quelques validations de l'approche A Contrario RANSAC *comparée à d'autres approches de type* RANSAC. *Nous y présentons des tests synthétiques au paragraphe 3.1, puis un test sur une séquence réelle au paragraphe 3.2.*

Lorsque l'on utilise des algorithmes suivant l'approche RANSAC, il est nécessaire de régler plusieurs paramètres à la main (la distance maximale pour déclarer une mesure compatible avec le modèle évalué, le taux d'outliers permettant de fixer le nombre d'itérations, et la taille minimale du meilleur ensemble de consensus pour considérer la recherche réussie), ce qui peut se révéler être une entreprise délicate. Récemment, Moisan et Stival [MS04b] ont proposé une nouvelle approche de type RANSAC afin d'estimer la géométrie épipolaire. Leur approche repose sur une mesure statistique qui ne nécessite pas de régler de paramètre et dont la performance est démontrée en présence de taux importants de données aberrantes. Nous avons présenté ce modèle dans le chapitre 1, et nous abordons maintenant la validation de ses performances. Rappelons (voir paragraphe 1.3.3.2) qu'il s'agit de trouver l'ensemble de consensus minimisant le nombre de fausses alarmes ε_2 :

$$\varepsilon_2(\alpha, n, k) := 3(n-7)\binom{n}{k}\binom{k}{7}\alpha^{k-7}. \tag{3.1}$$

avec α la rigidité de l'ensemble considéré, k sa taille, n le nombre total d'appariements.

L'étude réalisée par Moisan et Stival avait porté uniquement sur des appariements déterminés à la main. Pour valider leur méthode *A Contrario* RANSAC, nous avons réalisé à la fois des tests synthétiques et des évaluations dans le cas réel. Par rapport à l'algorithme présenté au chapitre 1, section 1.3.3.3, une optimisation, appelée ORSA dans [MS04b], est mise en place lorsqu'un groupe est trouvé avec un $\varepsilon_2 \leqslant 1$: dans cette variante, dès qu'un tel groupe est trouvé, les échantillons minimaux de 7 appariements sont tirés dans le précédent groupe le plus significatif plutôt que dans tout l'ensemble de candidats. Le nombre d'itérations effectuées pour l'optimisation est fixé à 1000 dans toutes les expériences.

Ce travail a été l'objet d'une publication dans [NSB07b]. Les tests synthétiques ont permis d'évaluer les gains obtenus en robustesse, précision et vitesse par rapport aux approches Ransac classiques. Avec l'étude dans le cas réel, nous avons vérifié l'efficacité de la méthode à partir d'appariements extraits à partir des descripteurs SIFT [Low04].

3.1 Tests synthétiques

3.1.1 Protocole expérimental

Le banc de test est réalisé en utilisant un corpus de 1400 appariements synthétiques : après un tirage uniforme des coordonnées 3D dans un intervalle donné, le point est projeté dans les deux plans images ; si chacun des points 2D trouvés est respectivement dans la fenêtre du plan constituant l'image, le point est conservé et les coordonnées 2D ainsi calculées permettent d'établir un nouvel appariement (appelé *inlier*).

On applique ensuite un bruit uniforme sur les coordonnées 2D, d'une amplitude fixée. On simule ainsi le bruit d'extraction des appariements.

Enfin, on remplace une proportion du corpus par des *outliers*, points dont les coordonnées sont tirées uniformément sur le rectangle défini par les points d'abscisses et d'ordonnées minimales et maximales calculées lors de la génération des appariements.

Le calcul de la matrice fondamentale est effectué sur la première moitié des points, et une phase de ré-estimation non linéaire est effectuée après la sélection de l'ensemble d'inliers par les méthodes robustes. L'évaluation de la précision est réalisée sur la seconde moitié, en prenant en compte la moyenne des mesures de distance à l'épipolaire pour chaque appariement cohérent.

3.1.2 Test de robustesse

Le test est répété 100 fois pour chaque proportion d'outliers pour établir des taux de réussite. On fixe ici l'amplitude du bruit uniforme appliqué aux coordonnées à 1, ce qui correspond à une précision d'extraction plausible pour les points d'intérêts. L'évaluation de la robustesse est réalisée en calculant la distance moyenne à l'épipolaire pour les couples de points corrects de l'ensemble de validation. Si celle-ci est inférieure au bruit d'extraction maximum (ici égal à 1), on considère que le modèle (la matrice fondamentale) est correctement évalué.

L'algorithme AC-RANSAC montre sa supériorité en obtenant un taux de réussite proche de 1 jusqu'à 80% d'outliers (avec un autre protocole, on atteint même les 90% dans [MS04b]) contre au mieux 65% pour un RANSAC classique (cf. figure 3.1). Les seuils des algorithmes RANSAC et MSAC ont été réglés de manière à obtenir un meilleur comportement en moyenne. Rappelons que l'algorithme AC-RANSAC ne nécessite pas de seuil. Les algorithmes RANSAC 7 points et MSAC montrent leur supériorité sur l'approche utilisant un échantillon composé de 8 points : il est en effet nécessaire de considérer moins de cas pour estimer un modèle correct à partir de 7 points que pour 8.

3.1.3 Test de précision

La précision de l'estimation du modèle s'avère bien meilleure avec l'algorithme AC-RANSAC qu'avec les algorithmes RANSAC dès que le taux t d'outliers dépasse 50% (cf. figure 3.2).

3.1.4 Tests de vitesse d'exécution

On réalise ici des moyennes sur 30 expériences pour chaque précision d'appariement et taux de points aberrants.

Pour l'algorithme AC-RANSAC, la vitesse d'exécution dépend essentiellement de la présence d'outliers (cf. figure 3.3), tandis que pour l'algorithme RANSAC, la précision des appariements joue un rôle prépondérant (cf. figure 3.4). La phase d'optimisation réalise un nombre fixé d'itérations, sans test d'arrêt : le temps d'exécution minimal d'AC-RANSAC est donc plus long.

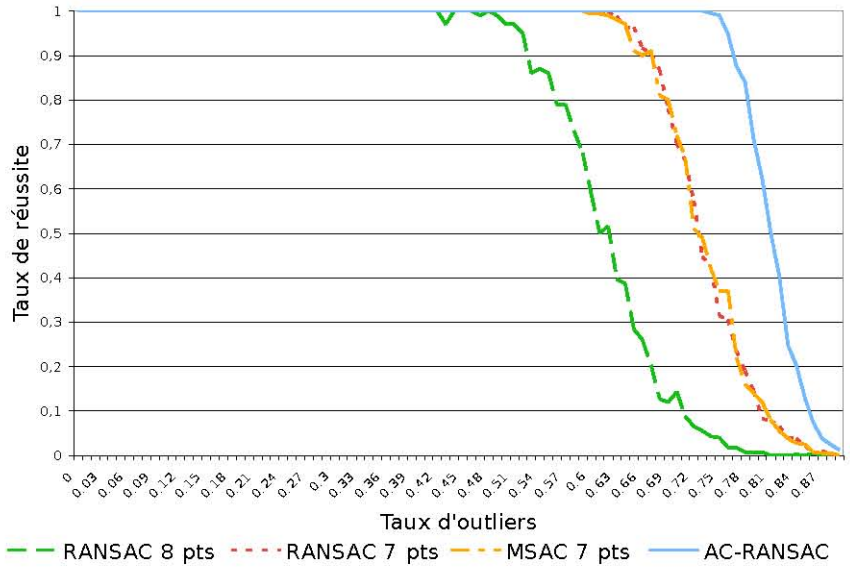

FIGURE 3.1 – Taux de réussite dans le calcul de la matrice fondamentale en fonction du taux d'outliers dans la base d'appariements. L'algorithme AC-RANSAC présente une meilleure robustesse que l'évaluation de la matrice fondamentale par les algorithmes Msac et Ransac (que le modèle soit obtenu par la méthode des 7 points ou celle des 8 points).

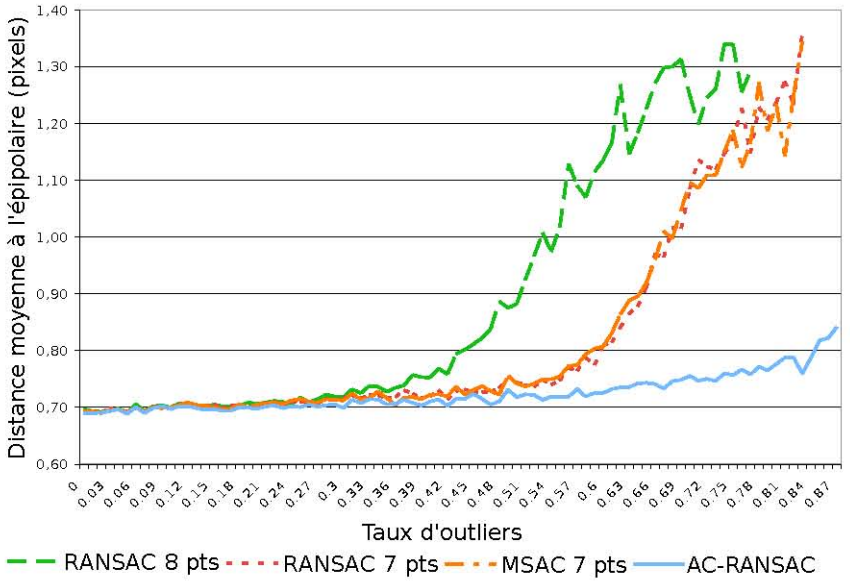

FIGURE 3.2 – Distance moyenne à l'épipolaire en fonction du taux d'outliers. L'algorithme AC-RANSAC permet une précision bien meilleure que les algorithmes RANSAC et MSAC dès que le taux d'outliers dépasse 50%.

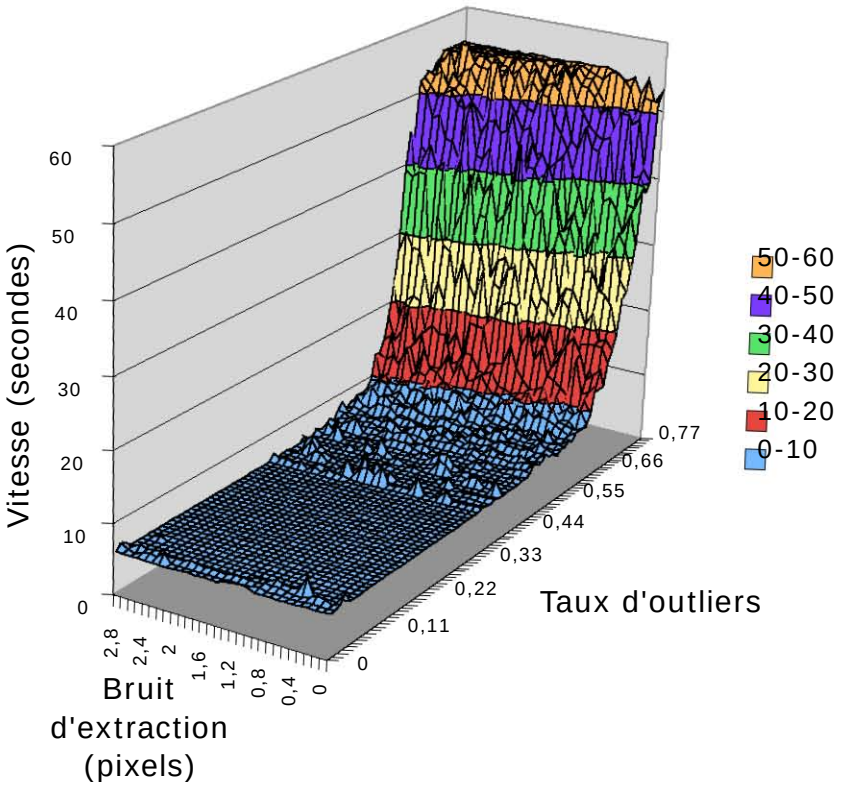

FIGURE 3.3 – Temps d'exécution de l'algorithme AC-RANSAC en fonction du bruit d'extraction sur les points d'intérêt et de la proportion d'outliers. Le temps d'exécution est stable par rapport au bruit, et augmente avec le taux d'outliers. Ceci s'explique par le plus grand nombre d'itérations alors nécessaire avant de trouver un ensemble 1-significatif qui permet de rentrer dans la phase d'optimisation.

Par contre, pour AC-RANSAC, le temps d'exécution ne varie pas en fonction de la précision des appariements : c'est une conséquence de l'absence de seuils. Le paramètre ε_2, qui inclut l'évaluation de la qualité du tirage, permet un arrêt de l'évaluation avec un ensemble de consensus plus petit car de précision garantie. La meilleure résistance aux outliers d'AC-RANSAC se paie par l'augmentation du temps de calcul pour des valeurs du taux d'outliers plus élevées que pour Ransac.

Le gain en vitesse global pour AC-RANSAC est dû à la phase d'optimisation. Le nombre de fausses alarmes, alors en dessous de 1, garantit l'existence d'une solution correcte et permet de ne considérer ensuite qu'un nombre réduit de cas.

3.2 Tests sur une séquence vidéo

Après cette étude sur des données synthétiques, on évalue maintenant le comportement de l'approche AC-RANSAC pour une séquence vidéo, en utilisant le détecteur SIFT pour déterminer des points d'intérêt et les appariements entre ces points (voir figure 3.5). On réalise des tests sur une série d'images dont on connaît par ailleurs les paramètres des caméras : la vérité terrain permet de calculer la matrice fondamentale théorique. Il est alors possible d'évaluer la qualité des appariements sélectionnés en calculant la distance à l'épipolaire pour cette matrice fondamentale. On compare les moyennes de cette distance réalisée sur tous les inliers, pour chaque couple d'images.

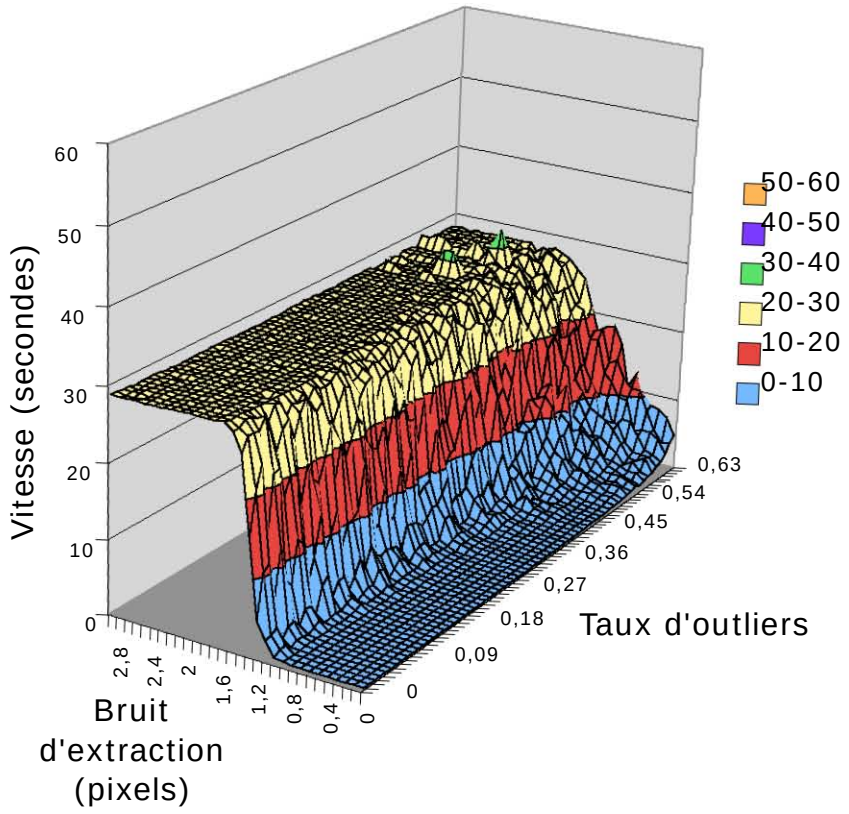

FIGURE 3.4 – Temps d'exécution de l'algorithme RANSAC en fonction du bruit d'extraction sur les points d'intérêt et de la proportion d'outliers. Le temps d'exécution augmente légèrement avec le taux d'outliers mais dépend fortement de la précision sur les points d'intérêt. On observe un effet de seuil lorsque le bruit devient relativement plus élevé que le seuil sur la distance au modèle au dessus duquel les échantillons sont rejetés de l'ensemble de consensus.

FIGURE 3.5 – Exemple d'appariements obtenus pour un des couples d'images de la séquence. On estime la matrice fondamentale sur des couples séparés par un intervalle de 50 images. Les segments verts représentent le mouvement apparent.

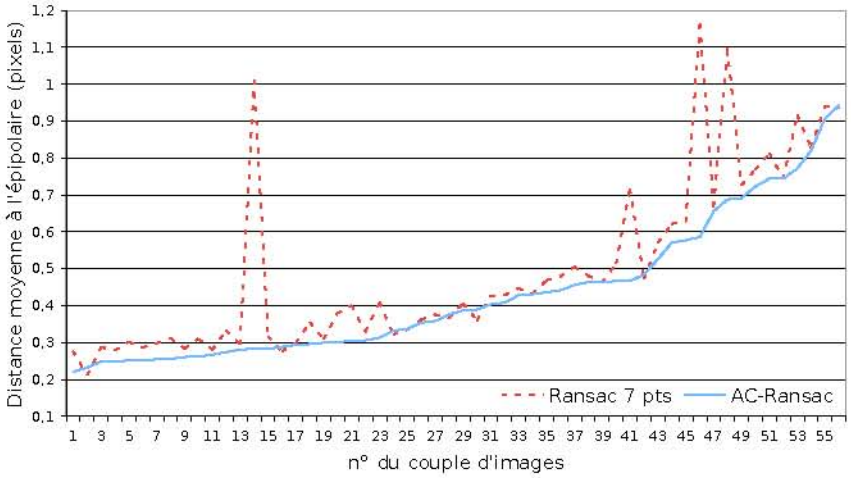

FIGURE 3.6 – Précision obtenue pour la séquence vidéo : l'algorithme AC-RANSAC donne systématiquement des résultats meilleurs que l'algorithme RANSAC 7 points. Les pics indiquent les images de la séquence pour lesquelles les seuil fixés pour RANSAC sont inadaptés, ce qui a alors un impact sur la précision de la solution trouvée.

La figure 3.6 présente les précisions des modèles calculés en fonction du numéro du couple d'images considéré. Le graphique a été réalisé en classant les couples d'images en fonction de la précision obtenue par le modèle *a contrario*. Afin de réaliser une comparaison équitable, les seuils pour RANSAC ont été choisis "au mieux". On obtient une confirmation des résultats synthétiques : au mieux, la précision de la méthode RANSAC est celle de l'approche AC-RANSAC. Les seuils choisis pour RANSAC permettent un bon comportement dans la plupart des cas, mais donnent une évaluation du modèle très mauvaise dans certains cas. Encore une fois, AC-RANSAC n'impose de régler aucun seuil.

3.3 Conclusion

Dans ce chapitre, nous avons réalisé une validation de l'algorithme AC-RANSAC [MS04b] pour l'estimation robuste de correspondances sous contrainte épipolaire. Cependant, cette avancée ne suffit pas pour résoudre des cas difficiles, tels la figure 3.7, que nous souhaitons traiter dans la suite. Les motifs de la moquette sont répétés, et la mise en correspondance utilisant uniquement l'infor-

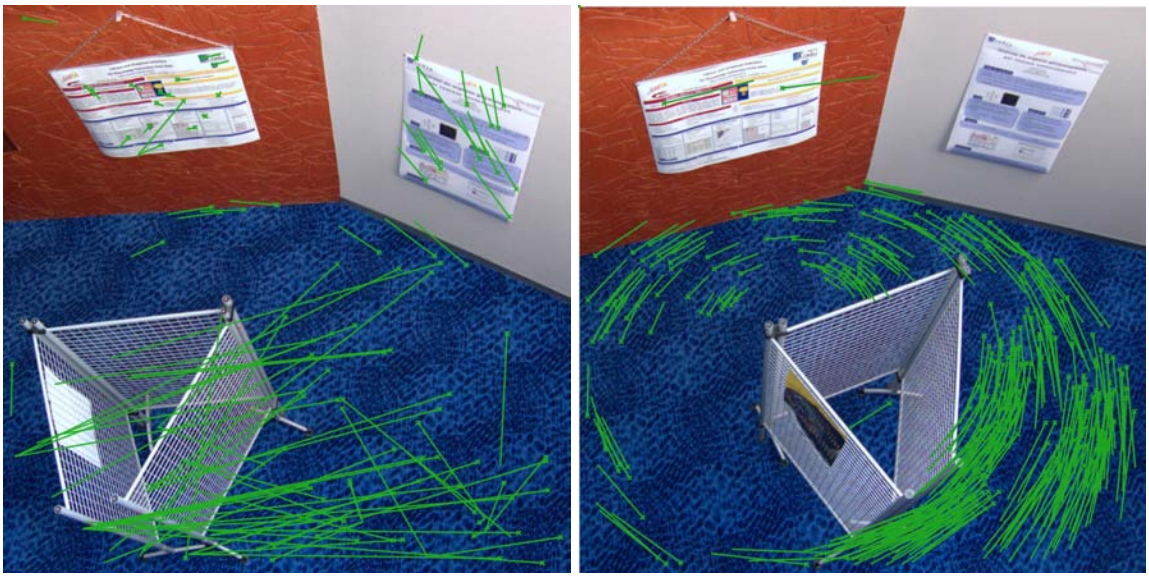

FIGURE 3.7 – *Loria*. Les points d'intérêt SIFT sont signalés par une croix, le segment représente le mouvement apparent avec le point d'intérêt apparié dans l'autre image. Sur la gauche, l'approche SIFT standard pour la mise en correspondance (NNT dans le texte) produit sur la moquette des correspondances de points d'intérêt incorrectes pour la plupart. Sur la droite, l'approche que nous allons développer dans les chapitres suivants, qui utilise à la fois les contraintes photométriques et géométriques (ici une homographie), trouve des appariements corrects en dépit d'une forte ambiguïté perceptuelle.

mation photométrique ne suffit pour produire des appariements corrects. Ainsi, aucun algorithme de sélection robuste ne réussira à extraire les appariements corrects vus que les appariements à filtrer sont pour la plupart aberrants, comme montré sur l'image de gauche. Il est pourtant possible d'extraire des appariements cohérents, visibles sur l'image de droite.

Nous allons maintenant montrer comment réaliser l'estimation de mouvement sans candidats à l'appariement, ce qui permet de ne pas décider de façon prématurée de la fiabilité des appariements en présence d'ambiguïté due à l'aliasing perceptuel, et obtenir des correspondances fiables en présence de motifs répétés.

Chapitre 4

Modèle A Contrario, mise en correspondance robuste sans appariement a priori

Ce chapitre traite des schémas alternatifs à l'utilisation classique d'appariements de points d'intérêt et à leur filtrage robuste par une méthode de type RANSAC *afin d'estimer le mouvement entre deux vues, tel que présenté dans le chapitre 2 à la section 2.1. Après un état de l'art 4.1 sur les méthodes d'appariements, nous présentons un modèle* A Contrario *mêlant contraintes épipolaires et photométriques en section 4.2, puis terminons ce chapitre en précisant les algorithmes et choix techniques en section 4.3. La validation expérimentale sera présentée au chapitre suivant.*

Afin d'apparier des points d'intérêt munis de descripteurs locaux, il est nécessaire de disposer :
- d'une distance entre descripteurs,
- d'un critère d'appariement utilisant cette distance.

Nous recherchons une distance suffisamment fiable, telle que les descripteurs de voisinages perceptuellement semblables soient à des distances faibles. Ce n'est pas toujours le cas, et ce pour plusieurs raisons :
- l'invariance limitée aux changements d'illumination,
- la présence d'occultations sur les images,
- les distances usuelles (L_1, L_2, χ^2) sont sensibles aux erreurs de localisation ou d'orientation des descripteurs.

Si on se contente d'une étape de mise en correspondance reposant uniquement sur le plus proche voisin conditionné par un seuil sur le rapport de distance entre plus proche voisin et second plus proche voisin (comme dans l'approche d'appariement SIFT habituelle), les ambiguïtés sont retirées lors de cette étape (leur rapport sera en effet toujours proche de 1). De plus, il n'y a aucune garantie que l'appariement qui est conservé soit correct, comme mis en évidence sur la figure 3.7.

Une façon d'éviter cet écueil est d'utiliser un modèle statistique afin d'apparier les descripteurs, en prenant en compte les difficultés de comparaison d'histogrammes circulaires. Rabin et al. [RDG09a] présentent une méthode *A Contrario*, que nous décrivons succinctement en section 4.1.1 pour débuter notre état de l'art. Cette méthode permet d'obtenir de meilleurs performances pour l'appariement par rapport au seuil sur le rapport de distance, mais il reste très difficile de mettre en correspondance d'une façon fiable des motifs répétés ou des points décrits de façon ambiguë entre images.

Afin de pouvoir réaliser la mise en correspondance et le calcul du mouvement dans ces cas

difficiles, il est nécessaire de garder suffisamment de candidats à l'appariement, plutôt que de les rejeter a priori s'ils sont ambigus : l'appariement devient alors un problème combinatoire à la complexité importante. Cet aspect fait l'objet de la deuxième partie de notre état de l'art, présenté en section 4.1.5.

Il s'agit ici d'une vision extrême des approches de mise en correspondance : d'un côté, un schéma en plusieurs étapes avec des seuils éliminant les appariements ambigus, de l'autre une conservation de toutes les données et une recherche combinatoire coûteuse. Cette dernière approche est dite "sans correspondance" dans la suite. Ces deux approches sont complémentaires. Relaxer les seuils permet de conserver l'ambiguïté tant que l'on a pas assez d'information pour prendre une décision, et de limiter la complexité par des heuristiques guidant la recherche. C'est d'ailleurs la voie suivie par Hsiao et al. [HCH10] dans le domaine de la reconnaissance d'objets, qui élargissent l'ensemble des hypothèses afin de laisser la mise en correspondance plus libre.

4.1 État de l'art : méthodes de mise en correspondance

Nous présentons dans cette section d'abord une méthode permettant la sélection d'appariements candidats via une approche utilisant des descripteurs photométriques sans tenir compte de contrainte géométrique, puis des méthodes d'estimation du mouvement reposant sur des stratégies combinatoires, guidées par la géométrie et / ou la photométrie.

4.1.1 Distance CEmd de Rabin et al. et modèle *A Contrario* associé

Rabin et al. [RDG09a] présentent une approche pour la mise en correspondance entre points d'intérêt munis de descripteurs locaux qui se compare favorablement à la méthode du rapport des plus proches voisins. Pointant la faiblesse de ce critère en présence d'ambiguïtés, sa méthode cherche à rendre plus précis le filtrage des points dans ce cas en utilisant une distance du terrassier circulaire, qui prend en compte la particularité des descripteurs.

Afin de mesurer la dissimilarité d'un couple de descripteurs $(D(x), D(y))$, composés chacun de N histogrammes $D^i(x)$ et $D^i(y)$, $\mathsf{dist}(D(x), D(y)) = \sum_{i=1}^{N} \widetilde{\mathsf{dist}}(D^i(x), D^i(y))$, Rabin et al. définissent l'hypothèse nulle $\mathscr{H}_0^j$ telle que chacune des distances $\widetilde{\mathsf{dist}}(D_j^i(x), D^i(y))$ est une variable aléatoire, qui sont toutes mutuellement indépendantes, et dont on note les densités de probabilité p_i^j, j étant l'indice indiquant qu'on parcourt tous les descripteurs candidats.

Sous hypothèse nulle $\mathscr{H}_0^j$, la probabilité que la distance entre deux descripteurs soit plus petite qu'un seuil quelconque mais fixé δ s'écrit :

$$\Pr(\mathsf{dist}(D_j(x), D(y)) \leqslant \delta | \mathscr{H}_0) = \int_0^{\delta} \bigotimes_{i=1}^{N} p_i^j(x) dx \tag{4.1}$$

avec $\bigotimes$ le produit de convolution. En effet, la densité de probabilité de la somme de variables réelles indépendantes est le produit de convolution de leurs densités.

Pour définir le nombre de fausses alarmes, il est nécessaire de définir le nombre total de tests où l'on peut observer l'évènement correspondant à l'hypothèse nulle. Ici, comme il s'agit d'appariements, celui-ci est égal au produit du nombre de points d'intérêt détectés (N_1 et N_2) dans chacune des deux images.

$$\mathrm{NFA} := N_1 N_2 \Pr(\mathsf{dist}(D(x), D(y)) \leqslant \delta | \mathscr{H}_0^i) \tag{4.2}$$

Ce nombre de fausses alarmes permet de fixer un seuil pour déterminer quels candidats à l'appariement sont retenus, après une estimation empirique des p_i^j.

4.1.2 Réaliser la mise en correspondance par une méthode à base de graphes

4.1.3 Contraintes de voisinage et relaxation

Avant la généralisation de l'utilisation des descripteurs pour la phase d'appariement, les méthodes à base de corrélation ont été souvent utilisées. Ces méthodes sont peu fiables dès lors que le changement d'apparence s'intensifie. Il est alors nécessaire de guider la recherche par des contraintes locales (nombre de points détectés dans le voisinage, orientation relative de ceux-ci, points avec un mouvement apparent proche...). Ces approches sont appelées méthodes d'appariement par relaxation. Par ailleurs, même avec l'utilisation de descripteurs invariants, on peut souhaiter ajouter de l'information supplémentaire via ces contraintes.

La relaxation, proposée par Rosenfeld et al. dans [RHZ76] est une méthode d'optimisation qui utilise l'information fournie par le voisinage de chaque point pour accroître la *cohérence* et réduire l'*ambiguïté* de la mise en correspondance. Plutôt qu'appliquer la stratégie classique du "winner takes all", cette stratégie évite de retenir de prime abord les appariements avec un bon score de similarité lorsque ceux-ci sont ambigus.

Zhang et al. [ZDFL95] ainsi que Schmid et Mohr [SM97] et Tell et Carlsson [TC02] ont recours à des contraintes locales pour la mise en correspondance de points d'intérêt. La différence principale dans ces méthodes est la mesure de similarité employée : Schmid utilise la conservation des angles. l'angle défini entre le point courant et deux points voisins doit être constant dans toutes les vues où ils apparaissent. Tell recherche la conservation de l'ordre cyclique d'attributs d'intensité aux voisinages des points. Zhang pondère le score de corrélation calculé sur une fenêtre représentant le voisinage du point en le multipliant par le nombre d'appariements dont les positions relatives sont identiques. On remarque que ces mesures de similarité combinent les contraintes géométriques et photométriques, mais qu'elles ne sont pas invariantes à des déformations ayant plus de degrés de liberté qu'une similitude.

Ces méthodes ont presque disparu depuis l'apparition des descripteurs à base d'histogrammes et leurs meilleures performances. Cependant, Sidibe et al. [SMJ08] a montré que la relaxation permet d'obtenir une meilleure mise en correspondance au moyen de la similarité obtenue via les distances entre descripteurs, en particulier en présence de fortes ambiguïtés : les correspondances ambiguës ne sont en effet pas rejetées une fois pour toute comme avec la méthode du rapport de distances, et la relaxation permet de ne pas toujours apparier les points avec les descripteurs les plus ressemblants ensemble.

Deng et al. [DMSD06] propose de pondérer les distances entre descripteurs en fonction du nombre de points d'intérêt détectés au voisinage du couple pour lequel on est en train d'estimer la distance entre descripteurs, ce qui permet de différencier des appariements qui auraient eu des distances proches entre descripteurs autrement.

4.1.4 Graphes d'adjacence

Gold et Rangarajan [GR96] présentent une une approche de mise en correspondance de graphes. Appliquée aux points d'intérêts, le but est d'apparier les graphes d'adjacence de ces points, en considérant que la topologie de ceux-ci ne varie que modérément entre les images. Cette méthode est proche des méthodes de relaxation, mais réalise l'affectation des correspondances dans les deux sens. Elle impose qu'un sommet d'un graphe soit apparié à au plus un sommet dans l'autre graphe, ce qui contraint plus la solution que dans les méthodes de relaxation, où la contrainte n'est vérifiée que dans un seul sens. Le problème de l'appariement revient donc ici à déterminer une matrice de permutation. Pour déterminer cette matrice, ils introduisent des suites paramétrées pour chaque coefficient, et montrent par des propriétés de convergence qu'ils obtiennent ainsi une matrice de

permutation. Un algorithme appelé *Softassign* [Sin64] est utilisé afin de satisfaire cette condition. Suivant les applications, l'initialisation de cette matrice peut être réalisée différemment. Une façon est par exemple d'utiliser des scores de ressemblance des voisinages des points considérés comme poids M.

Le fonctionnement de l'algorithme est le suivant. Avec un ensemble de N variables $\{X_{ai}\}$ de $\mathbb{R}$, on associe des poids $M_{ai} \in \{0,1\}$ tels que $\forall a, \sum_{i=1}^{N} M_{ai} = 1$ et $\forall i, \sum_{a=1}^{N} M_{ai} = 1$. On cherche à trouver (pour les colonnes) une matrice de permutation telle que :

$$M_{aj} = \begin{cases} 1 & \text{si } X_{aj} \text{ est le maximum des } \{X_{ai}\} \\ 0 & \text{autrement} \end{cases} \tag{4.3}$$

Et de même en faisant varier a (pour les lignes).

Pour trouver celle-ci, on utilise un algorithme itératif qui calcule les coefficients de la matrice via des suites paramétrées. Le problème continu à résoudre est le suivant : avec le paramètre de contrôle $\beta > 0$, qui sert à rapprocher petit à petit au cours du processus itératif la matrice courante d'une matrice de permutation, on détermine M tel que :

$$M_{aj} = \frac{\exp(\beta X_{aj})}{\sum_{i=1}^{N} \exp(\beta X_{ai})} \tag{4.4}$$

La normalisation alternative des M_{aj} suivant les lignes et les colonnes permet de satisfaire les contraintes imposées, grâce au résultat de [Sin64] : toute matrice carrée dont tous les éléments sont positifs converge vers une matrice doublement stochastique en itérant la normalisation des lignes et des colonnes. La gestion des outliers est réalisée en ajoutant une colonne et une ligne supplémentaire qui servent à marquer quand aucune correspondance n'est trouvée pour le point courant.

Dans [GR96], cette méthode est appliquée sur des graphes synthétiques générés aléatoirement, ou sur un graphe extrait manuellement constitué de différents types de sommets représentant des points situés sur des lignes droites, courbes, ou encore à des discontinuités. Les arêtes extraites entre les sommets dépendent de la topologie de l'image : sont reliés les points situés sur la même ligne, ou voisins sur celle-ci, ou séparés par une faible distance définie par un seuil. Dans [GRL+98], cette approche est réutilisée, mais cette fois-ci en utilisant un contrainte géométrique globale (une transformation affine). Les limitations de cette méthode sont qu'elle ne produit pas toujours une matrice de permutation (la convergence étant lente, ils se contentent d'un nombre limité d'itérations), et qu'elle n'est pas robuste aux changements d'apparence qui modifient la topologie du graphe, ce qui arrive rapidement en pratique.

Aguilar et al. [AFE+09] proposent de remplacer l'étape de filtrage robuste en recherchant une mise en correspondance des points d'intérêt via une méthode, appelée *Graph Transformation Matching*, reposant sur un graphe d'adjacence. Ce graphe d'adjacence est un graphe reliant chaque point à ses K (fixé à 4 ou 5 suivant les expériences) plus proches voisins dans l'image, la proximité étant limitée par la médiane des distances entre points de l'image. L'hypothèse utilisée est que le changement d'aspect est suffisamment limité pour conserver les relations de voisinages entre points détectés. L'algorithme consiste alors à éliminer itérativement les appariements qui perturbent ces relations de voisinage. Cette méthode a l'avantage d'être robuste à la présence d'outliers, avec des performances analogues à RANSAC, voire supérieures pour les forts taux d'outliers (supérieurs à 75%). Elle ne prend cependant pas en compte de contrainte géométrique forte.

4.1.5 Méthodes d'appariement en une étape

Ces méthodes sont dites en une étape car elles ne reposent pas sur un filtrage robuste d'appariements candidats extraits par une extraction bas-niveau.

Plusieurs articles cherchent à dépasser les difficultés de la mise en correspondance par étapes successives (comme décrit au chapitre 2) en procédant autrement. Une manière de résoudre le problème d'analyse de la structure et du mouvement sans correspondance est d'utiliser un échantillonnage systématique du mouvement, plutôt que guidé par les données, comme réalisé par les approches à la RANSAC vues précédemment au chapitre 2 en section 2.4. L'échantillonnage est alors réalisé :

– soit en considérant toutes les correspondances possibles parmi les points d'intérêt extraits,
– soit directement en explorant l'espace des mouvements possibles.

La difficulté repose alors souvent dans la complexité de la recherche, du fait du grand nombre d'hypothèses à explorer.

Approche par échantillonnage des correspondances La recherche d'appariements sans étape préalable de mise en correspondance à l'aide d'informations photométriques a été étudiée par Dellaert et al.[DSTT03] : leur approche est purement combinatoire, tous les points sont considérés comme correspondants potentiels. Leur contribution est la conception d'un algorithme EM (*Expectation Maximisation*) avec une étape de recuit simulé afin d'éviter les minima locaux pour l'erreur de reprojection. L'étape M consiste à résoudre le problème d'analyse de la structure et du mouvement pour un jeu d'appariement courant : pour les appariements supposés, la géométrie est calculée, soit pour un modèle orthographique, soit pour un modèle para-perspectif suivi d'un ajustement de faisceau. L'étape E est l'élaboration d'un nouveau jeu de données et consiste à réaliser un nouveau tirage des correspondances entre points d'intérêts. Ce tirage est réalisé en utilisant une heuristique de *Metropolis*, une méthode de Monte-Carlo pour les chaînes de Markov. Une hypothèse importante est que tous les points à mettre en correspondance sont visibles dans toutes les images, i.e. qu'il n'y a ni mesure aberrante ni occultation. Nous considérons que cette hypothèse est trop restrictive : les occultations et mauvaises détections se produisent souvent en pratique. De plus, la combinatoire du problème est vraiment limitante, et résoudre un problème d'analyse de la structure et du mouvement est rarement trivial.

Approches par échantillonnage du mouvement Un moyen de réduire la complexité du balayage de l'espace des transformations géométriques est d'utiliser de l'information photométrique combinée avec des contraintes géométriques.

Domke et Aloimonos [DA06] proposent une solution consistant à établir un modèle probabiliste a priori de la distribution des appariements. Pour chaque pixel de la première image, une carte des probabilités de correspondance dans la seconde image est établie.

La contrainte photométrique Ils utilisent des filtres de Gabor suivant des directions γ et des échelles l différentes, dont les réponses sont utilisées comme descripteur photométrique. La mise en correspondance est réalisée en supposant que les réponses aux filtres utilisés se ressemblent à échelle et orientation fixée : cette méthode ne supporte donc que de faibles modifications en rotation ou en échelle. Chaque probabilité est le produit des probabilités des réponses (phases $\phi_{l,\gamma}$ des filtres de Gabor) pour les pixels s les plus proches de $\hat{q}$; on en déduit l'expression de $\rho_s(\hat{q})$:

$$\rho_s(\hat{q}) \propto \prod_{l,\gamma} (\exp(-[\phi_{l,\gamma}(s) - \phi_{l,\gamma}(\hat{q})]^2_\pi) + 1) \tag{4.5}$$

L'addition de la constante 1 augmente la robustesse au bruit.

En pratique, les pixels ne se correspondent pas parfaitement d'une image à l'autre. Il est donc nécessaire d'estimer la probabilité qu'un point s corresponde non pas à un pixel $\hat{q}$, mais à un point quelconque q, qui n'a donc pas obligatoirement des coordonnées entières.

$$\rho_s(q) \propto \max_{\hat{q}} \rho_s(\hat{q}) \exp(-|q-\hat{q}|^2) + \alpha \tag{4.6}$$

α est une constante permettant de prendre en compte que la mesure de la phase peut être aberrante, en particulier en présence d'une occultation.

La contrainte géométrique Pour une hypothèse de mouvement (i.e. une matrice essentielle E) donnée, il est ensuite possible de calculer sa probabilité comme proportionnelle au maximum des probabilités de correspondance des appariements satisfaisant la contrainte épipolaire pour ce mouvement.

$$\rho_s(E) \propto \max_{q:q^t Es=0} \rho_s(q) \tag{4.7}$$

La contrainte épipolaire simple est retrouvée dans le cas où $\rho_s(q)$ vaut 1 pour les appariements connus et 0 pour les autres.

Contrainte finale et méthode de résolution En remplaçant $\rho_s(q)$ par son expression définie par l'équation (4.6), et en remarquant qu'il n'est pas nécessaire de déterminer précisément la position du point q, mais que la distance entre le point $\hat{q}$ et la droite Es suffit, on peut déduire :

$$\rho(E) \propto \prod_s (\max_q \rho_s(q) \exp(-(\text{dist}(q, Es)^2) + \alpha) \tag{4.8}$$

Des hypothèses de mouvement sont ensuite tirées au hasard en échantillonnant l'espace à 5 dimensions de tous les mouvements possibles (pour le cas calibré), via le tirage de 2500 hypothèses de mouvements différentes (choix d'un vecteur t et d'un angle ω sur la sphère unité tel que $E = [t]_\times R(\omega)$. Leur approche a un coût calculatoire élevé, comme pour [DSTT03]. Les auteurs insistent cependant sur le fait que la partie la plus coûteuse n'est pas liée à l'échantillonnage des mouvements, mais au calcul de la fonction de distribution des appariements. il est d'ailleurs possible d'accélérer l'échantillonnage des hypothèses quand une estimée courante du mouvement est disponible, celle-ci réduisant l'espace des solutions possibles.

Le même genre d'idée a été aussi utilisé précédemment par Roy et Cox [RC96], qui calculent la vraisemblance photométrique que le lieu des points est la droite épipolaire correspondante, et cherchent à la maximiser pour tous les mouvements possibles. Dans le même style, Antone et Teller [AT02] ont appliqué cette idée dans le contexte de réseaux d'images omni-directionnelles. Ils estiment la ligne de base entre deux vues en considérant toutes les correspondances possibles de points qui satisfont la contrainte épipolaire. Pour résoudre ce problème d'une grande complexité, ils contraignent la recherche en fonction de la ressemblance des caractéristiques des points.

Des travaux très récents discutent l'utilisation de la transformée de Radon pour estimer le mouvement sans correspondance. Lehmann et al. [LBC+07] s'intéressent au calcul de matrice fondamental dans le cas affine (i.e. pour le modèle de projection orthographique d'une caméra). Les transformées de Fourier de différentes vues sont reliées à travers les paramètres du mouvement de la caméra : les rotations et translations de l'image produisent une rotation du spectre et un décalage de phase de la transformée calculée. Les paramètres du mouvement sont déterminés en mettant en correspondance les droites dans le domaine de Fourier avec un algorithme EM dédié. Les résultats expérimentaux

sont prometteurs. Pour le moment, cette approche reste limitée au cas orthographique qui est moins complexe que le modèle épipolaire complet. Les occultations ne sont pas gérées par cette approche. La même idée gouverne le travail de Makadia et al. [MGD07] ; la transformée de Fourier est utilisée pour générer une fonction de vraisemblance globale sur l'espace de tous les mouvements de caméras possibles, ce qui est réalisé dans le domaine de Fourier, pourvu que la discrétisation soit choisie avec précaution. Les résultats sont présentés sur des images de caméras catadioptriques.

Toutes ces méthodes "sans-correspondance" calculent le mouvement de la caméra par une mise en correspondance globale, via une fonction de coût minimisée sur un espace dense de paramètres, et ne sont donc pas robustes aux occultations, ou quand seulement une petite partie est commune aux deux images.

Une autre façon d'échantillonner les mouvements possibles est proposée par Georgel et al. [GBN09]. Leur méthode discrétise l'espace des mouvements pour estimer simultanément appariements entre points d'intérêt et mouvement de la caméra. Ici, les appariements sont établis uniquement à l'aide de la contrainte géométrique via une mise en correspondance guidée par la contrainte épipolaire, sous réserve de respect d'une contrainte d'unicité d'utilisation des points d'intérêt. La photométrie ne sert que pour l'extraction des points. Cette approche est pour le moment limitée à des mouvements restreints de caméra dans le cas calibré : la caméra est posée sur une table, permettant translation et une rotation, ce qui fait 2 degrés de liberté pour la matrice essentielle, l'échelle étant alors inconnue.

Une autre façon d'incorporer les contraintes photométriques et géométriques pour l'analyse de la structure et du mouvement a été étudiée par Stein et Shashua dans [SS00]. Le flot optique fournit un champ de mouvement dense qu'on peut assimiler à de l'information photométrique mais souffre du problème d'ouverture (*aperture problem*), qui est gênant pour les scènes avec des longs bords droits, et souffre aussi de l'hypothèse d'intensité constante, qui est au contraire résolue par les descripteurs invariants aux changements de contraste tels que SIFT. Pour éviter ce problème, les auteurs de [SS00] proposent de construire une contrainte de *tensor brightness* qui repose à la fois sur le flot optique et le tenseur trifocal, qui encode la géométrie relative à trois vues [HZ00]. Cependant, comme toute méthode basée sur le flot optique, leur travail ne peut être étendu à de forts changements de points de vue, qui rendent l'estimation du flot optique incertaine. Cette méthode a l'avantage de ne nécessiter aucune correspondance de points d'intérêt.

Nos choix Utiliser des points d'intérêt, grâce à leur caractère local, permet d'éviter une des limitations des approches globales, leurs faiblesses en présence d'occultations. Au regard des difficultés liées à la complexité de la recherche de solution sans appariement a priori, ou de l'exploration directe de l'espace de mouvement, nous préférons utiliser un cadre où des candidats à l'appariement sont pré-filtrés, même si les appariements ne sont pas encore affectés du fait de la présence d'ambiguïté. Le filtrage des groupes de candidats à l'appariement est guidé par les données, à l'opposé d'une recherche systématique dont l'état de l'art nous a montré à quelle point elle est coûteuse. Le filtrage robuste est réalisé à l'aide d'un algorithme *A Contrario*-RANSAC combinant contrainte géométrique (la contrainte épipolaire pour le cas projectif que nous traitons) et photométrique – cette dernière contrainte profite des travaux de Rabin et al. rappelés plus haut (section 4.1.1) – et guidé par des heuristiques afin de limiter la combinatoire du problème.

4.2 Appariement *A Contrario* sous contraintes épipolaires et photométriques

Dans ce chapitre, nous proposons une méthode basée sur un modèle *a contrario*. Elle repose sur la méthode *A Contrario*-RANSAC d'appariement de points d'intérêt de Moisan et Stival [MS04b] vu au

chapitre 1 et par quelques aspects au modèle *A Contrario* de mise en correspondance de descripteurs de type SIFT par une distance du terrassier de Rabin et al. [RDG09a, RDGM10]

Comme vu au chapitre 1, le premier article [MS04b] se concentre sur les contraintes géométriques et suppose que les appariements entre points d'intérêt sont donnés par une étape préalable. Il nous indique aussi une piste pour établir des appariements, à partir de la géométrie et de la photométrie (appelé critère de *colored rigidity*). Notre contribution est une généralisation de l'algorithme de Moisan et Stival afin d'incorporer à la fois la contrainte épipolaire et la consistance photométrique. Nous explicitons aussi l'implantation ainsi que les heuristiques utilisées afin de limiter le coût calculatoire nécessaire pour réaliser la mise en correspondance. En particulier, cette mise en correspondance nécessite l'utilisation de distances adaptées aux comparaisons d'histogrammes. Les articles [RDG09a, RDGM10] prouvent que la distance du terrassier est plus performante que les mesures de (dis-)similarité existantes entre descripteurs SIFT, et analysent plusieurs modèles *A Contrario*. Cependant, la mise en place de contraintes géométriques est toujours considérée comme une étape à part après la mise en correspondance, et l'accent est mis sur la reconnaissance d'objet.

Définissons quelques notations. On suppose disposer de deux vues (les images $\mathscr{I}_1$ et $\mathscr{I}_2$) de la même scène. Pour chaque image, un algorithme (par exemple SIFT) donne un ensemble de points d'intérêt, munis de leurs descripteurs. Soient $(x_i, D(x_i))_{1\leq i\leq N_1}$ (resp. $(y_j, D(y_j))_{1\leq j\leq N_2}$) les N_1 (resp. N_2) couples de $\mathscr{I}_1$ (resp. $\mathscr{I}_2$) tels que x_i (resp. y_j) soit les coordonnées d'un point d'intérêt, et $D(x_i)$ (resp. $D(y_j)$) le descripteur local correspondant. Selon les circonstances, nous notons par x_i le point d'intérêt lui-même, ses coordonnées en pixel, ou ses coordonnées homogènes dans le plan projectif.

Nous supposons aussi que la position de la caméra a changé entre les vues, de telle façon que les points 3D correspondants aux points d'intérêt ne soient pas coplanaires. Nous traiterons ce cas particulier, l'homographie, dans la section 4.4. Ceci posé, si x_i et y_j sont les projetés dans $\mathscr{I}_1$ et $\mathscr{I}_2$ du même point 3D, alors y_j est sur la droite épipolaire associée au point x_i. Cette droite est représentée par son vecteur normal qui s'écrit $F \cdot x_i$, avec F la matrice fondamentale de $\mathscr{I}_1$ vers $\mathscr{I}_2$. De manière duale, le lieu de x_i est la droite épipolaire $F^T \cdot y_j$, la matrice fondamentale de $\mathscr{I}_2$ vers $\mathscr{I}_1$ étant la matrice transposée F^T. Si les descripteurs locaux étaient invariants aux transformations projectives, alors $D(x_i)$ et $D(y_j)$ seraient identiques. Cependant, une telle invariance n'est pas atteignable en pratique, comme vu au chapitre 2. On souhaite que $D(x_i)$ et $D(y_j)$ soient "suffisamment similaires".

Notre but est donc de trouver un sous-ensemble $\mathscr{S}$ de $\{1,\dots,N_1\}\times\{1,\dots,N_2\}$ et une matrice fondamentale F de $\mathscr{I}_1$ vers $\mathscr{I}_2$ minimisant le nombre de fausses alarmes du modèle *A Contrario*, ce qui fixe automatiquement les seuils δ_D et δ_G suivants. Ces sous-ensembles sont tels que :

1. La distance entre les descripteurs mis en correspondance est inférieure à un seuil déterminé automatiquement δ_D, garantissant que les voisinages locaux des points se ressemblent :

$$\forall (i,j)\in\mathscr{S},\ d_D(D(x_i),D(y_j))\leq\delta_D. \tag{4.9}$$

2. La distance entre un point de la première image et la droite épipolaire associée avec le point mis en correspondance dans la seconde image est inférieure à un autre seuil δ_G, lui aussi fixé automatiquement, garantissant que la contrainte épipolaire est satisfaite :

$$\forall (i,j)\in\mathscr{S},\ d_G(x_i,y_j,F) := \max\{d_G(y_j,F\cdot x_i), d_G(x_i,F^T\cdot y_j)\}\leq\delta_G. \tag{4.10}$$

Nous allons donner dans la suite en sections 4.2.2 et 4.2.3 une définition de ces deux distances (ou mesures de dissimilarité) d_D et d_G, ainsi que des seuils δ_D et δ_G, qui seront eux fixés automatiquement grâce au modèle *a contrario* proposé. Celui-ci équilibre automatiquement la géométrie et la photométrie, et permet de déduire les deux seuils relatifs à un ensemble $\mathscr{S}$.

4.2.1 Le modèle *A Contrario*

Avant de définir les distance d_D et d_G, nous décrivons le modèle statistique que nous utilisons pour prendre des décisions. Dans la méthodologie *A Contrario*, comme vue au chapitre 1, les groupes de points d'intérêt sont dits *significatifs* si leur probabilité est très faible sous l'hypothèse $\mathscr{H}_0$.

Supposons qu'un ensemble $\mathscr{S}$ d'appariements est fourni, ainsi qu'une matrice fondamentale F. Nous supposons aussi que la matrice fondamentale F est estimée à partir d'un échantillon minimal s parmi $\mathscr{S}$ comme pour le paradigme RANSAC. Nous pouvons alors définir l'hypothèse $\mathscr{H}_0$.

Définition 10. *Avec $(x_i, D(x_i))$ et $(y_j, D(y_j))$ des variables aléatoires, nous définissons l'hypothèse $\mathscr{H}_0$ comme :*

1. *$(d_D(D(x_i), D(y_j)))_{(i,j)\in\mathscr{S}}$, et $(d_G(x_i, y_j, F))_{(i,j)\in\mathscr{S}\setminus s}$ sont des variables aléatoires mutuellement indépendantes.*
2. *$(d_G(x_i, y_j, F))_{(i,j)\in\mathscr{S}\setminus s}$ sont identiquement distribuées et leur fonction de répartition empirique est notée f_G*
3. *$(d_D(D(x_i), D(y_j)))_{(i,j)\in\mathscr{S}}$ sont identiquement distribuées et leur fonction de répartition empirique est notée f_D.*

Bien sûr, $(d_G(x_i, y_j, F))_{(i,j)\in s}$ sont aussi identiquement distribués mais ne suivent pas la même fonction de distribution f_G que les variables de $\mathscr{S}\setminus s$ puisque F est estimée à partir de s, ce qui induit les conditions suivantes : $d_G(y_j, F\cdot x_i) \simeq 0$ et $d_G(x_i, F^T\cdot y_j) \simeq 0$ pour tout $(i,j)\in s$.

De plus, nous cherchons à fixer automatiquement deux seuils δ_G et δ_D comme dans les équations (4.9) et (4.10), ce que permet justement la modélisation *A Contrario*. La probabilité est alors la suivante :

$$p(\mathscr{S}, F, \delta_G, \delta_D) := \Pr\big(\forall (i,j)\in\mathscr{S}, d_G(x_i, y_j, F) \le \delta_G \text{ et } d_D(D(x_i), D(y_j)) \le \delta_D \mid \mathscr{H}_0\big). \tag{4.11}$$

Dans notre modèle, un sous-ensemble s de $\mathscr{S}$ composé de 7 correspondances est utilisé pour estimer F [Zha98]. Cette méthode produit au plus trois solutions pour chaque échantillon. Remarquons qu'il est aussi possible d'utiliser la méthode linéaire dite des 8 points [LH81] en modifiant la combinatoire. Par conséquence :

Proposition 1.

$$p(\mathscr{S}, F, \delta_G, \delta_D) = f_D(\delta_D)^k f_G(\delta_G)^{k-7} \tag{4.12}$$

où k est le cardinal de $\mathscr{S}$.

Démonstration.

$$\begin{aligned} p(\mathscr{S}, F, \delta_G, \delta_D) &= \prod_{(i,j)\in\mathscr{S}\setminus s} \Pr\big(d_G(x_i, y_j, F) \le \delta_G\big) \cdot \prod_{(i,j)\in\mathscr{S}} \Pr\big(d_D(D(x_i), D(y_j)) \le \delta_D\big) && (4.13)\\ &= f_D(\delta_D)^k f_G(\delta_G)^{k-7} && (4.14) \end{aligned}$$

L'équation (4.13) dérive du point 1 de la définition 10 et de l'équation (4.14) des points 2 et 3. □

Dans le cadre du test d'hypothèse classique, on rejette l'hypothèse nulle $\mathscr{H}_0$ dès que $p(\mathscr{S}, F, \delta_G, \delta_D)$ est inférieur à un seuil de confiance fixé à l'avance. Cependant, cela signifie alors, sans autre hypothèse, que les grands groupes $\mathscr{S}$ seront favorisés. Dans le cadre de l'approche *A Contrario*, on ne manipule pas directement les probabilités mais plutôt l'espérance de l'évènement associé, i.e. le *Nombre de Fausses Alarmes*. Il correspond au nombre moyen de groupes compatibles avec F, δ_G, δ_D sous l'hypothèse $\mathscr{H}_0$.

Définition 11. *Un ensemble $\mathscr{S}$ d'appariements est dit ε-significatif s'il existe*

1. *deux seuils δ_G et δ_D tels que :*

$$\forall (i,j) \in \mathscr{S},\ d_G(x_i, y_j, F) \leq \delta_G, \tag{4.15}$$

$$\forall (i,j) \in \mathscr{S},\ d_D(D(x_i), D(y_j)) \leq \delta_D, \tag{4.16}$$

2. *une matrice fondamentale F calculée à partir de 7 points de $\mathscr{S}$, telle que :*

$$NFA(\mathscr{S}, F, \delta_G, \delta_D) := 3\left(\min\{N_1, N_2\} - 7\right) k! \binom{N_1}{k}\binom{N_2}{k}\binom{k}{7} f_D(\delta_D)^k f_G(\delta_G)^{k-7} \leq \varepsilon \tag{4.17}$$

où k est le cardinal de $\mathscr{S}$.

Il est possible de démontrer [CLM+08, DMM08] que le nombre moyen d'ensembles ε-significatifs est, sous l'hypothèse $\mathscr{H}_0$, majoré par ε. Cela justifie l'expression "Nombre de Fausses Alarmes".

Une estimation du NFA peut s'établir de la manière suivante : il y a $\min\{N_1, N_2\} - 7$ choix pour $k \geq 7$, $\binom{N_1}{k}$ choix pour les points d'intérêt dans l'image 1, $\binom{N_2}{k}$ choix pour les points d'intérêt dans l'image 2, $k!$ choix pour les appariements, $\binom{k}{7}$ choix pour l'échantillon minimal servant à estimer F, et chaque échantillon produit jusqu'à trois matrices fondamentales [Zha98].

La définition 11 a été évoquée dans [MS04b] (*colored rigidity*) mais n'a ni été approfondie plus en avant ni implantée. Il est en effet indiqué qu'il est nécessaire de résoudre deux difficultés :

- disposer d'une mesure de dissimilarité indiquant la proximité visuelle entre deux points,
- diminuer le coût combinatoire, celui-ci étant suffisamment grand pour rendre même une recherche stochastique inopérante.

Comme f_D et f_G sont croissantes, on déduit le corollaire suivant de la définition précédente.

Proposition 2. *Un ensemble $\mathscr{S}$ d'appariements est ε-significatif s'il existe une matrice fondamentale F estimée à partir de 7 appariements parmi $\mathscr{S}$ tel que :*

$$NFA(\mathscr{S}, F) := 3\left(\min\{N_1, N_2\} - 7\right) k! \binom{N_1}{k}\binom{N_2}{k}\binom{k}{7} f_D(\delta_D)^k f_G(\delta_G)^{k-7} \leq \varepsilon \tag{4.18}$$

avec $\delta_G = \max_{(i,j)\in\mathscr{S}} \max\{d_G(y_j, F\cdot x_i), d_G(x_i, F^T\cdot y_j)\}$, $\delta_D = \max(d_D(D(x_i), D(y_j))$, où k est le cardinal de $\mathscr{S}$.

L'objectif de l'algorithme discuté dans la section 4.3 est de trouver l'ensemble d'appariements le plus significatif, c'est-à-dire l'ensemble d'appariements $\mathscr{S}$ avec le plus faible NFA($\mathscr{S}$). Trouver un minimum pour le nombre de fausses alarmes, dans notre cas, n'est vérifié que sur l'ensemble des cas testés, qui dépend des heuristiques de simplification combinatoire choisies. L'équation (4.18) équilibre l'influence entre la probabilité $f_D(\delta_D)^k f_G(\delta_G)^{k-7}$ et le nombre d'ensembles possibles de taille k. Si δ_D et δ_G sont fixés, lorsque k grandit, le premier diminue tandis que le second a tendance à augmenter.

4.2.2 Modélisation de la contrainte géométrique

Moisan et Stival [MS04b] proposent de définir $d_G(y, F\cdot x)$ comme la distance euclidienne entre y et la droite épipolaire $F\cdot x$. La fonction f_G est alors définie (de façon un peu abusive, voir chapitre 1, section 1.3.3) comme :

$$f_G(d_{\text{euc}}(y, F\cdot x)) = \frac{2D}{A} d_{\text{euc}}(y, F\cdot x) \tag{4.19}$$

avec D et A respectivement le diamètre et l'aire des deux images (qu'on suppose ici avoir les même dimensions). Ce choix vient du modèle *A Contrario* que l'on a présenté au chapitre 1, qui est plus spécifique que celui de la section précédente. Moisan et Stival supposent non seulement l'indépendance, mais aussi que les points d'intérêt sont distribués uniformément dans les images. Ceci les conduit à estimer la probabilité qu'un point choisi aléatoirement (tiré à partir d'une distribution uniforme) soit situé à une distance inférieure à δ de la droite épipolaire correspondante. La figure 1.4 présentant ce raisonnement géométrique a été présentée au chapitre 1.

Avec l'équation (4.10), on déduit alors :

$$\begin{aligned} \Pr(d_G(x,y,F) \leq \delta_G) &= \Pr(\max\{d_G(y, F\cdot x), d_G(x, F^T\cdot y)\} \leq \delta_G) && (4.20)\\ &= \left(\frac{2D}{A}\delta_G\right)^2 && (4.21)\end{aligned}$$

en supposant que les distances euclidienne entre y et $F\cdot x$ et entre x et $F^T\cdot y$ sont indépendantes, telle qu'habituellement pour l'approche *A Contrario*.

Pour une raison que nous expliciterons au chapitre 5, nous utilisons :

$$f_G(\delta_G) = \left(\frac{2D}{A}\delta_G\right)^{2\alpha} \tag{4.22}$$

avec $\alpha = 5$ fixé une fois pour toute.

Remarquons que $\frac{2D}{A}\delta_G$ peut être plus grand que 1 : c'est en effet une majoration de la fonction de répartition. Afin d'accélérer la recherche à la manière des tests préemptifs de Chum [CM02] et Nistér [Nis05], nous décidons d'éliminer a priori les groupes tels que cette probabilité est plus grande que 5%. Pour des images de dimensions 500×500, cela correspond à $\delta_G > 12.5$ pixels.

4.2.3 Modélisation de la contrainte photométrique

Nous définissons ici d_D et f_D, c'est-à-dire la distance entre descripteurs photométriques locaux et leurs fonctions de répartition associées.

L'espace des descripteurs n'est ni isotrope ni homogène, comme indiqué au chapitre 2. Il n'est alors pas conseillé d'utiliser une simple distance euclidienne pour mesurer la ressemblance entre descripteurs. On a habituellement plutôt recours à une étape de mise en correspondance au plus proche voisin.

Entre la distance euclidienne et la mise en correspondance au plus proche voisin, il nous apparaît judicieux de prendre en compte une certaine proximité de $D(x)$ telle que la valeur $d_D(D(x),D(y))$ ait la même signification en ce qui concerne la ressemblance de chaque descripteur $D(x)$.

Rabin et al. [RDG09a] exploitent cette approche en définissant un modèle *A Contrario* dédié à la mise en correspondance de descripteurs de type SIFT. Leur approche permet de fixer automatiquement des seuils de distance adaptatifs aux descripteurs. Au contraire du modèle *A Contrario* présenté dans ce chapitre, ils ne prennent pas en compte les contraintes géométriques.

En suivant [RDG09a], et à partir des travaux antérieurs [MSC+06, NSB07b, NSB10b] nous définissons :

$$d_D(D(x),D(y)) = \phi_{D(x)}(\mathsf{dist}(D(x),D(y))) \tag{4.23}$$

où dist est une distance quelconque (ou mesure de dissimilarité) sur l'espace des descripteurs (la distance euclidienne ou une plus sophistiquée comme on le verra par la suite), $\phi_{D(x)}$ est la fonction de répartition de $\mathsf{dist}(D(x),D(\cdot))$ lorsque $D(\cdot)$ parcourt l'ensemble des descripteurs de l'image $\mathscr{I}_2$.

Remarquons que, pourvu que $\phi_{D(x)}$ soit connue complètement et que $\mathsf{dist}(D(x),D(y))$ soit une réalisation du processus aléatoire sous-jacent, alors $d_D(D(x),D(y))$ est distribué uniformément sur

l'intervalle unitaire $[0,1]$ (voir la proposition 3, qui suit). Cette distance s'adapte ainsi automatiquement à l'hétérogénéité de l'espace des descripteurs comme une "mesure de dissimilarité contextuelle". La distance croît à mesure que le nombre de descripteurs perceptuellement proches grandit.

Proposition 3. *Supposons que l'espace des descripteurs* SIFT *est muni d'une métrique* dist*. Soient un descripteur D et un autre descripteur choisi aléatoirement D' tel que* $\mathsf{dist}(D,D')$ *soit une variable aléatoire de fonction de distribution cumulée f_D (supposée continue et strictement croissante). Nous définissons alors une nouvelle métrique $d(D,D') := f_D(\mathsf{dist}(D,D'))$. Alors $d(D,D')$ est distribuée uniformément sur l'intervalle unité* $[0,1]$.

Démonstration. On a en effet pour tout $t \in [0,1]$:

$$\Pr(d(D,D') \leqslant t) = \Pr(f_D(\mathsf{dist}(D,D') \leqslant t) = t \tag{4.24}$$

d'après la proposition 4. □

Proposition 4. *Si X est une variable aléatoire réelle et F sa fonction de répartition, alors pour tout nombre réel positif x :*

$$\Pr(F(X) \leq x) \geq x \tag{4.25}$$

et l'égalité est vraie si F est continue et croissante.

Démonstration. Soit $F^{-1}(x) = \arg\inf_{t\in\mathbb{R}}\{F(t) \geq x\}$, définie car F est croissante. On en déduit que $F(t) \leq x$ si et seulement si $t \leq F^{-1}(x)$.

Par conséquence :

$$\Pr(F(X) \leq x) = \Pr(X \leq F^{-1}(x)) = F(F^{-1}(x)) \geq x \tag{4.26}$$

et l'égalité est vraie si F est continue et croissante, car dans ce cas, F^{-1} est l'inverse de F. □

$\phi_{D(x)}$ n'est pas connue, mais peut être estimée empiriquement. Les descripteurs SIFT sont de grande dimension (usuellement 128), composés de la concaténation de 16 histogrammes de dimension $N = 8$. Rabin et al. [RDG09a] exploitent cette caractéristique pour réduire la dimensionalité en explorant deux définitions possibles de la distance entre descripteurs, lorsque l'on dispose d'une distance appropriée $\widetilde{\mathsf{dist}}$ entre histogrammes :

$$\mathsf{dist}(D(x),D(y)) = \sum_{i=1}^{N} \widetilde{\mathsf{dist}}(D^i(x),D^i(y)) \tag{4.27}$$

ou :

$$\mathsf{dist}(D(x),D(y)) = \max_{i=1\ldots N} \widetilde{\mathsf{dist}}(D^i(x),D^i(y)) \tag{4.28}$$

avec $(D^1(x),D^2(x),\ldots,D^N(x))$ l'ensemble des N histogrammes de $D(x)$.

Nous appelons dist-SUM (resp. dist-MAX) la mesure de dissimilarité de l'équation (4.27) (resp. (4.28)).

Notons pour tout $i \in [1,N]$, $\phi_{D^i(x)}$ la fonction de distribution de $\widetilde{\mathsf{dist}}(D^i(x),D^i(y))$, lorsque $D^i(y)$ parcourt l'ensemble des descripteurs de l'image $\mathscr{I}_2$. Suivant les principes de la méthodologie *A Contrario*, et sous conditions d'indépendance, nous définissons $\phi_{D(x)}$ tel que :

$$\phi_{D(x)}(\delta) = \int_0^{\delta} \bigotimes_{i=1}^{N} \phi_{D^i(x)}(t)dt \tag{4.29}$$

pour le cas dist-SUM (où $\otimes$ est le produit de convolution), et tel que :

$$\phi_{D(x)}(\delta) = \prod_{i=1}^{N} \int_0^{\delta} \phi_{D^i(x)}(t)dt \tag{4.30}$$

pour le cas dist-MAX.

En effet, dans le premier cas, $\mathsf{dist}(D(x), D(y))$ est définie comme la somme de N variables aléatoires (dont la distribution de probabilités est le produit de convolution des N distributions marginales sous hypothèse d'indépendance), tandis que dans le second cas, la distance est définie comme le maximum des variables aléatoires (dont la fonction de répartition est le produit des N fonctions cumulées marginales sous hypothèse d'indépendance).

En pratique, la fonction de répartition $\phi_{D^i(x)}$ est estimée empiriquement sur l'ensemble de tous les $D^i(y)$ avec y à valeurs dans l'ensemble des points d'intérêt extraits de l'image $\mathscr{I}_2$.

D'après l'équation (4.23), $d_D(D(x), D(y)) = \phi_{D(x)}(\mathsf{dist}(D(x), D(y)))$. Afin de satisfaire les conditions de la section 4.2.1, il est encore nécessaire de définir la fonction de répartition f_D. Comme

$$f_D(t) = \Pr(\phi_{D(x)}(\mathsf{dist}(D(x), D(y))) \leq t) = t \tag{4.31}$$

si $\phi_{D(x)}$ est continue et croissante (une propriété habituelle des fonctions de distribution cumulées, rappelée dans la proposition 3 qui précède. Nous fixons ici simplement $f_D(t) = t$.

4.2.4 Les distances photométriques utilisées

4.2.4.1 Nomenclature des distances

Dans les sections suivantes, nous utilisons les abréviations suivantes pour les choix de distance d_D (et f_D) vues au chapitre 2, section 2.3.1 :

- EUC-MAX : distance Euclidienne avec la loi du max (produit simple)
- EUC-SUM : distance Euclidienne avec la loi de la somme (produit de convolution)
- MAN-MAX : distance de Manhattan avec la loi du max (produit simple)
- MAN-SUM : distance de Manhattan avec la loi de la somme (produit de convolution)
- CHI2-MAX : distance du χ^2 avec la loi du max (produit simple)
- CHI2-SUM : distance du χ^2 avec la loi de la somme (produit de convolution)
- CEMD-SUM : distance du terrassier circulaire avec la loi du max (produit simple)
- CEMD-MAX : distance du terrassier circulaire avec la loi de la somme (produit de convolution)

4.2.4.2 Métrique sur les ensembles mis en correspondance

Pour deux ensembles de descripteurs provenant de deux images, on recherche l'ensemble d'appariements le plus significatif. Le but est de trouver un ensemble $\mathscr{S}$ et une matrice fondamentale F estimée à partir d'un échantillon de 7 points tel que le Nombre de Fausses Alarmes (NFA) soit le plus faible parmi tous les ensembles possibles. En suivant la proposition 2, le NFA est défini tel que :

$$\mathrm{NFA}(\mathscr{S}, F) := 3\left(\min\{N_1, N_2\} - 7\right) k! \binom{N_1}{k} \binom{N_2}{k} \binom{k}{7} f_D(\delta_D)^k f_G(\delta_G)^{k-7} \tag{4.32}$$

avec

$$\delta_G = \max_{(i,j)\in\mathscr{S}} \max\{d_G(y_j, F \cdot x_i), d_G(x_i, F^T \cdot y_j)\} \tag{4.33}$$

et

$$\delta_D = \max_{(i,j)\in\mathscr{S}} (d_D(D(x_i), D(y_j)). \tag{4.34}$$

où f_G et d_G sont fixée en section 4.2.2, et f_D et d_D sont définis comme un des choix (EUC, MAN, CHI2, CEMD / MAX ou SUM) vus précédemment.

Par exemple, le NFA associé avec CEMD-MAX est :

$$\begin{aligned}\mathrm{NFA}(\mathscr{S},F) := 3\left(\min\{N_1,N_2\}-7\right)k!\binom{N_1}{k}\binom{N_2}{k}\binom{k}{7}\cdot \\ \left(\max_{(i,j)\in\mathscr{S}}\prod_{k=1}^{N}\int_0^{d_{\text{CEMD-MAX}}(D(x_i),D(y_j))}\phi_{D^k(x_i)}(t)dt\right)^k\cdot \\ \left(\frac{2D}{A}\max_{(i,j)\in\mathscr{S}}\max\{d_{\text{euc}}(y_j,F\cdot x_i),d_{\text{euc}}(x_i,F^T\cdot y_j)\}\right)^{2\alpha(k-7)}\end{aligned} \tag{4.35}$$

4.3 Choix techniques et algorithmes

4.3.1 Discussion : le NFA, un critère pour estimer le consensus

Les ensembles avec un petit NFA sont les plus pertinents, en particulier lorsque le NFA est inférieur à 1. Dans cette section, nous montrons que rechercher un ensemble ε-significatif est réalisable en pratique, étant donnée la complexité du problème et les probabilités dont nous disposons. Cette discussion complète celle sur la *colored rigidity* dans [MS04b]. Afin de simplifier la présentation, nous supposons ici que $N_1 = N_2 = N$.

Nous notons

$$M(k,N) := 3(N-7)k!\binom{N}{k}^2\binom{k}{7}. \tag{4.36}$$

La figure 4.1 montre la courbe de $-\log_{10}(M(k,N))/k$ en fonction de k pour quelques valeurs typiques de N. A partir de l'équation (4.18), on en déduit la valeur maximale du logarithme de $f_D(\delta_D)f_G(\delta_G)^{1-7/k} \simeq f_D(\delta_D)f_G(\delta_G)$ telle que le groupe correspondant $\mathscr{S}$ soit 1-significatif (pour le cas $N_1 = N_2 = N$). On a en effet :

$$\mathrm{NFA}(\mathscr{S}) \le 1 \quad \text{ssi} \quad \log_{10}\left(f_D(\delta_D)f_G(\delta_G)^{1-7/k}\right) \le -\log_{10}(M(k,N))/k. \tag{4.37}$$

Utiliser le NFA comme critère remplit les deux conditions que l'on attendait :

- Pour une valeur de N fixée, plus k est petit, plus faible est le produit des probabilités. Cette situation peut être rencontrée lorsque le taux d'outliers est élevé et lorsque l'on recherche un groupe significatif avec un k petit devant N. Comme f_D et f_G sont croissantes, cela signifie que les seuils δ_D et δ_G doivent être plus stricts dans ce cas.
- Quand k/N est fixé, plus N est grand, plus faible est le produit des probabilités (et alors plus les seuils sont stricts). Cette propriété est intéressante lorsque l'on recherche des appariements dans des images de taille fixée : plus les appariements candidats sont nombreux, plus ils doivent être précis pour les critères géométriques et photométriques.

4.3.2 Accélération de la recherche d'ensembles significatifs

Quand on cherche le groupe le plus significatif d'appariements (que ce soit pour la matrice fondamentale ou l'homographie), une approche naïve consiste à tester tous les ensembles possibles d'appariements. Cependant, si $N = 100$ points d'intérêt sont extraits dans chaque image, il y a

$$\sum_{k=0}^{N} k!\binom{N}{k}^2 \simeq 10^{164} \tag{4.38}$$

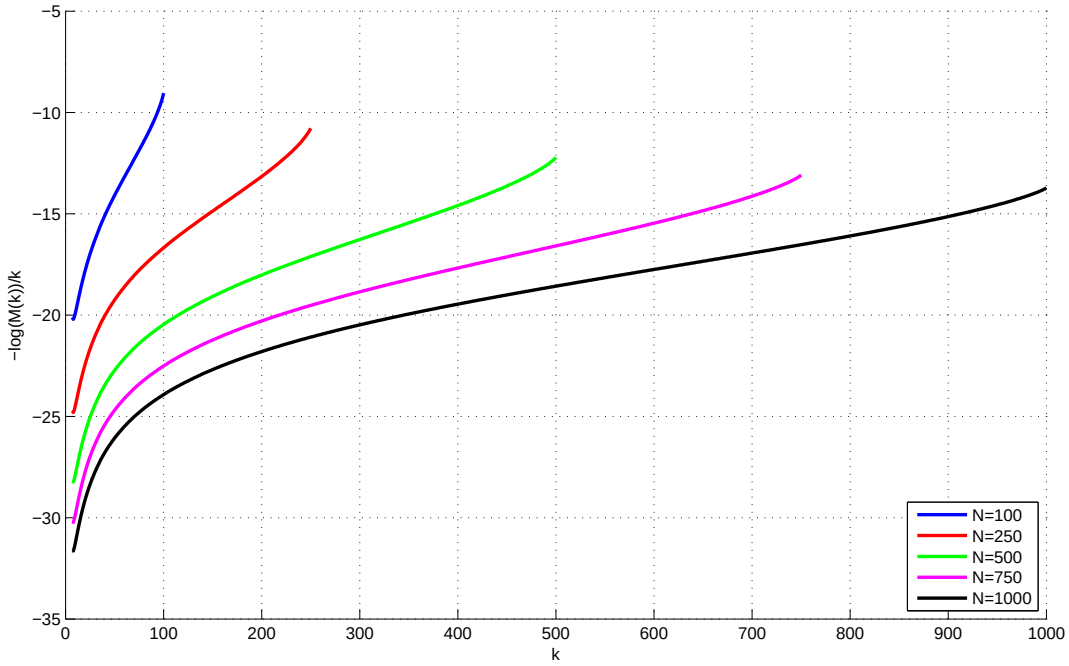

FIGURE 4.1 – $-\log_{10}(M(k,N))/k$ en fonction de k, pour différentes valeurs de N. Ceci montre l'ordre de grandeur du logarithme de $f_D(\delta_D)\cdot f_G(\delta_G)$, tel qu'il soit possible de trouver un ensemble d'appariements 1-significatif.

ensembles de ce type (comme indiqué dans [AT02]). Puisque tester toutes les possibilités est hors de question, une recherche guidée par une heuristique est utilisée. D'abord (section 4.3.2.1), nous utilisons un seuil (assez tolérant) sur la contrainte photométrique afin de réduire l'ensemble des candidats à l'appariement pour un point d'intérêt donné de l'image $\mathscr{I}_1$. Comme le nombre de correspondances possibles est encore important, nous utilisons une méthode d'échantillonnage statistique, i.e. une heuristique de type RANSAC (section 4.3.2.2).

4.3.2.1 Réduction de la combinatoire

Afin de réduire la complexité du problème, nous ne considérons pas tous les appariements possibles $y_1,\dots,y_{N_2}$ dans l'image $\mathscr{I}_2$ pour un point d'intérêt x_i de l'image $\mathscr{I}_1$, mais seulement un ensemble d'appariements candidats $y_{j_1},\dots,y_{j_{N_i}}$ tels que la distance entre les descripteurs associés soit inférieure à un seuil fixé. Bien sûr, ce seuil de mise en correspondance doit réduire la taille de l'ensemble des possibilités uniquement pour des raisons de complexité algorithmique. Il doit alors être suffisamment grand afin que la vrai décision de mise en correspondance ne soit pas réalisée à cette étape, tout en retirant des appariements qui sont non pertinents sans ambiguïté.

Afin d'éviter la fixation de seuils arbitraires, nous utilisons la méthode *A Contrario* fournie par [RDG09a]. Dans ce cas, y_j est candidat à la mise en correspondance avec x_i, si, avec les notations de l'équation (4.23) :

$$N_1 N_2 d_D(D(x_i),D(y_j)) \leq \widetilde{\varepsilon}. \tag{4.39}$$

La valeur de $\widetilde{\varepsilon}$ ne dépend pas du protocole expérimental et est discutée en détail dans [RDG09a] pour la distance CEMD-SUM. Remarquons que la proposition 3 (section 4.2.3) montre un argument à propos de l'auto-adaptabilité de cette quantité aux descripteurs considérés. Dans cette thèse, nous

fixons $\widetilde{\varepsilon} = 10^{-2}$, ce qui donne une quantité raisonnable de candidats à la mise en correspondance pour chaque x_i dans une image typique.

4.3.2.2 Algorithme d'échantillonnage statistique

A cette étape chaque point d'intérêt x_i de l'image $\mathscr{I}_1$ est mis en relation avec un ensemble de N_i correspondances potentielles $y_{j_1}, \ldots, y_{j_{N_i}}$ de l'image $\mathscr{I}_2$. Maintenant, l'objectif est de choisir un (ou zéro) $y_{j(i)}$ parmi cette liste. Comme le nombre de possibilités reste important, nous utilisons un algorithme d'échantillonnage statistique. Nous reprenons ainsi le schéma utilisé par les approches de type RANSAC, en corrigeant leurs principales faiblesses dues à l'utilisation de seuils déterminés à priori. Il s'agit d'un algorithme itératif en deux étapes, qu'on décrit dans le cas de la matrice fondamentale :

A tirage d'un échantillon de sept appariements pour estimer F,

B Recherche du groupe le plus significatif composé à partir d'un sous-groupe compatible avec F, composé des candidats à l'appariement précédents.

A. Tirage d'un échantillon de sept appariements. Sept points x_i sont tirés uniformément, et sont alors associés à un point correspondant candidat $y_{j(i)}$. En suivant les résultats expérimentaux qui montrent ses bonnes performances, nous utilisons la mise en correspondance par la méthode du plus proche voisin (au sens de la photométrie). Nous pourrions aussi ne pas avoir biaisé l'algorithme avec ce choix, et choisi à la place de prendre pour chaque i le point correspondant $y_{j(i)}$ en le tirant aléatoirement parmi l'ensemble $y_{j_1}, \ldots, y_{j_{N_i}}$ où y_{j_i} a le poids $K/d_D(D(x_i), D(y_{j_i}))$ (K est un paramètre de normalisation). Cette approche sélectionnerait alors plutôt les plus proches voisins mais permettrait aussi de retenir des appariements dont les descripteurs ne sont pas plus proches voisins. Cependant, le taux d'outliers est significativement plus élevé pour les non plus proches voisins (ce qui est vérifié expérimentalement) ; cette seconde approche nécessite alors plus d'itérations pour fournir un résultat non pollué.

La matrice fondamentale est alors estimée en utilisant l'algorithme des sept points.

Remarquons que le détecteur SIFT peut extraire plusieurs points d'intérêt au même endroit, mais avec des orientations différentes. Afin d'éviter de calculer F pour un échantillon contenant plusieurs fois le même point, nous vérifions que l'échantillon choisi ne contient pas de tels points. Nous avons vérifié expérimentalement que ces points multiples ne modifient pas les groupes trouvés pour modifier les décisions prises à l'issue du calcul du NFA pour les groupes dont nous décrivons la construction ci-dessous.

B. Recherche des groupes significatifs. Des appariements sont ajoutés aux sept précédents afin de construire un groupe le plus significatif que possible. Contrairement à l'approche *A Contrario* de Moisan et Stival, il n'est pas possible de construire un tel groupe de façon systématique, en les ordonnant par leur rigidité géométrique. Nous utilisons une heuristique, qui vise à guider la recherche de solution, et consiste en l'itération des étapes suivantes.

1. Pour chaque x_i, sélectionner :

$$y_{j(i)} = \underset{y_{j_k}}{\operatorname{argmin}} \left\{ f_D(d_D(D(x_i), D(y_{j_k}))) \cdot f_G(d_G(F, x_i, y_{j_k})) \right\} \tag{4.40}$$

et trier les appariements $(x_i, y_{j(i)})$ dans l'ordre croissant par rapport à cette dernière valeur, tel qu'on obtienne une série de groupes imbriqués constitués de $k = 7, 8, 9, \ldots, N_1$ appariements.

Ce choix est basé sur l'observation, que pour un k fixé, le groupe le plus significatif minimise le produit $f_D(\delta_D)f_G(\delta_G)$.
Cette étape peut produire des appariements entre N x_i et un seul y_j, ce qui ne devrait pas exister. Ainsi, nous décidons d'en garder un seul parmi ces appariements, à savoir le $(x_i, y_{j(i)})$ tel que le produit de probabilité mentionné plus haut soit minimisé.

2. Calculer le NFA pour chacun des groupes imbriqués mentionnés plus haut (en suivant l'équation (4.32), avec δ_G et δ_D donnés par les équations (4.33) et (4.34)), et en gardant le groupe le plus significatif.
3. Trier les appariements $(x_i, y_{j(i)})$ en ordre croissant par rapport à $f_G(\delta_G(F, x_i, y_{j(i)}))$ pour construire un nouvel ensemble de groupes imbriqués, calculer le NFA et sélectionner le groupe le plus significatif. Nous avons remarqué expérimentalement que cette étape permet d'éliminer des correspondances incorrectes qui étaient introduites pour un petit k à l'étape 1, parce que la distance photométrique est très bonne et prend l'avantage sur la (mauvaise) distance géométrique.
4. Renvoyer le groupe le plus significatif trouvé à l'issue des étapes 2 ou 3.

Les étapes 1 et 2 ne garantissent évidemment pas que le groupe obtenu est le plus significatif pour une matrice F fixée (au contraire de l'approche *A Contrario*-RANSAC de[MS04b] où seul le critère géométrique intervient).

Remarquons que l'étape 1 permet de sélectionner des appariements dont les descripteurs ne sont pas plus proches voisins. Utiliser des heuristiques successives pour tester les ensembles d'appariements est fondé, puisqu'on recherche le plus petit NFA, quelle que soit la façon dont le groupe est construit.

D'autres stratégies d'échantillonnages dans un contexte similaire sont décrites dans [ZK06].

Remarquons que lorsque l'on recherche une homographie, calculer H nécessite 4 appariements au lieu de 7 pour F. Ce point mis à part, l'algorithme est exactement le même.

4.3.3 Algorithme

Pour résumer la discussion, l'algorithme complet est présenté ici. Nous supposons que deux vues de la même scène 3D sont données.

1. Utiliser l'algorithme SIFT pour extraire des points d'intérêt et des descripteurs invariants (zoom+rotation / changements de contraste) pour chaque vue : $(x_i, D(x_i))_{i\in\{1,\dots N_1\}}$ and $(y_j, D(y_j))_{j\in\{1,\dots N_2\}}$.
2. Pour chaque $i \in \{1, \dots N_1\}$,
   (a) construire la distance empirique d_D (section 4.2.3, équation (4.23)),
   (b) définir un ensemble de candidats à la mise en correspondance (section 4.3.2.1).
3. Itérer :
   (a) choisir sept points x_i et sélectionner les sept points correspondants $y_{j(i)}$ (heuristique A de 4.3.2.2),
   (b) calculer les trois matrices fondamentales F possibles à partir de ces sept appariements et aller au point (c) pour chacune de ces matrices,
   (c) sélectionner le groupe le plus significatif (heuristique B de 4.3.2.2).

Enfin, renvoyer le groupe le plus significatif parmi ceux évalués.

Le nombre d'itérations de la recherche par échantillonnage est fixé au chapitre 5 à une grande valeur (en fonction de la proportion estimée d'outliers) afin que le groupe renvoyé soit effectivement le groupe le plus significatif avec une confiance importante.

Remarquons que le modèle probabiliste proposé et l'algorithme associé ne sont pas spécifiques aux descripteurs SIFT et peuvent être adaptés facilement à tout autre descripteurs invariants à base d'histogrammes.

4.4 A propos des homographies

Le modèle *A Contrario* précédent traite de la mise en correspondance de points sous contrainte épipolaire. Cependant, dans certains cas dégénérés (i.e. lorsque la caméra tourne autour de son centre optique, ou lorsque les points 2D sont situés sur un même plan), les appariements sont reliés par une homographie.

Adapter la proposition 2 (section 4.2.1) nous amène à dire qu'un ensemble $\mathscr{S}$ d'appariements est ε-significatif dans ce cas s'il existe une homographie H estimée à partir de 4 appariements parmi $\mathscr{S}$ telle que :

$$\mathrm{NFA}(\mathscr{S},H) = \big(\min\{N_1,N_2\}-4\big)\,k!\binom{N_1}{k}\binom{N_2}{k}\binom{k}{4} f_D(\delta_D)^k f_G(\delta_G)^{k-4} \leq \varepsilon \tag{4.41}$$

en utilisant les mêmes notations que dans la proposition 2.

Il est alors uniquement nécessaire d'adapter la définition de $f_G(\delta_G)$ d'un appariement point à droite (équation (4.22)) vers un appariement point à point. On a ainsi :

$$f_G(\delta_G) = \left(\frac{\pi\delta_G^2}{A}\right)^{2\alpha}. \tag{4.42}$$

En effet, $\pi\delta_G^2/A$ est la probabilité qu'un point positionné aléatoirement et uniformément dans une image d'aire A soit à une distance inférieure à δ_G d'un point fixé. Dans [RDGM10], un modèle proche est utilisé pour réaliser la détection en cascade de transformations : cela permet de mettre en correspondance plusieurs plans différents qui auraient tendance à être considérés comme un seul et unique plan autrement.

Comme dans le cas de la matrice fondamentale, $\alpha = 5$ est fixé une fois pour toute.

4.5 Conclusion

Dans ce chapitre, nous avons présenté un modèle statistique *A Contrario* afin de déterminer des appariements entre points d'intérêt détectés dans deux images. Alors que la plupart des méthodes de la littérature traitent la consistance photométrique et les contraintes géométriques séparément, la méthode proposée les intègre en une métrique simple, le Nombre de Fausses Alarmes.

Remarquons que la mise en correspondance de points d'intérêt est limitée intrinsèquement par l'invariance incomplète des points SIFT. D'autres points d'intérêt, munis de descripteurs (pseudo) affine invariants, comme [MS04a, MTS+06], pourraient aussi être utilisés. Les méthodes de simulation de points pour générer des représentations invariantes, comme dans Lepetit et al [LF06] ou A-SIFT [MY09b], nous paraissent intéressantes pour la mise en correspondance avec de forts changements de point de vue, puisqu'elles permettent une meilleure invariance en générant des transformations affines de l'image. C'est d'ailleurs le sujet du chapitre 6, pour lequel nous présentons une méthode *A Contrario* reposant sur A-SIFT permettant la mise en correspondance en présence de forts changements de points de vue et de motifs répétés.

Chapitre 5

Résultats expérimentaux pour le modèle AC de mise en correspondance

Ce chapitre présente des expériences visant à évaluer la pertinence de la méthode proposée en section 4.3.3 ainsi qu'étudier l'influence des choix des paramètres qu'elle utilise. Nous testons la recherche d'appariements sous contrainte épipolaire avec la matrice fondamentale ou avec l'homographie. L'organisation du chapitre est la suivante : nous commençons par décrire les séquences utilisées ainsi que les critères d'évaluation choisis en section 5.1. Nous présentons l'influence du paramètre α en section 5.2.1, puis celle de la distance entre les images en section 5.2.2. Nous étudions ensuite le comportement des différentes distances entre descripteurs en section 5.3. Puis, nous montrons des résultats de mise en correspondance en présence d'aliasing perceptuel en section 5.4. Enfin, nous concluons en section 5.5.

5.1 Approche expérimentale

Description des séquences utilisées

Les séquences utilisées décrivent une grande variété de situations pour la recherche d'appariements de points d'intérêt. Dans ces expériences, nous mettons l'accent sur les possibilités de l'algorithme à gérer des motifs répétés, mais nous montrons aussi son efficacité dans le cadre de séquences plus génériques. Tout d'abord, nous présentons des résultats sur une séquence synthétique appelée *cube*, qui consiste en un survol (translation et rotation) d'un cube texturé d'une alternance de triangles clairs et foncés : ce motif répété fourni de nombreux points d'intérêt à l'aide de la méthode SIFT. Une autre séquence, *Pegase*, présente cette fois-ci une scène composée de quatre plans texturés de motifs répétés (d'après Escher), prise à l'aide d'un appareil photo numérique via un mouvement libre de translation et de légère rotation. Une série de paires d'images, *Sears*, *Flat Iron*, *Bâtiment du Loria* sont typiques pour leur présence de motifs répétés en environnement urbain, ou en intérieur avec *Corridor*. En dehors de ces séquences où la grande majorité des zones texturées sont composées de motifs répétés, nous avons aussi évalué notre méthode sur une séquence générique, *Medusa* de Pollefeys.

Critères d'évaluation

Pour évaluer notre approche, nous utilisons différents critères d'évaluation :

– Sur les images, nous affichons les appariements retenus par des croix bleues claires, suivi d'un segment indiquant le mouvement apparent à suivre pour trouver le point d'intérêt correspondant sur l'autre image. Cela permet une validation visuelle.

– Nous présentons aussi des tracés de droites épipolaires : nous visualisons ainsi la contrainte géométrique point - droite trouvée. Des couples de points de contrôle sont entrés à la main. Ils doivent se situer sur les droites épipolaires correspondantes dans l'autre vue, ce qui permet une autre vérification de cohérence visuelle.
– Afin d'établir la pertinence de notre approche pour la sélection d'appariements rejetés par la méthode NNT (voir définition 9, section 2.3), nous présentons des histogrammes de ces rapports pour l'ensemble des appariements candidats, ainsi qu'ensuite pour ceux retenus.
– Enfin, des tableaux fournissent le nombre d'appariements retenus ainsi que leur rang de ressemblance photométrique, pour montrer qu'il est important de ne pas uniquement conserver les plus ressemblants. Les valeurs indiquées de la contrainte géométrique δ_G, de la contrainte photométrique δ_D (en log), et de la significativité retenue $\widetilde{\varepsilon}$ montrent qu'elles s'adaptent automatiquement en fonction de la précision des points d'intérêt extraits et des paramètres du modèle de mouvement qu'il est possible de calculer. Les tableaux présentent des valeurs moyennées ainsi que les écarts types entre parenthèses obtenus en répétant 100 fois chaque expérience.

Nomenclature des différentes méthodes

Les différentes méthodes comparées sont les suivantes :

– la mise en correspondance SIFT classique (i.e. plus proche voisin + seuil sur le rapport de distances, entre 0.6 et 0.8 afin de produire les meilleurs résultats possibles), suivi d'un RANSAC d'après [MS04b]. Nous appelons cette approche NNT+F si RANSAC impose la géométrie épipolaire (matrice fondamentale), ou NNT+H pour la contrainte homographique.
– la méthode *A Contrario* du chapitre précédent, qui permet de résoudre l'ambiguïté perceptuelle, est appelée ACM+F ou ACM+H.

Les images présentées correspondent à des cas typiques parmi ces expériences. Pour chaque expérience, le nombre d'échantillons évalués est de 20000, ce qui permet d'obtenir une sélection robuste pour des taux d'outliers élevés (jusqu'à 85%) tout en conservant un temps de calcul raisonnable (de 5 à 10 secondes par expérience).

Pour les estimation d'appariements utilisant comme critère géométrique la matrice fondamentale, suivant les indications de [HZ00] une étape d'optimisation non linéaire est réalisée pour calculer une matrice fondamentale plus précise sur l'ensemble d'appariements filtrés. Nous utilisons pour cela l'algorithme de Powell avec la distance de Sampson. Les droites épipolaires présentées sont celles associées à cette nouvelle matrice fondamentale recalculée.

5.2 Influence des paramètres sur l'algorithme

Nous testons ici l'influence du paramètre α. Pour étudier le comportement de notre algorithme et l'influence de α, nous réalisons une série d'expériences pour des images synthétiques (séquence *cube*), ou réelles (séquences *pegase* et *medusa*), en recherchant des ensembles de correspondances sous contrainte épipolaire ou homographique. L'objectif de cette série d'expériences est de montrer que notre approche permet de trouver des ensembles d'appariements sélectionnés de façon robuste, dont les seuils de distance au modèle géométrique recherché, et de ressemblance photométrique varient automatiquement.

5.2.1 Influence de la valeur de α

Rappelons l'équation (4.22) faisant intervenir α :

$$f_G(\delta_G) = \left(\frac{2D}{A}\delta_G\right)^{2\alpha} \tag{5.1}$$

Nous considérons une paire d'images contenant de nombreux motifs répétés, extraite de la séquence *cube*, visible dans les figures 5.1 et 5.4. Cette séquence composée d'un cube texturé par un motif répété défini par une alternance de triangles clairs et foncés produit de nombreux points d'intérêt aux descripteurs photométriques proches. Il est donc difficile de mettre correctement ces points en correspondance sans la prise en compte de la contrainte géométrique. L'ambiguïté de cette scène est illustrée par l'histogramme des rapports de distance des descripteurs photométriques entre plus proche voisin et second plus proche voisin en figure 5.3, qui sont comparés à un seuil fixé usuellement à 0.6 pour décider de retenir ou non les candidats à l'appariement. En particulier, l'histogramme de droite des rapports pour les seuls appariements retenus montre que les appariements ambigus (avec un rapport proche de 1) sont nombreux.

Le tableau 5.1 nous montre que α permet de faire varier l'influence des contraintes géométriques et photométriques sur l'ensemble d'appariements sélectionnés. Pour de faibles valeurs de α, c'est la contrainte photométrique qui domine. Pour la valeur $\alpha = 1$, la précision obtenue pour le maximum des distances d'un point à la droite épipolaire correspondante – il s'agit du seuil géométrique fixé automatiquement par la méthode – est de 0.7 pixels pour -40 en logarithme pour la ressemblance photométrique (soit le maximum des distances entre descripteurs photométriques). Pour α entre 7 et 10, on obtient une meilleure précision géométrique (0.5 pixel) mais avec une ressemblance photométrique moindre : dans ce cas, la partie photométrie est "écrasée", son rôle est négligeable pour la construction de l'ensemble résultat. Elle a alors uniquement servi à fournir une hypothèse multiple pour l'appariement. Si l'on veut obtenir un filtrage robuste, c'est-à-dire un ensemble d'appariements résultat ne contenant plus d'outliers, on va chercher à privilégier la contrainte géométrique. La ressemblance photométrique ne fournit pas uniquement les hypothèses multiples que l'on cherche à filtrer, mais elle permet alors de pondérer le critère géométrique pour le choix des appariements en présence d'ambiguïté. La figure 5.1 montre des résultats pour la matrice fondamentale. Dans ce cas, pour une valeur de α inférieure à 3, on favorise les appariements entre zones les plus ressemblantes : le mouvement retrouvé est alors cohérent avec une contrainte géométrique pour laquelle le correspondant d'un point est situé sur une ligne de fuite correspondant à un alignement de motifs répétés, ce qui n'est pas le mouvement réel des caméras.

Dans le tableau 5.1, les valeurs de α entre 3 et 6 permettent d'équilibrer l'influence des deux contraintes : la géométrie est suffisamment précise pour assurer une sélection robuste, avec une photométrie aidant à résoudre les ambiguïtés.

La figure 5.2 montre le comportement obtenu pour l'homographie. On s'attend alors à trouver une homographie qui explique le mouvement global d'une face du cube. Là aussi, une petite valeur de α privilégie le choix de groupes d'appariements avec une ressemblance photométrique marquée : on obtient alors une homographie cohérente, mais pour un mouvement qui ne suit pas le mouvement de la caméra. Dans toutes les utilisations suivantes, nous retenons la valeur $\alpha = 5$ comme compromis raisonnable.

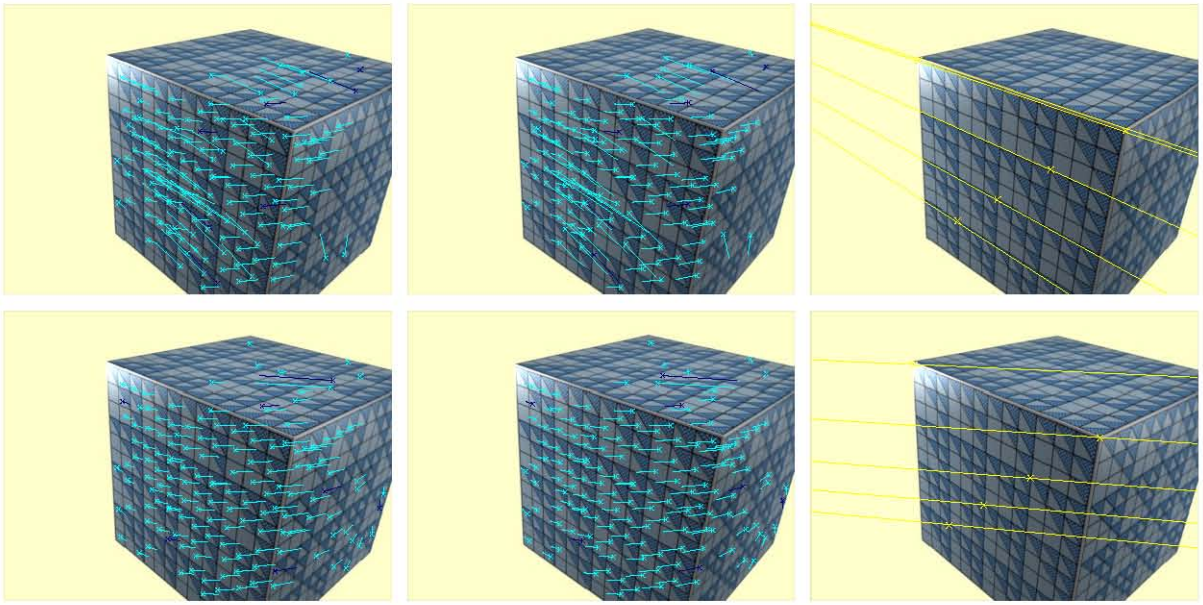

FIGURE 5.1 – *cube* - Influence de α pour la matrice fondamentale. En haut, $\alpha = 3$. Les 138 appariements trouvés (dont 31 ne sont pas plus proches voisins) sont ceux pour lesquels la ressemblance photométrique est la plus grande. Ils correspondent en effet au mouvement apparent le plus faible. Le mouvement retrouvé suit les lignes de fuite, comme indiqué par les droites épipolaires tracées à droite. En bas, $\alpha = 5$. Les 148 appariements trouvés (dont 47 ne sont pas plus proches voisins) suivent le mouvement global, ce que confirment les droites épipolaires.

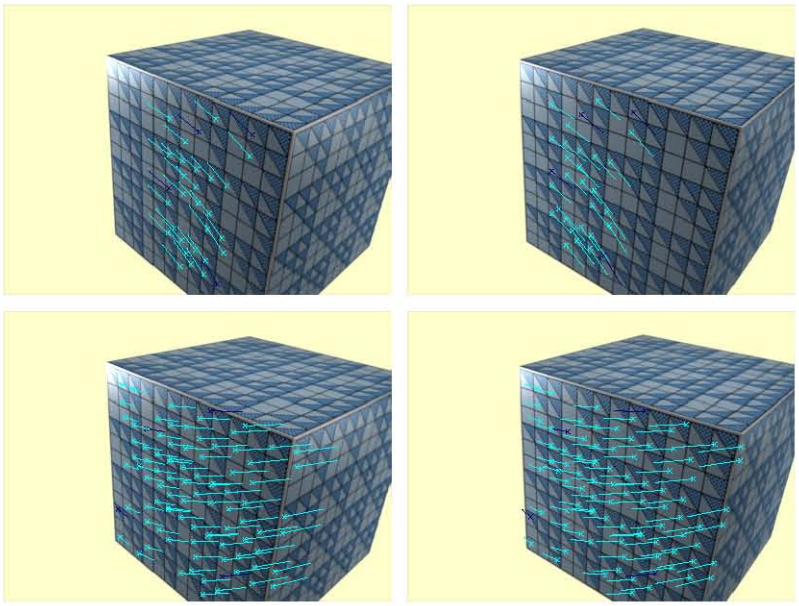

FIGURE 5.2 – Influence de α pour l'homographie. En haut, $\alpha = 1$. Le mouvement de la caméra trouvé fait que les 31 appariements sélectionnés (dont 4 ne sont pas plus proches voisins) sont ceux dont les distances photométriques sont les plus faibles (le changement d'apparence est plus faible pour le motif décalé apparié pour le cas du haut que pour le motif correspondant correct pour le cas du bas). Si la superposition des motifs est cohérente, elle ne suit pas le mouvement global. En bas, $\alpha = 3$. Les 91 appariements trouvés (dont 56 ne sont pas plus proches voisins) suivent le mouvement global.

α	Nb. points	δ_G	log δ_D
1	149.2 (1.8)	.71 (.12)	-40.4 (.5)
2	151.6 (3.8)	.70 (.14)	-39.5 (1.1)
3	162.1 (8.1)	.58 (.12)	-33.6 (3.2)
4	164.2 (5.2)	.49 (.13)	-30.5 (1.8)
5	172.3 (10)	.46 (.11)	-26 (4.2)
6	183.5 (6.2)	.49 (.09)	-21.2 (1.9)
7	182.9 (5.3)	.47 (.10)	-20.7 (.48)
8	183.7 (4.9)	.46 (.08)	-20.6 (.28)
9	184.5 (4.5)	.46 (.08)	-20.5 (.28)
10	185 (4.8)	.46 (.08)	-20.5 (.17)

TABLE 5.1 – *cube*, matrice fondamentale, α varie. Nombre d'appariements retenus en fonction de la valeur de α (Nb. points), distance maximale d'un point à sa ligne épipolaire correspondante (δ_G), probabilité maximale en log correspondant à la distance photométrique (log δ_D) Les écarts types sont indiqués entre parenthèses.

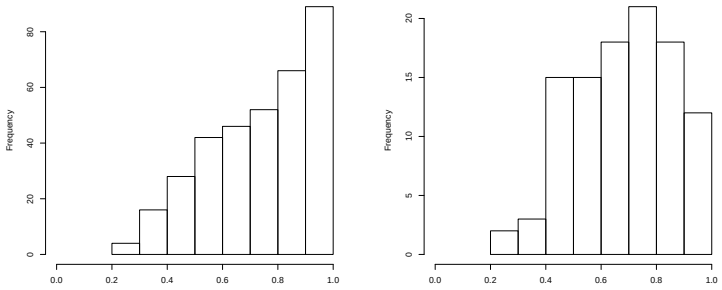

FIGURE 5.3 – *cube* - Histogramme des rapports de distance des descripteurs photométriques entre plus proche voisin et second plus proche voisin. A gauche, pour tous les points d'intérêt détectés. A droite, pour les points considérés inliers après filtrage robuste avec notre algorithme, restreints à ceux qui sont trouvés parmi ceux plus proches voisins. De nombreux points sont au dessus du seuil habituel de 0.6 et sont donc habituellement rejetés.

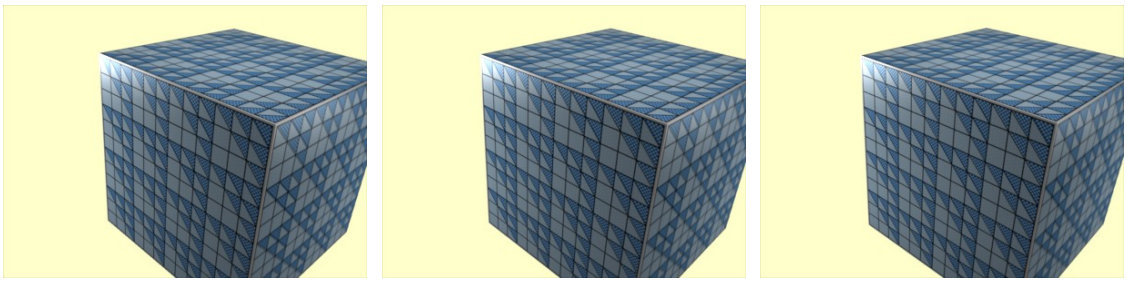

FIGURE 5.4 – *cube* - Variation de points de vue. Images 76, 79 et 82.

	Nb. points	δ_G	log δ_D	log $\widetilde{\varepsilon}$	
← Baseline croissante →	199.4 (4.6)	0.61 (0.11)	-24.1 (0.8)	-13908 (125)	Mouvement correct
	172.5 (10.7)	0.48 (0.13)	-25.9 (4.3)	-12620 (167)	
	169.0 (5.5)	0.59 (0.13)	-21.2 (1.4)	-11250 (182)	
	155.2 (4.0)	0.72 (0.15)	-19.8 (1.4)	-9757 (300)	
	151.9 (5.7)	0.79 (0.14)	-13.4 (1.3)	-8436 (277)	
	143.7 (4.5)	0.81 (0.18)	-15.6 (1.5)	-8251 (276)	
	130.0 (6.3)	0.84 (0.28)	-12.1 (1.1)	-6941 (282)	
	111.9 (12.8)	1.11 (0.62)	-13.0 (2.1)	-5773 (225)	
	124.9 (6.0)	1.62 (0.42)	-12.9 (0.9)	-5973 (155)	Lignes de fuite
	120.8 (7.3)	1.62 (0.46)	-12.1 (1.1)	-5681 (173)	

TABLE 5.2 – *cube*, matrice fondamentale. Nombre d'appariements considérés inliers détectés pour une distance entre caméras croissante, pour une scène présentant de nombreux motifs répétés. Les résultats sont des moyennes sur 100 tirages.

5.2.2 Influence de la distance entre caméras

Lorsque la distance entre caméras – la ligne de base, ou *baseline* – augmente, la partie de la scène qui est visible pour les deux caméras diminue. Cela a un impact sur le nombre de points d'intérêt qu'il est possible de mettre en correspondance, celui-ci allant en diminuant. L'apparence de la scène observée varie aussi selon l'angle sous lequel on la regarde : la précision de localisation des points détectés d'une image à l'autre, tout comme la ressemblance des voisinages sur lesquels les descripteurs photométriques sont calculés, diminuent.

Pour les approches de type RANSAC, il est alors nécessaire d'adapter la précision, et la taille des ensembles de consensus recherchés.

Le rôle de cette expérience est de montrer que les seuils déterminés automatiquement par notre approche varient significativement en fonction de la qualité des appariements qu'il est possible d'extraire entre deux images.

Le tableau 5.2 montre que notre approche est capable de trouver des groupes de tailles différentes (de 199 appariements à 121 ici), de précisions géométriques suffisamment différentes pour ne pas pouvoir être fixées à l'avance (entre 0.6 et 1.6 pixels), tout comme pour la contrainte photométrique (de -24 à -12). Pour les deux dernières images (valeurs séparées par un trait horizontal), des appariements le long des lignes épipolaires sont trouvés, comme montré dans les images du haut de la figure 5.1. Le tableau 5.3 indique lui le rang des appariements sélectionnés pour la photométrie, du plus proche voisin au k-ième plus proche. Plus la distance entre les caméras est importante, plus un nombre important d'appariements qui ne sont pourtant pas les plus ressemblants sont sélectionnés : cela montre l'importance de garder ces candidats à l'appariement. Il s'agit ici de données pour la matrice fondamentale.

On retrouve le même comportement pour l'homographie, comme indiqué dans le tableau 5.4. L'homographie fournissant une contrainte point/point, elle permet mieux de guider l'appariement que la matrice fondamentale : on obtient ainsi en proportion plus d'appariements venant de candidats de rangs de ressemblance photométrique plus élevés.

La séquence *pegase* nous permet de réaliser une expérience avec des images réelles contenant des motifs répétés. La figure 5.5 montre les appariements extraits pour un faible mouvement de translation, puis une translation plus forte. Dans le premier cas, le faible mouvement, la transformation géométrique trouvée est une homographie globale, telle que celles détectées pour les premières

← Baseline ↓	1	2	3	4	5
	164.6 (6.5)	4.3 (2.0)	4.2 (2.2)	0.6 (0.8)	0.1 (0.1)
	152.7 (5.3)	9.2 (0.9)	4.4 (0.8)	0.1 (0.2)	1.7 (0.6)
	124.0 (3.8)	18.8(1.7)	5.7 (0.7)	5.1 (0.4)	2.0 (0.4)
	114.1 (4.7)	26.7(1.8)	4.0 (0.8)	3.0 (0.2)	1.5 (0.7)
	97.0 (4.4)	29.3(1.6)	10.7(0.8)	5.4 (0.8)	0.8 (0.4)

TABLE 5.3 – *cube*, matrice fondamentale, rangs. Nombre d'appariements par rang de ressemblance photométrique (i-ième plus proche voisin) pour une distance entre caméras croissante. Ici, α est fixé à 5.

← Baseline croissante ↓	Nb. points	δ_G	log δ_D	log $\widetilde{\varepsilon}$
	185.2 (4.3)	3.79 (0.42)	-20.5 (0.3)	-17042 (222)
	168.4 (9.3)	4.90 (0.77)	-19.2 (3.7)	-14363 (358)
	111.8 (11.2)	1.61 (1.27)	-20.5 (0.6)	-12130 (160)
	95.1 (4.5)	0.82 (0.26)	-13.2 (1.4)	-10688 (126)
	99.1 (7.6)	1.17 (0.54)	-16.8 (0.8)	-10897 (182)
	89.7 (5.0)	0.77 (0.32)	-12.4 (1.6)	-10142 (181)
	93.5 (8.9)	1.37 (0.68)	-11.6 (1.0)	-9541 (201)
	98.3 (6.8)	2.10 (0.69)	-12.2 (0.2)	-9218 (204)
	88.5 (5.8)	2.14 (0.74)	-13.3 (0.8)	-8324 (266)

← Baseline croissante ↓	1	2	3	4	5
	169.6 (11.0)	6.1 (0.7)	6.6 (0.6)	1.9 (0.3)	0 (0)
	153.0 (7.5)	9.3 (1.7)	4.4 (0.7)	0 (0)	1.8 (0.5)
	86.1 (9.9)	15.0 (1.2)	4.1 (0.5)	5.0 (0)	1.6 (0.5)
	67.8 (4.1)	19.7 (0.7)	1.0 (0.1)	3.0 (0.1)	1.8 (0.4)
	58.4 (6.5)	25.7 (0.9)	9.9 (0.4)	4.3 (0.5)	1.0 (0.2)
	53.5 (3.7)	20.5 (0.9)	8.8 (0.4)	3.2 (0.7)	1 (0)
	47.5 (6.7)	24.3 (1.9)	13.2 (0.6)	4.7 (0.4)	1 (0)
	47.1 (6.0)	19.0 (0.9)	13.0 (0.8)	9.0 (0.2)	1.9 (0.3)
	34.5 (4.9)	17.3 (0.6)	22.3 (1.1)	8.0 (0.2)	0 (0)

TABLE 5.4 – *cube*, homographie, mouvement croissant - un nombre important d'appariements viennent de candidats qui ne sont pas plus proches voisins pour la ressemblance photométrique.

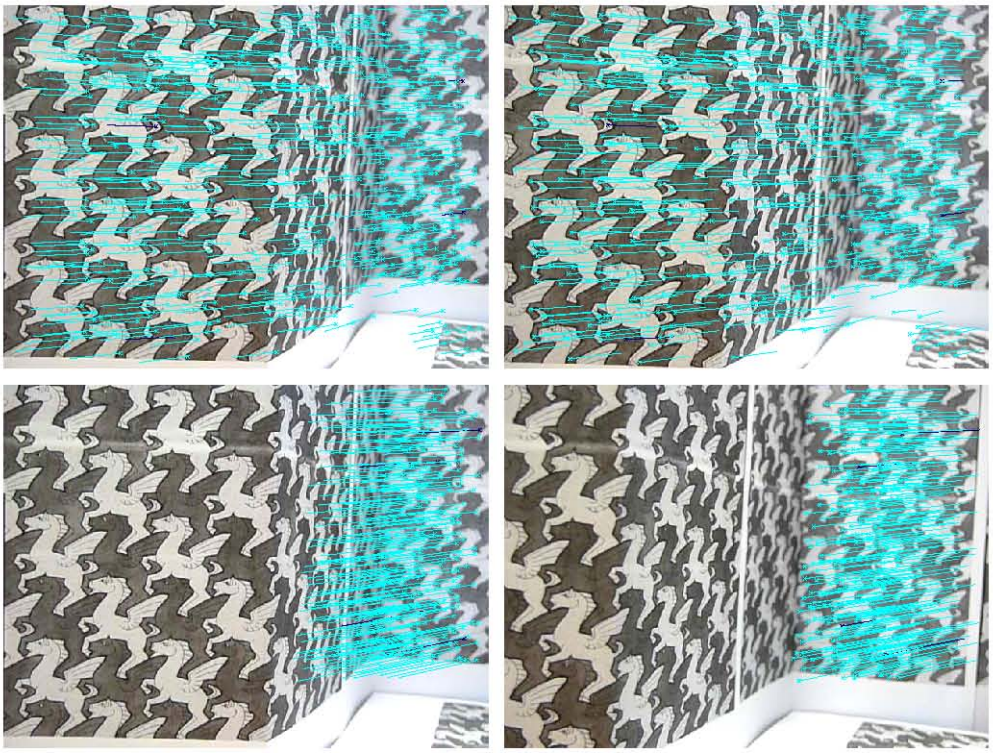

FIGURE 5.5 – *pegase*, homographie - En haut, faible mouvement de translation. En bas, mouvement de translation plus important. Pour une faible translation, l'homographie est approximée par une rotation équivalente, et permet d'obtenir un appariement fiable sur toute l'image. Pour une translation plus grande, l'homographie retrouvée permet la mise en correspondance des points qui sont situés sur le plan avec le plus grand support.

lignes du tableau 5.5. Ensuite, pour des mouvements de translation de plus grande amplitude, l'ensemble robuste sélectionné s'explique par une mise en correspondance de plans. Tout comme pour la séquence *cube*, l'homographie permet un bon guidage pour la recherche d'appariements, comme indiqué par le tableau 5.6. Les figures 5.6 et 5.7 montrent une ambiguïté importante dans les deux cas, celle-ci étant légèrement plus marquée pour la translation importante.

Après avoir testé notre approche sur des images contenant des motifs répétés, vérifions que son comportement est comparable sur des images réelles. D'abord, nous nous assurons que les performances obtenues sur une séquence ne contenant pas de motifs répétés sont analogues à d'autres approches telles que celle de Moisan et Stival [MS04b]. L'absence de motifs répétés est confirmé par l'histogramme des rapports de distance entre descripteurs présenté en figure 5.10. C'est l'objet du tableau 5.7 obtenu avec le début de la séquence *medusa*, pour laquelle nous estimons la matrice fondamentale. Un couple d'images de cette séquence est présenté en figure 5.8 pour ACM+F, et en figure 5.9 pour NNT+F.

← Baseline croissante ←	Nb. points	δ_G	log δ_D	log $\widetilde{\varepsilon}$	
	699.9 (31.9)	1.70 (0.34)	-20.2 (2.8)	-81797 (1744)	Homographie globale
	720.8 (21.1)	2.87 (0.37)	-18.9 (2.0)	-75699 (1596)	
	736.8 (11.9)	4.67 (0.52)	-14.4 (0.41)	-67004 (1557)	
	711.7 (15.8)	5.68 (0.67)	-15.2 (0.92)	-62432 (1158)	
	701.2 (11.2)	6.48 (0.49)	-14.5 (0.59)	-59131 (870)	
	658.1 (13.4)	7.48 (0.56)	-14.97 (0.46)	-53807 (839)	
	656.4 (27.4)	8.52 (0.55)	-14.7 (1.0)	-51798 (1827)	
	488.4 (120.5)	7.96 (1.91)	-16.9 (2.1)	-39756 (7687)	
	283.5 (47.3)	2.37 (2.18)	-16.0 (2.1)	-30102 (1668)	Homographie locale
	277.9 (37.5)	2.29 (1.87)	-16.4 (1.20)	-29598 (1469)	
	245.5 (10.5)	1.40 (0.49)	-15.9 (1.7)	-28052 (980)	

TABLE 5.5 – *pegase*, homographie - Les 8 premières lignes correspondent à un mouvement global : pour une faible translation, l'homographie est approchée par une rotation équivalente. La précision obtenue pour le groupe est de plus en plus faible. Pour les 3 dernières lignes, une homographie planaire est trouvée, d'où le nombre plus faible d'appariements retenus, mais avec une meilleure précision.

← Baseline croissante ←	1	2	3	4	5	6	7
	614.9 (20.2)	50.3 (5.5)	14.7 (2.1)	7.2 (1.8)	3.6 (1.1)	3.3 (0.8)	3.6 (1.8)
	629.1 (13.6)	41.8 (4.4)	17.6 (1.9)	10.8 (0.9)	7.2 (1.0)	8.4 (1.3)	1.2 (0.5)
	640.9 (10.5)	43.2 (1.8)	26.5 (0.8)	7.4 (0.9)	4.7 (0.5)	2.9 (0.4)	2.6 (0.5)
	610.3 (11.6)	50.3 (2.9)	22.9 (1.4)	7.9 (0.5)	6.4 (0.8)	4.0 (0.1)	4.0 (0.4)
	596.7 (9.0)	47.1 (1.2)	25.4 (1.3)	9.7 (0.5)	5 (0)	6.8 (0.7)	3.9 (0.4)
	552.0 (9.4)	67.1 (3.0)	14.0 (1.2)	7.3 (0.7)	7.7 (1.0)	2.0 (0.4)	2.5 (0.5)
	568.0 (21.2)	49.8 (3.0)	16.3 (2.4)	3.8 (0.9)	7.5 (1.3)	3.0 (0.2)	2.9 (0.5)
	414.1 (94.7)	34.1 (12.8)	18.9 (6.5)	7.8 (2.3)	4.5 (2.4)	4.3 (2.3)	1.5 (0.9)
	231.5 (46.1)	23.4 (2.5)	9.6 (1.0)	5.4 (1.2)	2.6 (0.7)	0.1 (0.2)	3.8 (0.7)
	222.4 (38.8)	23.2 (2.1)	15.4 (2.0)	2.0 (0.6)	5.0 (0.8)	2.4 (0.7)	3.6 (0.9)
	199.5 (8.5)	21.5 (1.5)	4.5 (0.6)	6.0 (0.6)	3.5 (0.8)	2.9 (0.3)	2.0 (0.7)

TABLE 5.6 – *pegase*, homographie, rang - Nombre d'appariements inliers en fonction de leur rang de ressemblance photométrique, du plus proche voisin au k-ième plus proche voisin.

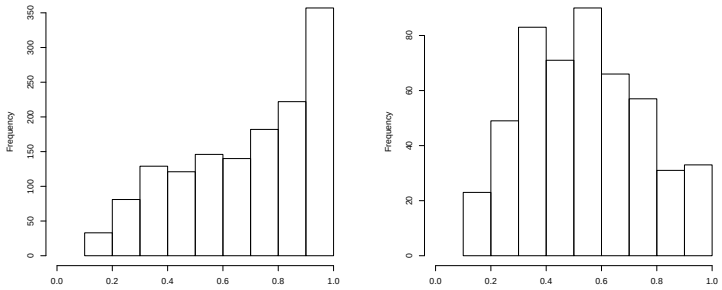

FIGURE 5.6 – *pegase*, petite translation - Histogramme des rapports de distance des descripteurs photométriques plus proche voisin et second plus proche voisin. A gauche, pour tous les points d'intérêt détectés. A droite, pour les points retenus après filtrage robuste, limité au plus proche voisin. Se priver des candidats ambigus à l'appariement en imposant un seuil sur la valeur du rapport de distance (un rapport $<$ à 0.6, par exemple), diminuerait sensiblement la taille du groupe retenu. Les points utilisés sont en effet situés partout dans l'histogramme.

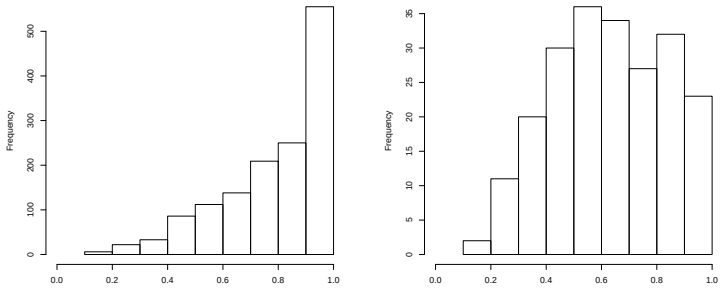

FIGURE 5.7 – *pegase*, grande translation - Histogramme des rapports de distance des descripteurs photométriques plus proches voisins et second plus proche voisin. A gauche, pour tous les points d'intérêt détectés. A droite, pour les points considérés inliers après filtrage robuste. Dans ce cas, malgré la présence de nombreux candidats difficiles à discriminer (valeurs proches de 1 sur l'histogramme de gauche), il est possible de sélectionner des appariements ambigus.

FIGURE 5.8 – *medusa* - Exemple de séquence sans motifs répétés pour ACM+F. Les deux images mises en correspondance sont disposées l'une sous l'autre. A droite, on peut voir le faisceau des droites épipolaires.

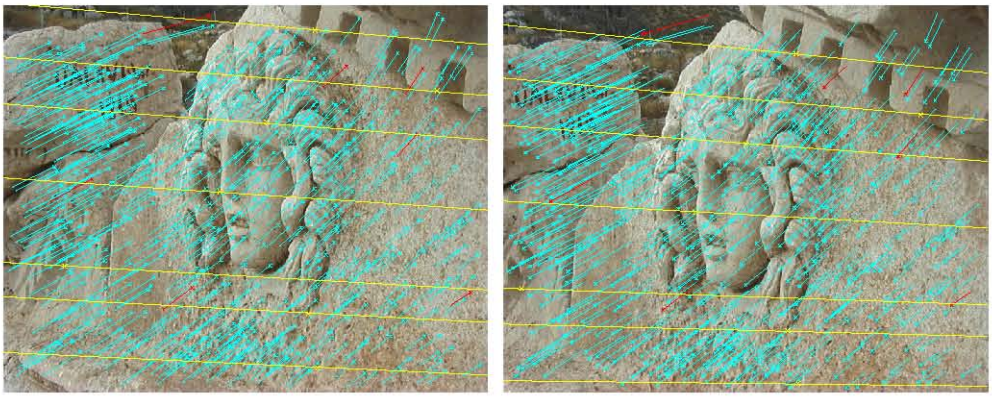

FIGURE 5.9 – *medusa* - Ici est presenté le résultat de mise en correspondance pour la méthode NNT+F. On obtient environ 600 appariements contre plus de 700 pour notre méthode. Comme sur la figure 5.8, chaque droite épipolaire trouvée passe par le point correspondant d'un couple d'appariements de contrôle.

	Nb. points	δ_G	log δ_D	log $\widetilde{\varepsilon}$
← Baseline croissante	1333.3 (1.6)	0 (0)	-74.6 (0.6)	-563281 (148)
	768.5 (22.9)	0.84 (0.15)	-22.5 (2.8)	-54609 (1099)
	765.1 (35.7)	0.77 (0.13)	-20.5 (3.7)	-53401 (1089)
	722.8 (33.9)	0.97 (0.16)	-22.4 (3.5)	-50158 (977)
	745.3 (42.5)	0.83 (0.10)	-20.8 (4.5)	-51651 (941)
	690.1 (18.2)	1.01 (0.14)	-16.6 (1.7)	-43777 (662)
	706.1 (15.6)	1.31 (0.22)	-14.7 (0.7)	-41717 (722)
	651.7 (25.3)	1.35 (0.34)	-14.8 (0.5)	-38311 (681)

	1	2	3	4	5
← Baseline croissante	1333.3 (1.6)	0 (0)	0 (0)	0 (0)	0 (0)
	767.1 (22.8)	1.4 (1.24)	0 (0)	0 (0)	0 (0)
	762.7 (34.6)	2.34 (1.49)	0.05 (0.22)	0.02 (0.20)	0.02 (0.10)
	720.9 (33.8)	1.83 (1.16)	0.04 (0.20)	0.01 (0.10)	0 (0)
	741.3 (41.4)	3.5 (1.5)	0.35 (0.50)	0.04 (0.20)	0.01 (0.10)
	688.9 (18.2)	1.14 (0.96)	0.04 (0.24)	0 0	0.03 (0.17)
	703.2 (15.3)	1.26 (1.03)	1.15 (0.83)	0.27 (0.47)	0.06 (0.24)
	647.0 (24.5)	3.53 (1.16)	0.28 (0.53)	0.17 (0.51)	0.10 (0.33)

TABLE 5.7 – *medusa*, matrice fondamentale, mouvement croissant - Cette scène contient peu de motifs répétés. Notre approche se comporte alors de façon très proche de la méthode *A Contrario* de Moisan et Stival.

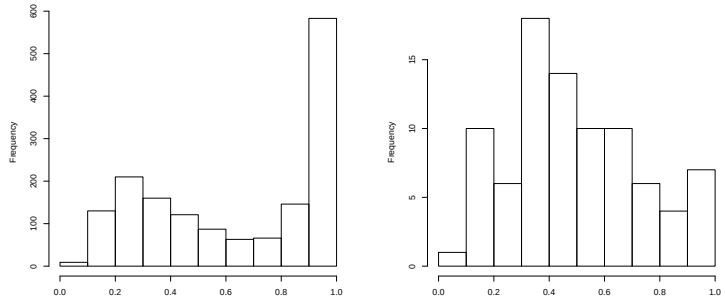

FIGURE 5.10 – *medusa* - Histogramme des rapports de distance des descripteurs photométriques plus proche voisin et second plus proche voisin. A gauche, pour tous les points d'intérêt détectés, avant le filtrage robuste. A droite, pour les points considérés inliers après filtrage robuste. Les appariements trouvés sont situés partout dans l'histogramme, même dans la partie proche de 1.

5.3 Influence des différentes distances entre descripteurs

Nous testons ici l'influence du paramètre de réduction combinatoire $\widetilde{\varepsilon}$ (section 4.3.2.1), en utilisant les différentes distances entre descripteurs rappelées en section 4.2.3. Rappelons l'équation 4.39 :

$$N_1 N_2 d_D(D(x_i), D(y_j)) \leq \widetilde{\varepsilon}. \tag{5.2}$$

Comme les descripteurs SIFT sont uniquement invariants aux zooms et aux rotations, le nombre de points d'intérêt qu'il est possible d'apparier diminue fortement dès que le changement de point de vue est trop fort, comme dans toute méthode reposant sur SIFT. Nous considérons donc un mouvement faible entre les deux vues, afin que le manque d'invariance des descripteurs SIFT n'interfère pas dans les comparaisons de distances entre descripteurs. Dans ce cas, les points en correspondance doivent avoir des descripteurs très proches. Le test est ici réalisé avec le modèle utilisant la matrice fondamentale. Pour l'illustrer, la figure 5.11 montre quelques résultats avec la distance CEMD-SUM. Remarquons que les images contiennent de nombreux motifs répétés, il s'agit d'une situation difficile.

Les tableaux 5.8 et 5.9 regroupent des statistiques à propos de cette expérience. Au sujet de l'influence de $\widetilde{\varepsilon}$, on peut voir que le réduire diminue aussi le nombre d'appariements candidats parmi lesquels l'ensemble le plus significatif est recherché, mais n'a presque aucun impact sur le cardinal de cet ensemble. En d'autres termes, diminuer la valeur de $\widetilde{\varepsilon}$ accélère la recherche tout en rejetant principalement des appariements incorrects. Remarquons qu'au moins 20-25% des mises en correspondance ne sont pas réalisées entre points dont les descripteurs sont plus proches voisins. Il aurait été impossible de les obtenir avec la procédure habituelle de mise en correspondance de SIFT (mise en correspondance au plus proche voisin, pourvu que le rapport de distance entre le plus proche et le second plus proche voisin est inférieur à un seuil fixé). Dans ce cadre expérimental particulier (un petit mouvement entre les caméras), ces tableaux montrent que le groupe le plus significatif est plus grand avec CHI2-SUM, CEMD-SUM, CHI2-MAX, CEMD-MAX. Avec ces distances, le pourcentage de correspondances de rang 1 est significativement plus petit pour EUC-SUM, EUC-MAX et MAN-MAX, ce qui prouve que ces distances trient les appariements de façon moins consistante que les autres. Cela confirme les résultats de [LO07] et de [RDG09a] avec un autre protocole expérimental.

A partir de ces résultats et d'autres expériences sur des images réelles, nous avons décidé de fixer pour la suite $\widetilde{\varepsilon} = 10^{-2}$, (ce qui réalise un bon compromis entre la réduction de la complexité et la taille du groupe le plus significatif), et d'utiliser la métrique CEMD-SUM. Nous avons constaté par ailleurs que la distance du CHI2 donne des résultats mitigés. Ainsi, l'algorithme proposé avec CEMD-SUM donne de bons résultats dans l'expérience difficile de la figure 5.14, alors qu'aucun groupe n'est produit par les autres métriques.

5.4 Mise en correspondance de points et aliasing perceptuel

L'objectif de cette section est de montrer que la méthode proposée nous permet d'obtenir plus d'appariements qu'un critère habituel de mise en correspondance robuste en présence de motifs répétés. En effet, une partie importante des points d'intérêt mis en correspondance ne correspond pas au descripteur du plus proche voisin, mais d'appariements avec un rang plus élevé. Nous comparons l'algorithme proposé avec une méthode standard utilisant les étapes 2 et 3 présentées au chapitre 2, en section 2.1 :

- mise en correspondance selon le rapport de distance des plus proches voisins (NNT)(distance Euclidienne, seuil sur le rapport fixé à 0.6).
- Sélection robuste avec l'approche *A Contrario*-RANSAC de [MS04b].

distance	$\widetilde{\varepsilon}$	# corr. poten.	# groupe le plus signif.	% de corr. de rang 1
MAN-SUM	1	3094	184.6 (19.6)	86.2
	10^{-2}	2253	190.0 (21.8)	85.2
	10^{-4}	1733	198.5 (20.4)	85.2
	10^{-6}	1336	204.0 (21.6)	86.4
	10^{-8}	1001	214.7 (19.2)	85.4
	10^{-10}	808	215.2 (17.9)	84.5
EUC-SUM	1	10197	167.3 (21.6)	60.7
	10^{-2}	9365	169.3 (22.5)	61.6
	10^{-4}	7247	173.5 (23.4)	62.1
	10^{-6}	5374	173.6 (22.8)	61.8
	10^{-8}	3922	177.1 (24.4)	62.1
	10^{-10}	2865	190.4 (21.3)	59.9
CHI2-SUM	1	3091	218.1 (6.1)	79.5
	10^{-2}	2088	218.1 (6.5)	79.6
	10^{-4}	1638	219.3 (6.0)	79.6
	10^{-6}	1296	219.1 (5.7)	79.6
	10^{-8}	1020	220.4 (4.5)	79.4
	10^{-10}	794	221.4 (4.3)	79.4
CEMD-SUM	1	2027	219.5 (6.4)	76.1
	10^{-2}	1409	220.6 (4.9)	76.2
	10^{-4}	999	218.3 (3.9)	76.6
	10^{-6}	663	202.7 (3.0)	78.4
	10^{-8}	407	172.8 (2.8)	85.5
	10^{-10}	274	146.2 (2.2)	90.3

TABLE 5.8 – Images *Synthétiques*. Comparaison des distances entre descripteurs et influence de $\widetilde{\varepsilon}$. De gauche à droite : les quatre distances entre descripteurs sont testées, avec six valeurs de $\widetilde{\varepsilon}$ dans la gamme 1 - 10^{-10}, le nombre de correspondances potentielles obtenues après l'étape de réduction combinatoire (section 4.3.2.1), le cardinal du groupe le plus significatif, et la proportion des plus proches voisins parmi ce groupe d'appariements.

distance	$\widetilde{\varepsilon}$	# corr. poten.	# groupe le plus signif.	% de corr. de rang 1.
MAN-MAX	1	10290	221.8 (19.5)	51.2
	10^{-2}	10290	219.6 (19.1)	51.7
	10^{-5}	10290	220.2 (19.8)	51.7
	10^{-10}	10270	221.3 (19.3)	52.0
	10^{-15}	5478	237.0 (23.8)	51.8
	10^{-20}	2529	259.9 (20.6)	54.7
EUC-MAX	1	10290	210.4 (18.2)	51.1
	10^{-2}	10290	213.3 (19.4)	50.3
	10^{-5}	10290	212.6 (19.3)	50.9
	10^{-10}	10019	212.6 (18.9)	51.1
	10^{-15}	5035	224.0 (20.6)	50.5
	10^{-20}	2248	259.6 (18.6)	55.1
CHI2-MAX	1	9984	228.8 (4.8)	81.4
	10^{-2}	7670	228.3 (4.8)	81.4
	10^{-5}	3783	228.0 (5.0)	81.4
	10^{-10}	1736	229.2 (4.5)	81.4
	10^{-15}	1018	228.5 (3.8)	81.9
	10^{-20}	606	229.0 (3.2)	82.3
CEMD-MAX	1	8096	225.1 (5.8)	75.3
	10^{-2}	4126	225.1 (5.9)	75.6
	10^{-5}	3215	224.5 (5.3)	76.2
	10^{-10}	1144	221.7 (3.8)	77.3
	10^{-15}	494	201.8 (3.0)	83.0
	10^{-20}	234	142.7 (2.0)	90.6

TABLE 5.9 – Images *Synthétiques*. Comparaison des distances entre descripteurs et influence de $\widetilde{\varepsilon}$ (suite). Quelques cas donnent 10290 appariements, ce qui correspond à un paramètre d'implantation : pour des questions d'occupation mémoire, nous ne conservons au maximum qu'au plus 30 candidats à la mise en correspondance par groupe, même si $\widetilde{\varepsilon}$ en fait garder plus. Cela signifie que $\widetilde{\varepsilon}$ réduit l'ensemble des appariements à une taille constante.

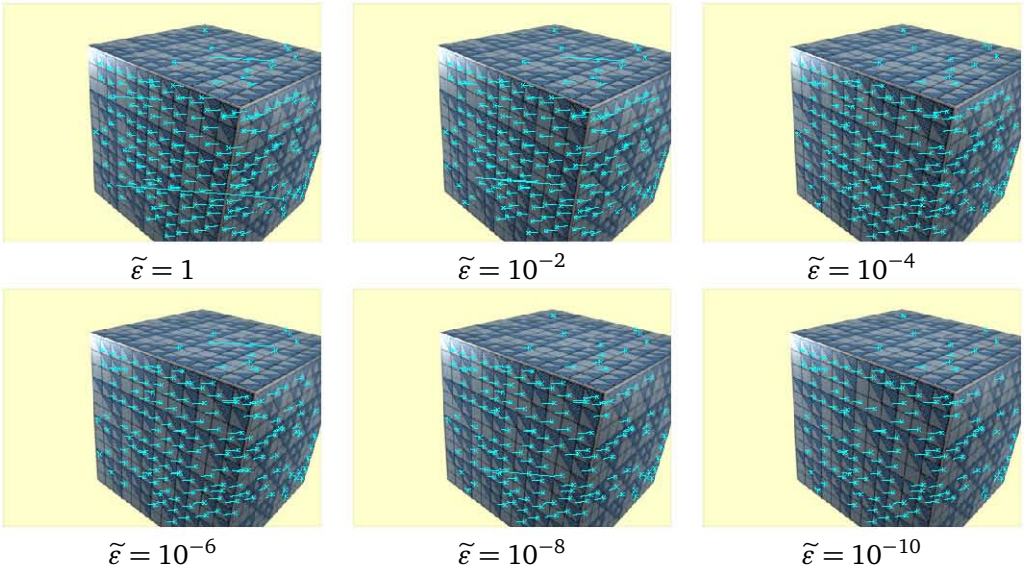

FIGURE 5.11 – Images *Synthétiques*. Dans cette expérience, nous recherchons le groupe le plus significatif compatible avec une matrice fondamentale entre deux vues, pour la distance CEMD-SUM, et six valeurs de $\widetilde{\varepsilon}$. Nous présentons ici uniquement les premières vues, les segments bleus correspondant au mouvement apparent d'un point d'intérêt (matérialisé par une croix) entre les deux vues. On constate que quelques appariements aberrants sont néanmoins sélectionnés. Un examen minutieux montre que ceux-ci sont situés le long des droites épipolaires qui leurs sont associées, et ne peuvent donc pas être détectés lors d'une mise en correspondance entre deux vues. Dans toutes les expériences (quels que soient la distance et $\widetilde{\varepsilon}$ comme dans les tableaux 5.8 et 5.9), la distance moyenne d'un point à une droite épipolaire est environ de 0.2-0.3 pixel. 343 points d'intérêt SIFT ont été extraits de l'image 1, et 321 de l'image 2.

5.4.1 Motifs répétés et homographie

Nous utilisons ici l'homographie comme contrainte géométrique. Lorsque l'on est confronté à des motifs répétés, le nombre d'appariements sélectionnés avec NNT est petit, comme montré dans l'image de droite de la figure 5.12 : les points d'intérêt "répétés" sont généralement éliminés dès cette première étape, et ne peuvent bien sûr pas être filtrés par l'algorithme RANSAC consécutif. Notre méthode, comme montré dans l'image de gauche de la figure 5.12, permet d'obtenir de plus nombreuses correspondances. Les nombreuses correspondances supplémentaires coïncident avec les appariements qui ne sont pas plus proche voisin pour la distance entre descripteurs. Comme indiqué par le tableau 5.10, nous montrons que pour 128 points mis en correspondance, 42 sont de rang 1 (plus proches voisins), contre 86 de rangs plus élevés. Les figures 5.13 et 5.14, *Sears* et *Flat Iron* présentent des cas réels de façade d'immeuble.

	Nombre d'appariements			
Rang	*Monkey*	*Loria*	*Sears*	*Flat Iron*
1	42	98	532	8
2	23	32	24	3
3	17	29	12	1
4	11	18	3	1
5	8	20	5	0
6	8	19	3	0
7	4	11	3	0
8	8	5	2	0
9	3	13	1	0
10	2	8	0	0
11	0	15	0	0
12	2	7	1	0
>12	0	37	3	0
Total	128	312	589	13

TABLE 5.10 – Nombre d'occurrences des n-ème plus proche voisins sélectionnés par la méthode ACM. *Monkey* correspond à la figure 5.12 (412 contre 445 points extraits), *Loria* à la figure 3.7 (2,562 vs 2,686), *Sears* à la figure 5.13 (2,787 vs 2,217), *Flat Iron* à la figure 5.14 (756 vs 598). Remarquons l'ambiguïté perceptuelle importante dans ces paires d'images. Comme indiqué à la figure 3.7, la méthode NNT ne fonctionne pas du tout pour l'expérience *Loria*. Des rangs plus grands que deux sont d'autant plus fréquents que la scène contient des motifs répétés, et ne peuvent pas être obtenus par la méthode NNT.

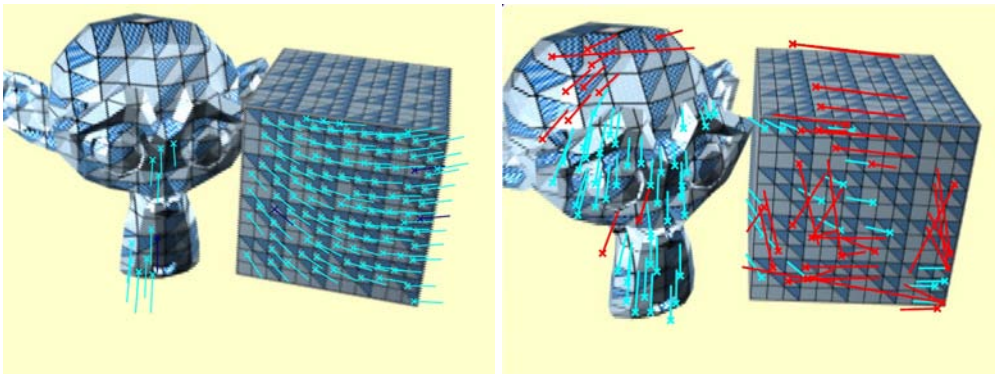

FIGURE 5.12 – *Monkey*, contrainte homographique. Deux images avec des motifs répétés. Sur la gauche, le modèle ACM proposé, la plupart des motifs sur le plan dominant sont détectés (les segments représentent le mouvement apparent entre les deux vues). Sur la droite, la seconde image avec les appariements obtenus avec NNT (les deux couleurs) et NNT+H (les inliers sont en bleu, les outliers en rouge). Beaucoup d'autres appariements sont obtenus avec l'algorithme ACM.

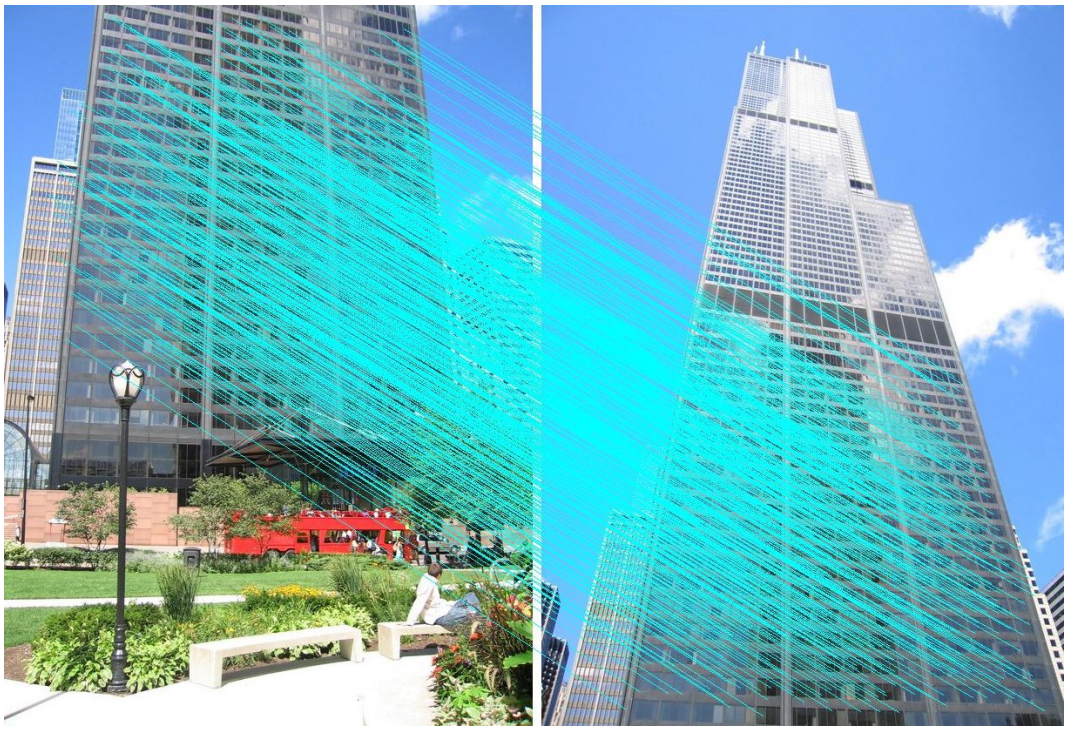

FIGURE 5.13 – *Sears*, contrainte homographique. 589 appariements sont trouvés avec la méthode ACM+H.

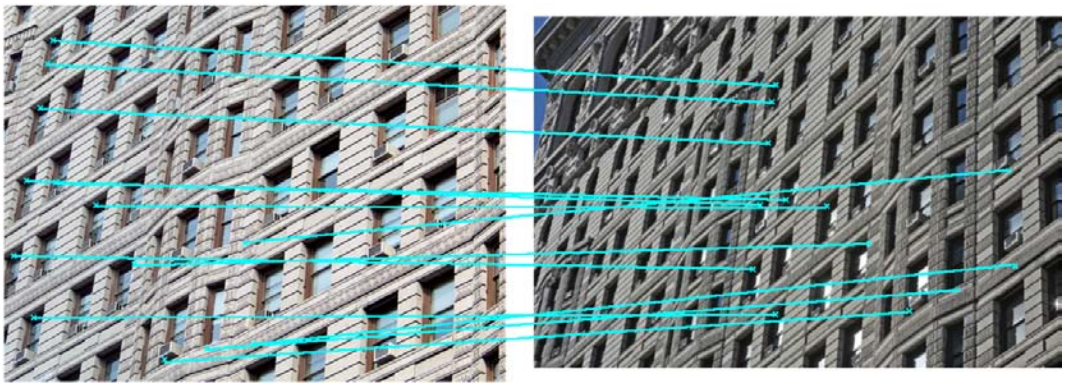

FIGURE 5.14 – *Flat Iron*, contrainte homographique. 13 appariements sont trouvés avec la méthode ACM, tous sont corrects. Ici NNT+H ne donne aucun groupe cohérent. C'est un problème très difficile qui nécessite d'augmenter le nombre de tirages effectués à 100 000 : remarquons la forte ambiguïté perceptuelle ainsi que l'important changement de contraste.

5.4.2 Motifs répétés et contrainte épipolaire

Dans cette section, nous testons le comportement de l'algorithme AC avec la contrainte épipolaire. Un exemple peut être vu sur la figure 5.15. Comme la contrainte épipolaire est moins stricte que l'homographie (c'est une contrainte point / droite), quelques appariements erronés sont inévitables. En effet, deux points peuvent être associés parce qu'ils sont "par hasard" à proximité de la droite épipolaire associée, bien qu'ils ne correspondent pas au même point en 3 dimensions. Une telle situation ne permet pas de lever l'ambiguïté dans le cas de deux vues. Une difficulté importante se pose alors lorsqu'on est confronté à de l'aliasing perceptuel. Si les motifs répétés sont presque alignés avec les droites épipolaires, alors les appariements erronés entre motifs similaires sont inévitables, menant à un ensemble significatif incohérent. Ce phénomène se produit quand les motifs répétés sont parallèles à la ligne de base entre les deux caméras, ce qui est une situation plutôt fréquente. De tels appariements sont visibles par exemple dans les figures 5.16 et 5.17.

Ces deux figures permettent une comparaison entre les méthodes NNT et ACM. Comme la méthode ACM est plus robuste à l'ambiguïté perceptuelle, nous obtenons plus d'appariements qui sont distribués de façon plus dense dans les images. Cependant, un examen minutieux montre que de nombreux appariements sont aberrants, malgré le fait que les points d'intérêt soient positionnés près des droites épipolaires correspondantes. Comme on peut le voir, cela permet cependant une estimation plus précise du faisceau des droites épipolaires. Nous avons sélectionné à la main des points (x, y) en correspondance dans chaque vue (en particulier dans les zones où presque aucune correspondance n'est obtenue avec NNT), et avons dessiné les droites épipolaires associées $(Fx, F^T y)$. La droite Fx (resp. $F^T y$) doit rencontrer le point y (resp. x). Ici F est réestimée sur l'ensemble de consensus obtenu par NNT+F ou ACM. La réestimation consiste en minimiser la distance de Sampson par l'algorithme de Powell [HZ00, Zha98]. Bien qu'une différence importante existe avec NNT (figure 5.16), la méthode ACM permet une estimation plus précise.

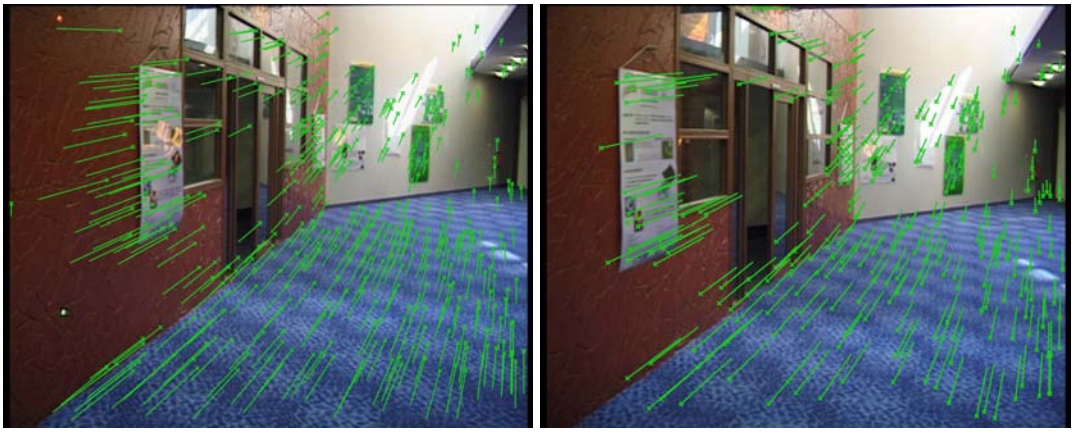

FIGURE 5.15 – *Corridor*, contrainte épipolaire. La méthode ACM (sur la gauche) permet d'obtenir 423 appariements. 405 appariements ont rang 1, 13 rang 2, 2 rang 3, 1 rang 4, 1 rang 6, 1 rang 10. La méthode NNT+O (sur la droite) permet d'obtenir 295 parmi 316 appariements de NNT. Les appariements supplémentaires sont sur la moquette et sur le mur. 1269 points d'intérêt ont été extraits de l'image 1, 1360 de l'image 2.

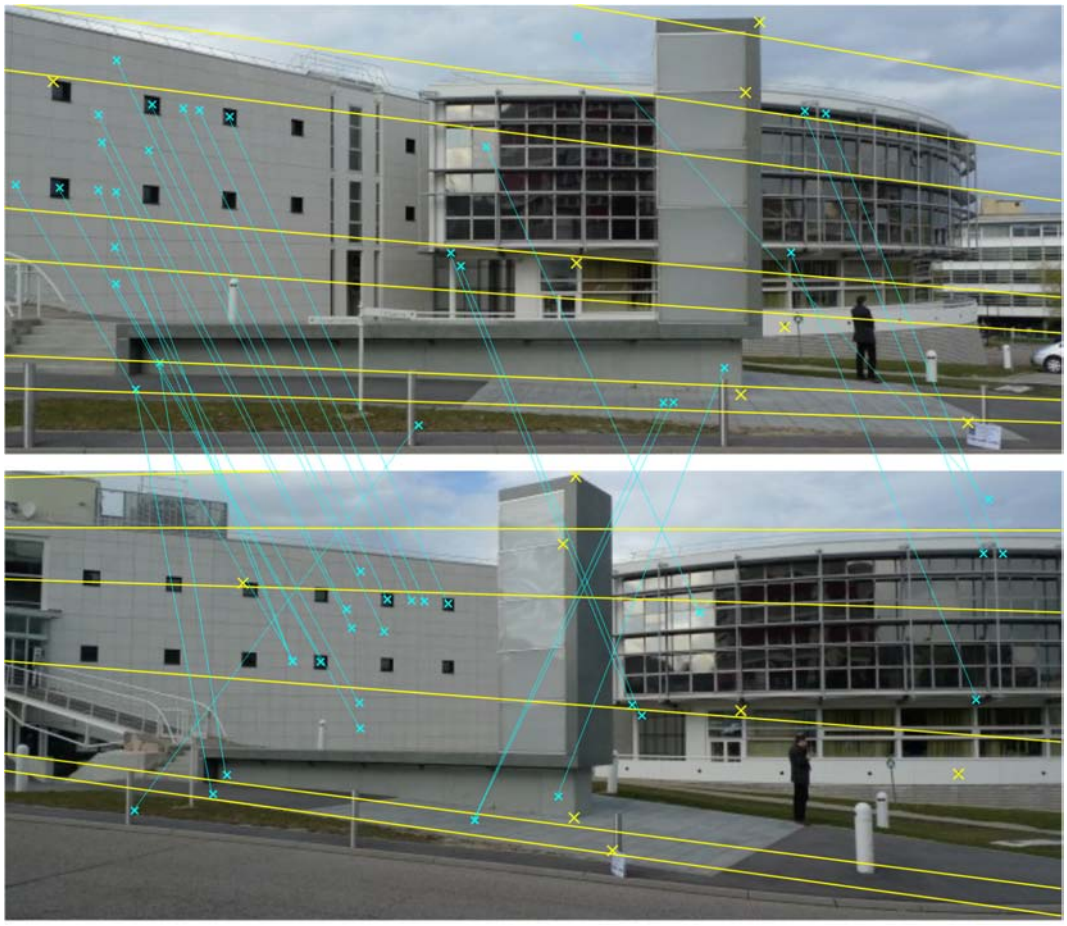

FIGURE 5.16 – *Bâtiment du Loria*. Méthode NNT+F (pour la contrainte épipolaire) entre les vues du haut et du bas. Les segments en bleu représentent le mouvement apparent des points d'intérêt, signalés par une croix. 630 points d'intérêt ont été extraits de l'image 1, 603 de l'image 2. Quelques appariements aberrants sont visibles. Cependant, leur distance à leur droite épipolaire respective est petite. Remarquons que le mouvement apparent du premier plan (une structure parallélépipédique) est très différent du mouvement de l'arrière plan. Quelques paires de points insérées à la main (en jaune) montrent que le faisceau épipolaire (réestimé à partir de l'ensemble de consensus) dévie de façon importante de celui correct. Les droites épipolaires devraient en effet rencontrer les croix jaunes. La distance mesurée entre points et droites épipolaires varie alors entre 4 et 20 pixels.

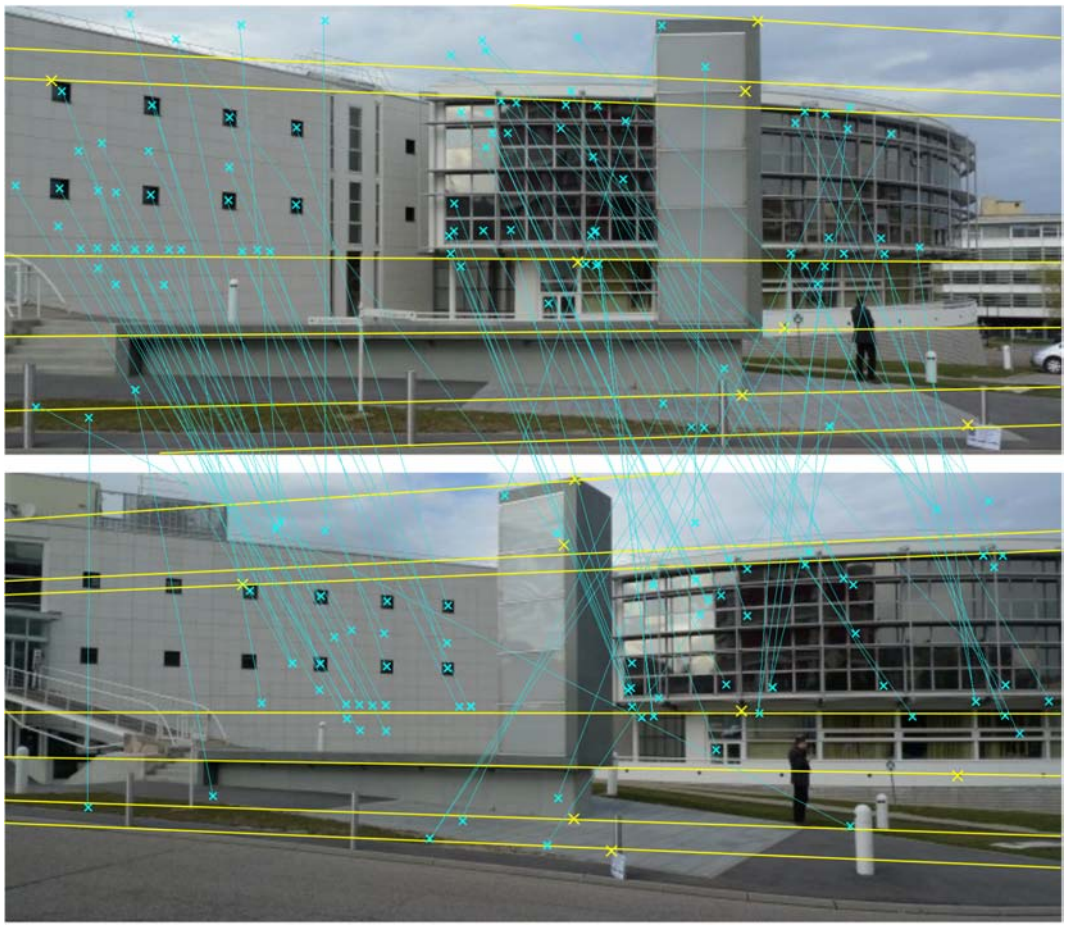

FIGURE 5.17 – *Bâtiment du Loria*. Méthode ACM+F (pour la contrainte épipolaire) entre les vues du haut et du bas. Comparé à la figure 5.16, plus d'appariements sont visibles. En particulier, les fenêtres répétées sur le côté gauche sont toutes détectées. Cependant, quelques faux appariements "inévitables" sont aussi obtenus, comme ceux entre les structures de la façade visible côté gauche qui sont décalés le long des droites épipolaires (comparé à la position des fenêtres, le même phénomène se produit en partant de la droite). Néanmoins, on peut voir sur les appariements saisis à la main (en jaune) que les droites épipolaires associées (réestimées) sont plus proches. La distance mesurée est de moins de 5 pixels, excepté pour un point sur la structure parallélépipédique qui est encore à 15 pixels.

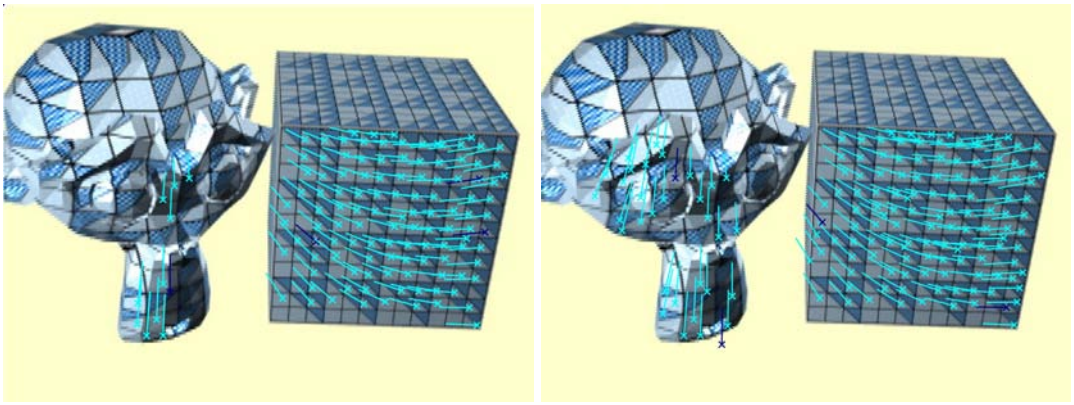

FIGURE 5.18 – *Monkey, homographie* - A gauche, notre méthode sans relâcher les seuils de détection de SIFT. A droite, avec les seuils de détection relâchés : un groupe d'appariements supplémentaires est visibles sur la gauche de la tête du singe.

5.4.3 Avec des seuils de détection plus permissifs

Dans cette section, nous montrons qu'il est possible de modifier certains seuils de détection de l'algorithme SIFT, afin d'extraire plus de points d'intérêt que nous sommes maintenant capables de mettre en correspondance correctement.

Dans l'approche SIFT, la valeur de la différence de gaussienne calculée pour un point d'intérêt doit être supérieure à un seuil fixé, assurant ainsi que le point détecté ne correspond pas à une zone peu contrastée dite "plate". Dans le code de D. Lowe, nous abaissons ce seuil de 0.04 à 0.01. Nous modifions aussi le rapport des valeurs propres du Hessien (dans l'espace échelle) associé avec chaque point d'intérêt. Il sert comme indicateur de la *cornerness* du point. Nous le relâchons de 10 à 20. Lorsque l'on fait de même pour la méthode NNT suivi d'un filtrage robuste a contrario, nous obtenons aussi quelques appariements supplémentaires : 47 appariements contre 41 pour la paire d'images *Monkey*, et plusieurs sont des outliers pour l'homographie. Au contraire, les résultats montrent un gain sensible avec notre méthode, comme illustré sur la figure 5.18. La figure 5.19 en est un autre exemple. Le tableau 5.11 regroupent des statistiques sur les rangs photométriques des inliers sélectionnés. Comparé au tableau 5.10, nous obtenons plus de 30% d'appariements supplémentaires. Prendre en compte les appariements avec un rang plus grand que 1 est d'autant plus utile, ils représentent alors plus de la moitié des appariements supplémentaires .

5.4.4 Mise en correspondance avec une ligne de base plus grande

Un des gains de notre modèle *a contrario* est une robustesse accrue pour la mise en correspondance avec une ligne de base plus grande. Si le changement d'aspect entre les deux images est trop important, alors l'invariance de SIFT, limitée aux zooms et rotations, rend difficile la mise en correspondance de façon fiable des points d'intérêt à l'aide de leurs descripteurs photométriques. Comme notre modèle a contrario permet de sélectionner des appariements moins ressemblants pour leurs descripteurs, pourvu que la contrainte géométrique soit satisfaite, il est plus robuste pour une ligne de base plus grande qu'avec NNT+F, comme visible sur la figure 5.20.

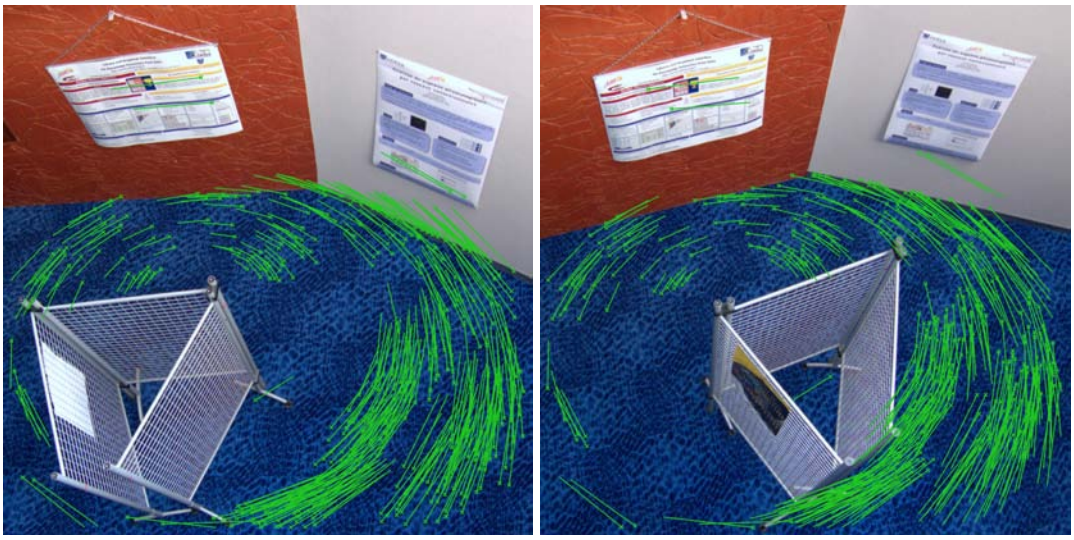

FIGURE 5.19 – *Loria, homographie* - 379 appariements avec les seuils de détection assouplis contre 312 sinon. Cette expérience illustrant le problème de l'aliasing perceptuel est très difficile : le rapport entre le nombre d'appariements de rang 1 et le nombre de points d'intérêt extraits est de seulement environ 7%. Notre méthode arrive cependant à extraire des correspondances cohérentes. Comme montré en figure 3.7, l'approche NNT en fonctionne pas ici.

	Nombre d'appariements		
Rang	*Monkey*	*Lab*	*Sears*
1	55	114	671
2	25	39	40
3	17	33	17
4	11	22	3
5	7	23	5
6	8	21	7
7	4	14	3
8	9	7	4
9	3	14	1
10	2	11	1
11	0	14	2
12	2	12	0
>12	0	55	6
Total	143	379	760

TABLE 5.11 – Nombre d'occurrences des n-ième plus proches voisins retenus par notre méthode lorsque les seuils de détection de SIFT sont assouplis, à comparer avec la table 5.10.

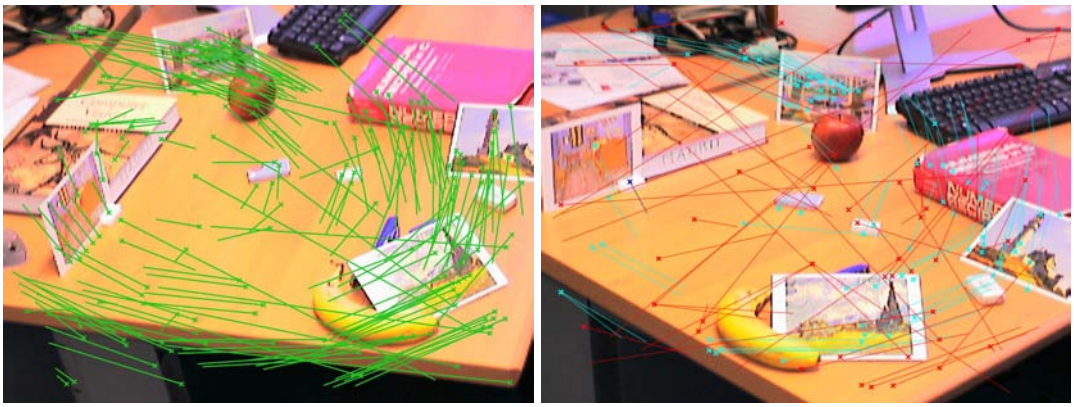

FIGURE 5.20 – *Bureau, contrainte épipolaire* - Mise en correspondance entre deux vues avec un fort changement de point de vue. Sur la gauche, on peut voir le mouvement apparent de 182 appariements (dont 126 sont premiers proches voisins pour la distance entre descripteurs). Sur la droite, 72 appariements avec NNT+F. Le seuil du rapport pour NNT est ici fixé à 0.8 (seulement 34 appariements sont obtenus avec un rapport de 0.6). Bien que quelques appariements incorrects soient visibles sur la gauche (à cause de points proches par hasard de leur droite épipolaire correspondante), la distribution des points est plus dense que sur l'image de droite, ce qui permet habituellement une estimation plus précise de la structure et du mouvement de la caméra.

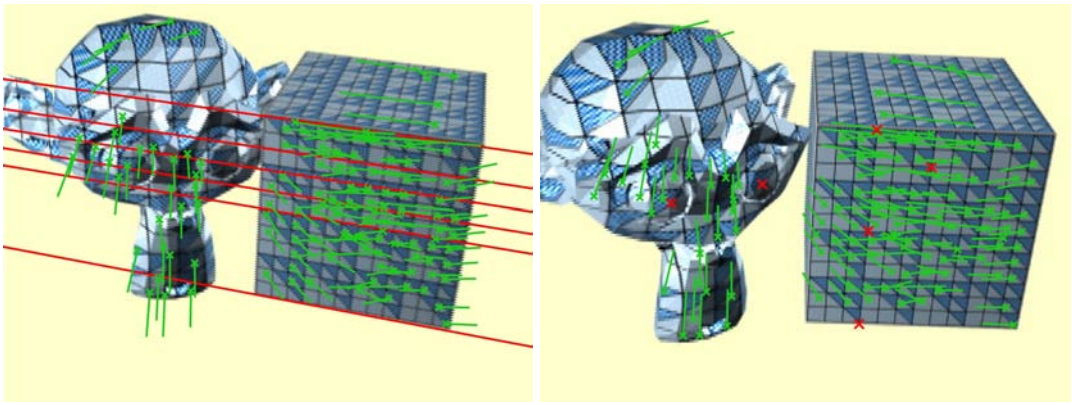

FIGURE 5.21 – *Monkey*, contrainte épipolaire. Cas d'erreur. La géométrie retrouvée correspond aux lignes de fuite. Dans ce cas, la contrainte point / droite ne permet pas de résoudre l'ambiguïté perceptuelle et donne alors de fausses correspondances. Quelques points insérés en rouge montrent le faisceau de droites épipolaires, qui correspond à l'alignement des motifs le long des ligne de fuites, et non pas au vrai mouvement.

5.4.5 Quand l'aliasing perceptuel ne peut être résolu

Si la quantité de motifs répétés est vraiment importante, et s'ils sont proches du faisceau des droites épipolaires, alors l'algorithme ACM ne produit pas un résultat convenable pour la contrainte épipolaire. En effet, lorsque l'on cherche des correspondances de points sous contrainte épipolaire, nous recherchons l'ensemble "le plus consistant" (au sens du NFA) par rapport au faisceau de droites. Cependant, dans cette situation, il est possible qu'il y ait trop d'appariements aberrants le long des droites selon lesquelles les motifs répétés sont alignés, comme on peut le voir en figure 5.22. Cette configuration correspond à la *double-nail illusion* de [KvdG80]. Le faisceau obtenu est alors dégénéré : il correspond à un groupe de lignes de fuite, qui concourent non plus en l'épipôle, mais en un point de fuite. Par exemple, dans la figure 5.21, on peut voir que le groupe le plus significatif est composé de faux appariements parmi les points qui sont situés sur un plan dominant le long des droites parallèles à un des côtés du cube. Remarquons que si la plupart des points d'intérêt ne se situent pas sur un plan dominant, alors l'algorithme ACM n'est pas attiré par les lignes de fuite et est alors capable de déterminer la géométrie correcte, même dans des cas où les images sont principalement composées de motifs répétés (comme dans la figure 5.11 où les points ne sont pas concentrés sur un unique plan). Remarquons aussi que la plus stricte contrainte point/point du cas homographique (comparez les figures 5.21 et 5.12) permet de déterminer un ensemble plausible, au contraire du cas utilisant la contrainte épipolaire. Nous terminons en remarquant que trouver des appariements à l'aide de la contrainte épipolaire est simplement impossible si les motifs répétés sont distribués le long de droites dans un plan dominant. Cette limitation est d'ailleurs commune à toutes les méthodes reposant sur RANSAC et cette contrainte.

5.5 Conclusion

La contribution apportée par cette méthode est un gain significatif pour le nombre d'appariements obtenus, en particulier en présence de motifs répétés (l'ambiguïté perceptuelle) et pour une ligne de base grandissante. Ces points supplémentaires sont obtenus en allant chercher des appariements de rangs supérieurs à 1, ou parmi ceux plus proche voisin qui sont rejetés par la méthode du

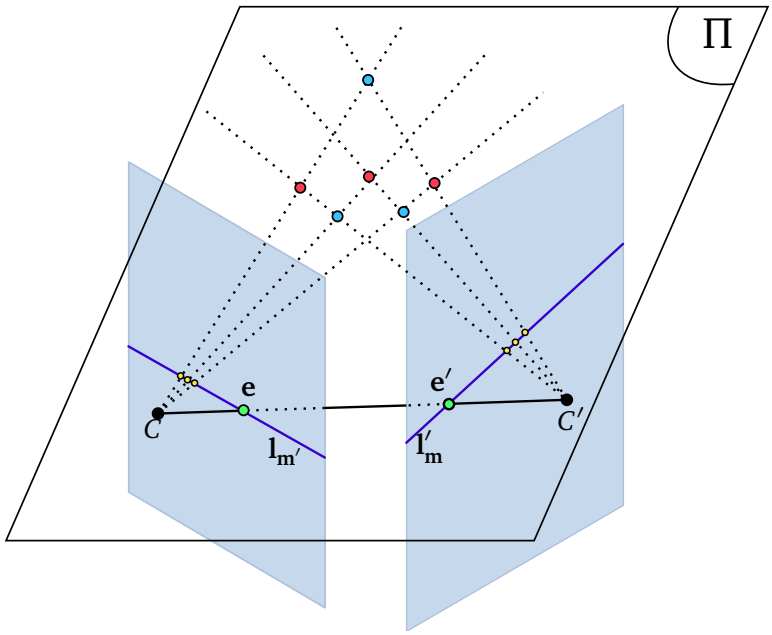

FIGURE 5.22 – Pour les 3 points jaunes visibles dans chaque vue, s'ils sont les projections de points 3D coplanaires de parties identiques en apparence de la scène, il n'est pas possible de distinguer si ce sont les images des points 3D bleus ou rouges. Une seule configuration est possible bien sûr, mais dans ce cas, ni la contrainte photométrique (les descripteurs sont proches), ni la contrainte géométrique (les points sont situés sur les droites épipolaires correspondantes), ne permettent de les distinguer.

seuil sur le rapport de distances.

De plus, cette méthode ne fait pas intervenir de paramètres difficiles à régler. L'équilibrage des contraintes géométriques et photométriques, l'augmentation de la distance entre les caméras, et l'influence des différentes distances photométriques ont ainsi été étudiés.

Quelques situations avec des motifs répétés restent insolubles avec cette méthode. Quand la répétition d'un motif est distribuée le long d'une droite épipolaire, il est impossible de mettre en correspondance correctement les motifs dans le cadre de la géométrie intrinsèque à deux vues. Utiliser une troisième vue convenablement choisie permettrait de résoudre l'ambiguïté dans certains cas. Prendre en compte des *shape context* comme dans [DMSD06] ou une distribution locale des points d'intérêt comme dans [Sch99] pourrait aussi aider.

Chapitre 6

Une méthode de mise en correspondance sous fort changement de point de vue

Après un état de l'art de la détection de points d'intérêt invariants aux transformations affines en section 6.1, en particulier avec la simulation de point de vue et l'approche ASIFT *de Morel et Yu [MY09b], nous expliquons en section 6.2 comment gérer les motifs répétés dans le cas de forts changements de points de vue. Puis, la section 6.3 décrit l'algorithme proposé. Nous améliorons aussi la sélection des images simulées pertinentes et l'étape de re-projection. Les expériences sont présentées en section 6.4. L'algorithme produit permet un gain en robustesse en présence de forts changements de points de vue.*

Comme on l'a vu au chapitre précédent, notre approche permet de résoudre le problème de l'appariement en présence de motifs répétés. Notre méthode reste cependant incapable de gérer des situations où le changement d'aspect est plus important. L'invariance des descripteurs n'est alors pas suffisante. Dans ce chapitre, nous traitons de méthodes qui permettent d'augmenter celle-ci en supposant le voisinage des points d'intérêt localement plan.

Il est possible d'atteindre une invariance aux transformations affines par trois types d'approches :

- les méthodes par normalisation ou rectification. Il s'agit des détecteurs invariants aux transformations affines, qui permettent d'extraire les points et descripteurs uniquement à l'aide de l'image 2D.
- les méthodes de simulation de point de vue. Toutes les transformations affines possibles et utiles pour réaliser la mise en correspondance sont effectuées, soit lors d'une étape préalable d'apprentissage, soit en ligne en les combinant avec les propriétés d'invariance d'un détecteur de points d'intérêt plus simple.
- En suivi séquentiel, la connaissance de la pose courante permet de doter les points détectés d'une normale permettant la rectification de la texture au voisinage du point.

6.1 Détection de points d'intérêt et invariance affine

Dans cette section, nous réalisons un état de l'art de différentes méthodes qui permettent d'obtenir des caractéristiques invariantes aux transformations affines. D'abord par normalisation à partir uniquement de l'information 2D d'une image en section 6.1.1, puis par simulation de points de vue en section 6.1.2, et ensuite par rectification en utilisant l'orientation relative des zones observées par rapport à la caméra en section 6.1.3. Nous présentons enfin l'approche ASIFT de Morel et

Yu [MY09b] en section 6.1.4. Cette approche combine normalisation et simulation de point de vue et permet de réaliser l'appariement en présence de forts changements de points de vue.

6.1.1 Points d'intérêt invariants aux transformations affines

Mikolajczyk et al. [MTS+06] ont présenté un article réalisant un état de l'art des détecteurs invariants aux transformations affines, et réalisant une comparaison de leur performance via l'appariement de zones planaires. Les détecteurs comparés sont :

- le détecteur Harris-affine [MS04a],
- le détecteur Hessien-affine [MS04a],
- MSER (*maximally stable extremal region*) [MCUP04],
- un détecteur de région à l'aide de contours EBR [TG04],
- un détecteur de région à l'aide d'extrema d'intensité IBR [TG04],
- un détecteur de région à l'aide de l'entropie (*salient regions*) [KZB04].

Une caractéristique commune à tous ces détecteurs est que les régions elliptiques détectées le sont indépendamment dans chaque image. Cette étude montre que les détecteurs ayant les meilleurs performances sont les MSER, suivis par les Hessien-affine. Cependant, le détecteur MSER nécessite des contours francs ce qui ne permet pas de l'utiliser dans de nombreux contextes (images trop bruitées, scènes naturelles) et le détecteur Hessien-affine est peu précis : son utilisation est plus indiquée dans un contexte de reconnaissance d'objet que pour de l'analyse de la structure et du mouvement.

6.1.2 Simulation de points de vue

Lepetit et Fua [LF06] proposent une approche qui ne repose pas sur des descripteurs, mais utilisent un classifieur entrainé à reconnaitre les voisinages des points d'intérêt sous différents points de vue. Les points sont détectés à l'aide d'extrema du Laplacien, via une approximation efficace en temps de calcul. Les voisinages des points utilisés pour l'apprentissage sont déformés à l'aide de transformations affines choisies aléatoirement. Elles s'écrivent :

$$A = R_\theta R_\phi^{-1} S R_\phi \tag{6.1}$$

avec R_θ et R_ϕ deux matrices de rotation d'angles θ et ϕ, et S une matrice diagonale de mise à l'échelle de paramètres λ_1 et λ_2. Les valeurs des angles des matrices de rotations sont tirées uniformément dans $[-\pi;\pi]$ et celles des λ dans $[0.6;1.5]$. Dans le cas d'un objet plan, une vue frontale suffit lors de l'apprentissage pour permettre l'appariement pour de forts changements de point de vue. Pour des objets plus complexes, il est nécessaire de disposer d'un modèle 3D et d'images du modèle selon différents points de vue.

L'apprentissage est réalisé au moyen de tests simples sur le voisinage du point d'intérêt que l'on cherche à caractériser. Par exemple, un des tests consiste à comparer l'intensité lumineuse pour deux points dont les positions sont choisies au hasard dans chaque voisinage. Suivant le résultat du test, d'autres tests sont appliqués en cascade, construisant ainsi des arbres aléatoires. Une évolution de cette approche, présentée dans [OCLF09], utilise un classifieur appelé *ferns*, qui combine le résultat du test décrit plus haut pour de nombreuses positions de points dans le voisinage choisies aléatoirement, en supposant l'indépendance des caractéristiques décrites par ces tests. Un classifieur par point est utilisé. Cette dernière méthode montre des performances équivalentes à SIFT et fonctionne en temps réel. Elle nécessite cependant un apprentissage préalable. Chaque point suivi définissant une classe, cela limite aussi le nombre total de points qu'il est possible de suivre, les articles mentionnant 400 points comme une valeur typique dans les scénarii évalués. Il faut noter que les *ferns*

sont confrontées au problème de la sélection d'appariements ambigus. Un travail récent [SOL+10] propose la gestion des hypothèses multiples d'appariements pour l'estimation d'homographie en présence de motifs répétés. Ils utilisent un processus itératif qui alterne estimation des appariements et d'une homographie à partir des appariements courants. Un filtre de Kalman permet de maintenir l'incertitude des points sélectionnés pour la pose courante, ce qui permet de définir des fenêtres de recherche pour lever les ambiguïtés de mise en correspondance.

6.1.3 Rectification via les normales aux patchs locaux

Les travaux de Molton *et al.* [MDR04] et de Wu *et al.* [WCL+08] proposent d'utiliser l'orientation des normales aux voisinages locaux des points, considérés plan, pour améliorer l'invariance aux changements de points de vue.

Chez Molton *et al.* [MDR04], des petites régions planaires de l'image sont rectifiées au moyen d'homographies dans une application de suivi séquentiel de SLAM monoculaire. Dans ce cadre, l'estimation à la volée du mouvement de la caméra et des normales du patch en 3 dimensions sont disponibles. Quand un nouveau point est détecté, la normale au patch est initialisée selon une direction parallèle à l'axe optique courant. Cette direction est ensuite affinée en utilisant le schéma habituel des systèmes SLAM de prédiction et de mise à jour. La prédiction est réalisée en appliquant au patch l'homographie résultant de l'orientation précédente du patch et de la nouvelle position de la caméra. La mise à jour consiste en un recalage du patch sur son observation courante, ce qui permet d'obtenir la correction à appliquer à la normale pour expliquer le mouvement. Il n'est pas nécessaire de générer toutes les rectifications possibles, rendant cette approche réalisable en temps réel. Ils produisent une description plus riche d'une scène 3D qu'avec les points d'intérêt utilisés habituellement.

Wu *et al.* [WCL+08] supposent disposer d'un modèle 3D texturé. Leur objectif est de pouvoir aligner différentes reconstructions partielles d'une même scène. Pour cela, ils introduisent la notion de *viewpoint invariant patch (VIP)*. Il s'agit d'une représentation locale invariante aux changements de point de vue qui permet le recalage de deux reconstructions à partir d'une unique correspondance VIP. Ce point d'intérêt regroupe sa position en 3D, la taille du patch local considéré, un vecteur normal au patch, un vecteur donnant l'orientation principale du patch en 3D, et un descripteur SIFT du patch après rectification. Il s'agit ici d'une première étape de combinaison des approches de normalisation et de rectification.

6.1.4 ASIFT

L'approche ASIFT combine normalisation et simulation de points de vue. La version de Morel et Yu [MY09b] que nous cherchons à améliorer, profite de l'invariance de SIFT pour s'épargner une partie de la simulation de points de vue. Dans ASIFT, l'invariance affine des descripteurs extraits de l'image est obtenue par une analyse de la décomposition en valeur singulière (SVD) qu'une transformation affine A (avec un déterminant strictement positif) peut se décomposer de la manière suivante :

$$A = \lambda R(\psi) \begin{pmatrix} t & 0 \\ 0 & 1 \end{pmatrix} R(\phi) \tag{6.2}$$

avec $\lambda > 0$, $R(\psi)$ et $R(\phi)$ des matrices de rotation, $\phi \in [0, 180^o[$, $t \geqslant 1$.

Ils utilisent la même décomposition des transformations affines que dans l'équation (6.1). Chez Lepetit et Fua [LF06], leur détecteur de points d'intérêt n'est ni invariant aux changements d'échelle ni aux rotations, ils doivent discrétiser l'ensemble complet des paramètres λ, t, ψ, ϕ.

Comme les descripteurs SIFT sont invariants aux changements d'échelle et aux rotations, une collection de descripteurs ASIFT invariants aux transformations affines d'une image I s'obtient en extrayant les descripteurs SIFT des images simulées suivantes :

$$I_{t,\phi} = \begin{pmatrix} t & 0 \\ 0 & 1 \end{pmatrix} R(\phi)(I). \tag{6.3}$$

En effet, le positionnement des points d'intérêt SIFT se veut covariant avec tout changement d'échelle et rotation $\lambda R(\psi)$ appliqué à $I_{t,\phi}$, et le descripteur associé ne change quasiment pas. La discussion présentée dans [MY09b] prouve qu'il est suffisant de discrétiser t et ϕ tels que : $t \in \{1, \sqrt{2}, 2, 2\sqrt{2}, 4\}$ et $\phi = \{0, b/t, \ldots, kb/t\}$ avec $b = 72^o$ et $k = \lfloor t/b \cdot 180^o \rfloor$.

L'étape suivante consiste à mettre en correspondance les points d'intérêt ASIFT ainsi générés entre les deux images I et I'. Morel et Yu proposent une approche utilisant deux échelles. D'abord, les images simulées $I_{t,\phi}$ et $I'_{t',\phi'}$ sont générées à partir des images de taille réduite (d'un facteur 3), ensuite les points d'intérêt SIFT extraits de chaque paire $(I_{t,\phi}, I'_{t',\phi'})$ sont mis en correspondance via l'algorithme habituel de SIFT, c'est-à-dire le rapport des distances euclidiennes entre plus proche et second plus proche voisin est inférieur à un seuil fixé (0.6 pour ASIFT). Les déformations correspondant aux M paires (M typiquement entre 5 et 15) qui mènent au plus grand nombre d'appariements sont utilisées sur les images de pleine résolution I et I', qui donnent des points d'intérêt SIFT qu'on met en correspondance par le critère que l'on vient de mentionner. Remarquons que la plupart des appariements ont pour origine la même paire d'images. Les appariements obtenus sont alors reprojetés dans I et I', pourvu que les appariements déjà détectés soient éloignés d'une distance de plus de $\sqrt{3}$. Cette stratégie est utilisée pour limiter le coût calculatoire et d'éviter les redondances entre points d'intérêt SIFT venants de différentes déformations. Une étape supplémentaire optionnelle consiste à éliminer les appariements aberrants avec RANSAC en imposant la compatibilité avec la contrainte épipolaire. Nous décrivons l'approche de manière synthétique dans l'algorithme 3.

Cette méthode utilise plusieurs paramètres : le nombre de tilts utilisés pour déformer les images, l'habituel seuil NNT de la mise en correspondance des descripteurs SIFT, le nombre de paires d'images déformées pour lesquelles on prend en compte les appariements générés, ainsi que celui de distance minimale de non occupation pour positionner un nouveau point. Comme indiqué plus haut, une des transformations est souvent la source de nombreux appariements. Plutôt que de reprojeter les appariements trouvés guidé uniquement par le nombre de ceux extraits dans chaque paire conservée, on souhaiterai utiliser une méthode assurant une meilleure confiance dans le processus de fusion. Par ailleurs, la distance fixée à moins de $\sqrt{3}$ pixels pour autoriser la reprojection d'un point n'évite pas les amas, qui apportent peu à l'estimation du mouvement au final.

6.1.5 Utilisation d'ASIFT

Hsiao et al. [HCH10] ont proposé un système de reconnaissance d'objet reposant sur ASIFT. Les approches habituelles de *bag of features* éliminent les points ambigus [SZ03a, JDS09]. Au contraire, Hsiao et al. insistent sur le fait que pour obtenir une meilleure reconnaissance d'objet, il faut prendre en compte les ambiguïtés en conservant les points ayant la même apparence. C'est seulement lors du calcul de la pose qu'il sera alors possible de lever l'ambiguïté.

Pour réaliser la mise en correspondance, les descripteurs qui ont été classifiés et quantifiés sont associés avec toutes les positions possibles des points correspondants à ces descripteurs dans les images. Ce n'est pas le cas habituellement où un descripteur correspond à un unique point. Ils ne suppriment ainsi pas les points ambigus de la base d'apprentissage. Pour les distinguer lors de la recherche d'objet finale, ils utilisent une méthode d'évaluation de pose 3D-2D appelée RANSAC à

Algorithme 3 : ASIFT

Entrées : deux images I et I'.
Sorties : Ensemble d'appariements filtrés
début
Phase 1 : basse résolution.
1. Pour chaque image, générer la nouvelle collection d'images $I_{t,\phi}$ et $I'_{t',\phi'}$ (eq. (6.3))
2. Extraire les points d'intérêt SIFT des images générées
3. Mettre en correspondance les points d'intérêt SIFT :
pour chaque paire de l'ensemble réduit de l'étape 1, mettre en correspondance chaque point d'intérêt dans $I_{t,\phi}$ avec son plus proche voisin dans $I'_{t',\phi'}$, pourvu que le rapport de distance entre descripteur plus proche et second plus proche voisin soit inférieur à 0.6
Phase 2 : haute résolution.
1. Pour chaque image, générer la nouvelle collection d'images $I_{t,\phi}$ et $I'_{t',\phi'}$ (eq. (6.3))
uniquement pour un nombre limité de $(t,\phi),(t'\phi')$,
venants des M paires ($M = 5$ dans l'article, dans le code diffusé entre 5 et 15) de déformations donnant le plus de points d'intérêt en basse résolution.
2. Extraire les points d'intérêt SIFT des images générées
3. Mettre en correspondance les points d'intérêt SIFT :
pour chaque paire de l'ensemble réduit de l'étape 1, mettre en correspondance chaque point d'intérêt dans $I_{t,\phi}$ avec son plus proche voisin dans $I'_{t',\phi'}$, pourvu que le rapport de distance entre descripteur plus proche et second plus proche voisin soit inférieur à 0.6
4. Re-projeter les points d'intérêt SIFT de de $I_{t,\phi}$ et $I'_{t',\phi'}$ dans I et I'
re-projeter les points d'intérêt appariés uniquement si il n'y a pas déjà un point positionné à une distance de moins de $\sqrt{3}$ pixels.
5. Retirer les appariements potentiellement erronés :
utiliser une méthode *A Contrario*-RANSAC pour vérifier la consistance avec la géométrie épipolaire uniquement (non mentionné dans [MY09b], mais obligatoire dans l'implantation de [MY09a])
fin

vue contrainte. Cette approche consiste à garder lors de l'apprentissage du modèle 3D la connaissance des points de vues pour lesquels chaque point de l'objet est visible. Cela permet d'initialiser l'échantillon de RANSAC de la façon suivante : une fois la première correspondance 3D-2D choisie au hasard, on se limite à choisir des correspondances parmi les points visibles dans les mêmes vues que le premier point 3D sélectionné. Si ce n'est pas suffisant pour calculer un échantillon minimal, on choisit les suivants au hasard sans remettre en causse les points dont on considère la visibilité contrainte par la vue courante. Cette méthode permet de limiter le nombre d'itérations de RANSAC en évitant de considérer des points qui ne peuvent être visibles une fois que certains sont choisis.

6.2 Ambiguïté perceptuelle et mise en correspondance de points

Pour résoudre le problème de l'ambiguïté perceptuelle, une approche comme ASIFT améliore l'invariance de la représentation, et permet donc d'obtenir une détection d'un même point 3D pour des positions de la caméra plus éloignées. Nous avons cependant vu, qu'en présence de motifs répétés, ceux-ci sont aussi décrits de manière similaire par les descripteurs invariants, bien qu'ils ne correspondent pas au même point 3D en pratique, ce qui fait échec à la plupart des méthodes de mise en correspondance habituelles.

ASIFT utilise le schéma en deux étapes 1) mise en correspondance au plus proche voisin, 2) un filtrage robuste à la RANSAC garantissant une cohérence géométrique. Si les images contiennent de nombreux motifs répétés, l'approche échoue comme montré en section 6.4, dans les figures 6.5 et 6.6.

Nous avons montré au chapitre précédent qu'il est possible d'utiliser une approche en une étape pour remplacer les points 1) et 2) ci-dessus. Nous proposons d'incorporer cet algorithme de mise en correspondance robuste résistant aux motifs répétés dans l'approche ASIFT.

Nous rappelons les notations du chapitre précédent. Soient N_1 points d'intérêt x_i et leur descripteur SIFT associé D_i de l'image I, et N_2 points d'intérêt x'_j et descripteurs D'_j de l'image I'. On souhaite construire un ensemble d'appariements $(x_i, x'_j)_{(i,j)\in S}$ avec S un sous-ensemble de $N_1 \times N_2$, qui est l'ensemble le plus consistant avec une homographie parmi tous les ensembles possibles d'appariements. Nous verrons à la section 6.3 pourquoi nous nous intéressons aux homographies. La consistance de S est mesurée par un Nombre de Fausses Alarmes, comme vu précédemment, via le modèle *A Contrario* présenté au chapitre précédent, section 4.4. On rappelle l'équation 4.41

$$\mathrm{NFA}(S,H) = \big(\min\{N_1,N_2\} - 4\big)\, k! \binom{N_1}{k}\binom{N_2}{k}\binom{k}{4} f_D(\delta_D)^k f_G(\delta_G)^{k-4} \tag{6.4}$$

avec :

- l'homographie H de I vers I' est estimée à partir de 4 paires de points de S,
- k est la cardinalité de S,
- $\delta_D = \max_{(i,j)\in S} \mathsf{dist}(D_i, D'_j)$ avec dist une métrique quelconque entre descripteurs SIFT,
- f_D est la fonction de distribution cumulée de δ_D,
- $\delta_G = \max_{(i,j)\in S} \max\{d(x'_j, Hx_i), d(x_i, H^{-1}x'_j)\}$ avec d la distance Euclidienne entre deux points,
- f_G est la fonction de distribution cumulée de δ_G.

6.3 Algorithme

Nous expliquons ici comment nous modifions l'algorithme ASIFT en incorporant le critère du NFA, qui nous permet de définir la méthode *Improved* ASIFT, I-ASIFT dans la suite. L'idée principale

Données : deux images I et I'.

1. Pour chaque image, générer la nouvelle collection d'images $I_{t,\phi}$ et $I'_{t',\phi'}$ (eq. (6.3))
I-ASIFT - utiliser $t, t' \in \{1, \sqrt{2}, 2\}$ et ϕ, ϕ' comme dans ASIFT (la gamme de t est la même que dans [HCH10], ce qui est suffisant si le changement de point de vue n'est pas trop extrême)
ASIFT - *d'abord en basse résolution, ensuite à pleine résolution uniquement pour un nombre limité de $(t, \phi), (t'\phi')$, comme expliqué en section 6.2.*

2. Extraire les points d'intérêt Sift des images générées

3. Mettre en correspondance les points d'intérêt Sift :
I-ASIFT - pour chaque paire $(I_{t,\phi}, I'_{t',\phi'})$ extraire le groupe d'appariements de points avec le plus petit NFA (eq. (6.4), voir la discussion).
ASIFT - *pour chaque paire de l'ensemble réduit de l'étape 1, mettre en correspondance chaque point d'intérêt $I_{t,\phi}$ avec son plus proche voisin dans $I'_{t',\phi'}$, pourvu que le rapport de distance entre descripteur plus proche et second plus proche voisin soit inférieur à 0.6*

4. Re-projeter les points d'intérêt Sift de de $I_{t,\phi}$ et $I'_{t',\phi'}$ dans I et I' :
I-ASIFT - garder uniquement les groupes avec $\log(\text{NFA}) < -50$, et ensuite les trier suivant leurs NFA croissants. En commençant avec le premier groupe selon cet ordre, reprojeter une paire de points d'intérêt mis en correspondance uniquement lorsque chaque point d'intérêt à insérer n'est pas situé à proximité de points déjà placés. La proximité est définie comme la reprojection dans I (resp. I') du cercle autour du point extrait dans les images simulées, avec un rayon égal à l'échelle de détection (minimum $\simeq 2$ pixels).

ASIFT - *re-projeter les points d'intérêt appariés uniquement si il n'y a pas déjà un point positionné à une distance de moins de $\sqrt{3}$ pixels.*

5. Retirer les appariements potentiellement erronés :
I-ASIFT - utiliser un algorithme *A Contrario*-RANSAC [MS04b] pour vérifier la consistance avec la géométrie épipolaire ou une homographie lorsque l'on considère la mise en correspondance entre deux plans.
ASIFT - *utiliser une méthode* A Contrario-RANSAC *pour vérifier la consistance avec la géométrie épipolaire uniquement (non mentionné dans [MY09b], mais obligatoire dans l'implantation de [MY09a]).*

Sortie : un ensemble d'appariements de points d'intérêt.

FIGURE 6.1 – ASIFT amélioré (**I-ASIFT**) et ASIFT standard.

est de remplacer la mise en correspondance au plus proche voisin entre les images générées par l'algorithme de mise en correspondance du chapitre précédent, i.e. chercher le groupe d'appariements consistants avec une homographie et de NFA le plus petit. La re-projection des points mis en correspondance dans les images initiales est aussi améliorée. Les algorithmes ASIFT standard et I-ASIFT sont décrits en figure 6.1, où les modifications introduites sont précisées. L'algorithme est illustré par un exemple d'exécution en figure 6.2. La figure 6.3 permet de comparer la mise en correspondance par l'approche SIFT habituelle et ASIFT.

Discutons les modifications effectuées. D'abord, nous remplaçons à l'étape 3 le critère du plus proche voisin par la méthode présentée plus haut.

La motivation qui justifie l'utilisation d'une contrainte homographique est la suivante : lorsque l'on simule des transformations affines, on s'attend à ce qu'une partie d'entre elles approximent correctement une partie des homographies reliant les parties planaires de la scène (peut-être même des plans *virtuels*, au sens que les points peuvent être distribués sur un plan qui n'a pas de signification physique). Alors, chaque groupe d'appariements entre les images simulées doit correspondre à des

points situés sur des structures planaires, et peut par conséquence être relié au moyen d'une homographie. Dans l'approche ASIFT standard, le nombre de groupes (i.e. de paires d'images simulées) est limité *a priori* à (au plus) quinze. Dans notre approche, ceci conduirait à sélectionner un nombre limité de morceaux de plans. Au contraire, nous gardons les groupes avec un log(NFA) inférieur à -50. Ceci revient en général à garder entre 5 (scène complètement plane) et $\simeq 70$ groupes (scène multi-planaire). Il n'est pas nécessaire de garder un plus grand nombre de groupes : en effet, les groupes avec le plus grand NFA sont faits de points redondants ou composés de quelques points appariés de façon incorrecte et seraient rejetés par l'étape RANSAC finale.

Pour améliorer la reprojection de l'étape 4, nous proposons d'utiliser le NFA comme critère de qualité d'ajustement. Comme remarqué par Hsiao *et al.* [HCH10], les méthodes de simulation de points de vue donnent de toute façon un grand nombre d'appariements, et une part d'entre eux sont concentrés dans des mêmes zones de tailles limitées. Le NFA permet d'équilibrer entre la taille d'un groupe et sa précision, comme on l'a expliqué précédemment. Il nous paraît plus justifié de favoriser les groupes avec le plus petit NFA que de favoriser systématiquement les grands groupes. Par ailleurs, lorsque que l'on reprojette les points d'intérêt, nous sélectionnons soigneusement les appariements à l'aide de leur échelle de façon à éviter les accumulations dans de petites zones (remarquons que notre critère utilise l'échelle de détection au lieu d'un paramètre fixe, ce qui est plus exigeant que dans ASIFT). Obtenir des appariements distribués uniformément et densément distribués dans la scène 3D est important pour les applications de SFM (comme dans [HCH10]).

Remarquons que les motifs répétés produisent des problèmes spécifiques que RANSAC ne peut pas résoudre, en particulier le cas des motifs distribués le long des droites épipolaires qu'on a vu au chapitre précédent (et comme dans la figure 6.6, ACM+F). En théorie, I-ASIFT peut aussi en souffrir. Cependant, cela nécessite que : 1) un des groupes constitué de points positionnés sur des motifs décalés soit consistant avec une homographie (comme dans le groupe 51 sur la figure 6.2) ; 2) ce groupe soit suffisamment grand et ait un très petit NFA (autrement la plupart des points sont redondants avec des points déjà détectés), et 3) que ces points soient disposés le long des droites épipolaires associées (sinon ils sont retirés par le RANSAC final). Ainsi I-ASIFT est plus robuste à ce phénomène.

6.4 Expériences

Dans cette section, nous comparons l'approche proposée I-ASIFT, notée I-ASIFT+H (resp. I-ASIFT+F) quand l'étape RANSAC finale repose sur une homographie (resp. une matrice fondamentale), avec :

- une mise en correspondance SIFT classique (i.e. plus proche voisin + seuil sur le rapport de distances, entre 0.6 et 0.8 afin de produire les meilleurs résultats possibles), suivi d'un RANSAC d'après [MS04b]. Nous appelons cette approche NNT+F si RANSAC impose la géométrie épipolaire (matrice fondamentale), ou NNT+H pour la contrainte homographique.
- la méthode *A Contrario* du chapitre précédent, qui permet de résoudre l'ambiguïté perceptuelle, est appelée ACM+F ou ACM+H ;
- ASIFT, en utilisant l'implantation fournie par Morel et Yu [MY09a].

Nous utilisons le code de Vedaldi pour SIFT [VF08]. Notre implantation utilise matlab et des fonctions écrites en C appelées par l'interface mex. L'algorithme bien qu'écrit pour bénéficier d'une parallélisation simple (boucle for en parallèle) pour le calcul des points d'intérêt pour chaque image transformée et pour l'extraction des appariements sur chaque couple d'images simulées, reste assez lent. Sur une machine quadri-cœur, l'analyse des correspondances pour un couple d'images 768×576 prend environ 5 minutes.

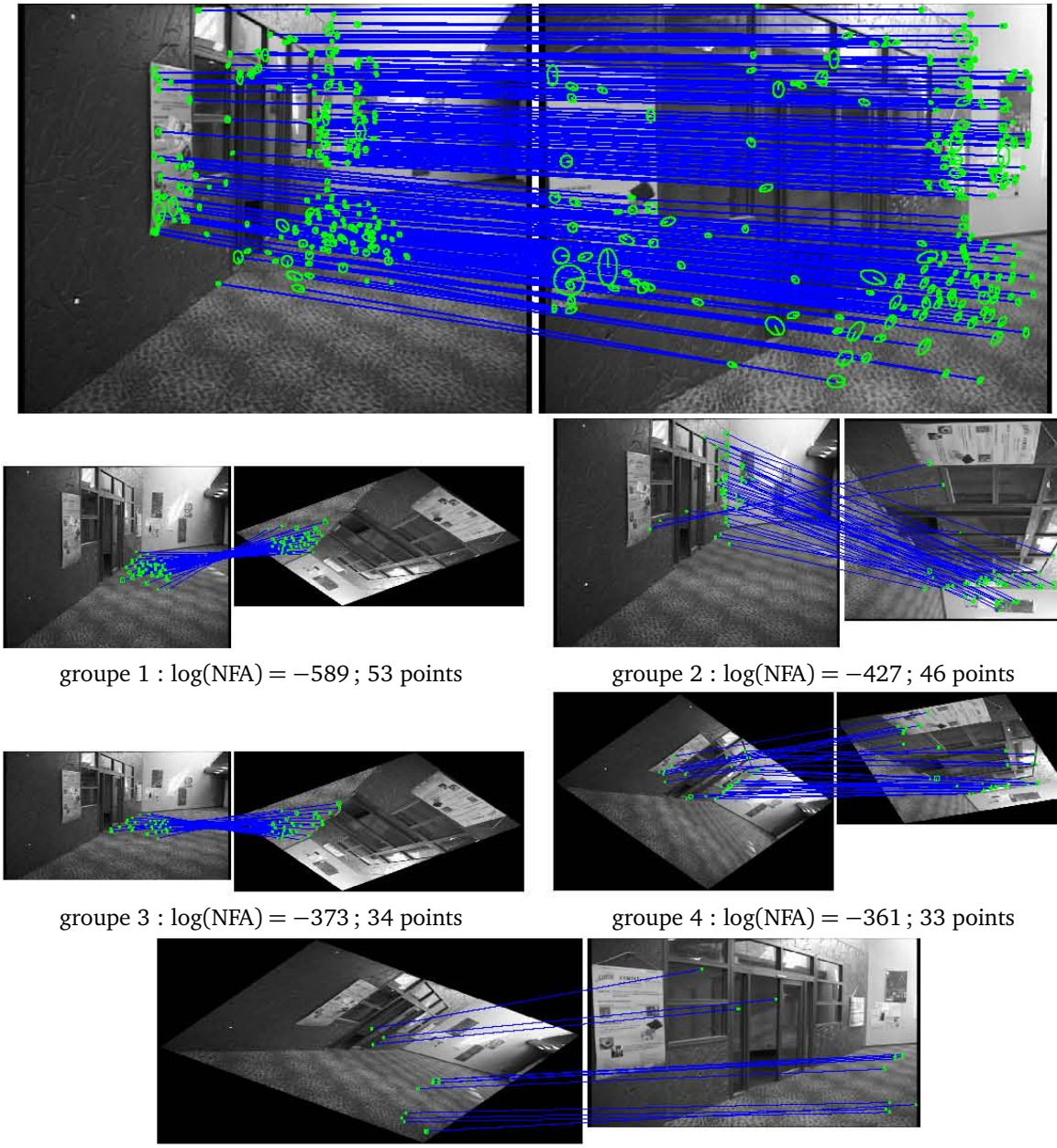

FIGURE 6.2 – *Exemple illustratif détaillé du fonctionnement d'*I-ASIFT. **Haut** : 210 appariements sont trouvés avec I-ASIFT. **En-dessous (groupes 1 à 4)** : les appariements des quatre paires $(I_{t,\phi}, I'_{t',\phi'})$ correspondants aux groupes avec les plus faibles NFA. On voit que ces groupes correspondent aux points situés sur des petits morceaux de plans. Dans cette expérience, 65 groupes de cette sorte sont considérés (avec $\log(\text{NFA}) < -50$). Les 10 derniers groupes produisent seulement 14 appariements. En comparaison, les quatre groupes présentés ici sont à l'origine de 116 appariements. Avec notre schéma, les points venants du groupe 3 (resp. 4) redondants avec ceux du groupe 1 (resp. 2) ne sont pas reprojetés dans I et I'. Remarquons que les points sur le mur ou sur la moquette sont répartis parmi plusieurs groupes. En effet, les homographies induites par le fort changement de point de vue doivent être approchée à l'aide de plusieurs transformations affines. Dans le **groupe 51**, l'algorithme de mise en correspondance est mis en échec par l'ambiguïté perceptuelle : les descripteurs se ressemblent mais les appariements sont compatibles avec une homographie "par hasard". 10 points de ce groupe sont reprojetés mais ils sont tous retirés par l'étape RANSAC finale qui qui impose la compatibilité avec le respect de la contrainte épipolaire.

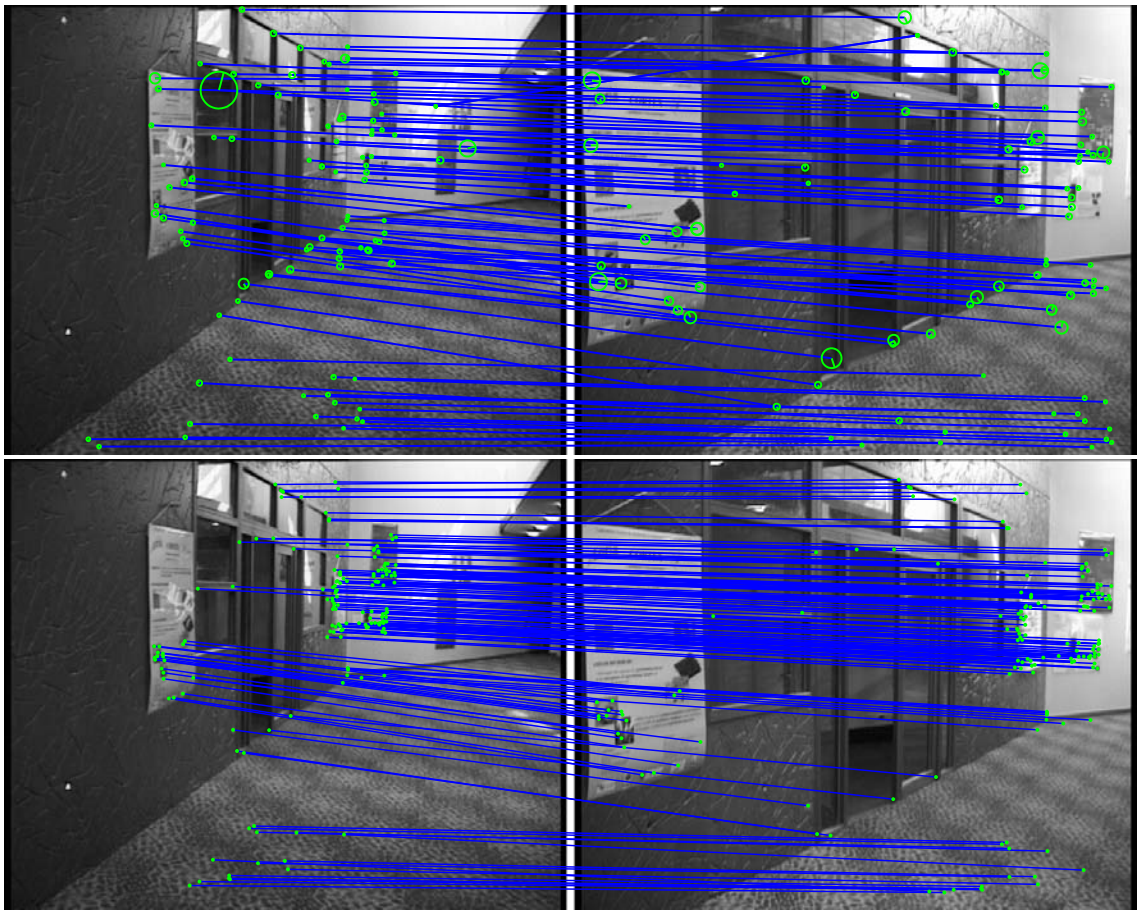

FIGURE 6.3 – *Exemple d'exécution*. **Haut** : mise en correspondance avec SIFT (plus proche voisin + seuil du rapport fixé à 0.8), filtré par la même approche RANSAC que dans ASIFT [MS04b]. **Bas** : ASIFT. 97 appariements sont trouvés avec SIFT, 153 avec ASIFT. Pour réaliser une comparaison équitable, ASIFT a été exécuté avec des images de même résolution que pour I-ASIFT (pas d'étape de sous-échantillonnage). Comme le changement de point de vue est limité, l'approche SIFT standard permet d'extraire une bonne quantité d'appariements. Remarquons que des points de la moquette ne sont pas mis en correspondance correctement. En particulier, quelques appariements aberrants sont visibles au premier plan : les motifs répétés sont a proximité des droites épipolaires associées.

La figure 6.4 est une vérification du fonctionnement d'I-ASIFT sur une paire d'images avec un fort changement de point de vue. NNT et ACM échouent dans ce cas. Les détecteurs Harris/Hessien Affine et MSER donnent moins de 2 appariements (voir [MY09b]). I-ASIFT produit plus d'appariements que ASIFT, et ils sont distribués de façon plus dense, alors qu'ASIFT en accumule dans de petites zones. Ceci est dû principalement au rapport de distance fixé à 0.6 pour ASIFT, qui rejette trop d'appariements candidats pour certaines images simulées. Cependant, utiliser une valeur plus grande génère un plus grand nombre d'outliers, en particulier quand confronté à des fortes déformations perspectives. Le schéma de mise en correspondance plus sophistiqué de I-ASIFT adapte automatiquement la métrique de ressemblance entre descripteurs pour chaque paire d'images simulées.

La figure 6.5 présente une expérience avec presque uniquement des motifs répétés qui peuvent néanmoins être appariés après un examen minutieux. Avec ces données, ASIFT échoue. Plus précisément, il donne quelques appariements corrects, mais ils sont enfouis dans un grand nombre de fausses correspondances, et ne sont pas retrouvés par le RANSAC final. NNT ne produit aucune correspondance. Comme le changement de point de vue n'est pas encore trop fort, ACM+H produit 21 appariements, tous corrects. I-ASIFT+H trouve 44 appariements, tous corrects. Le critère du NFA (eq. (6.4)) permet de mettre en correspondance des points d'intérêt dont les descripteurs ne sont pas plus proches voisins (30% pour I-ASIFT+H).

La figure 6.6 est une autre expérience avec des motifs répétés. A notre connaissance, I-ASIFT est le seul algorithme de mise en correspondance de points d'intérêt qui est capable d'extraire un nombre important d'appariements corrects sur les 3 faces visibles du cube. Cette caractéristique est très recherchée pour les applications d'analyse de la structure et du mouvement. La méthode ACM+H (utilisée à l'étape 3 de I-ASIFT) est capable de gérer les motifs répétés. Remarquons que dans cette expérience nous obtenons *une* des solutions consistantes. Cependant, nous ne pouvons vérifier l'hypothèse selon laquelle le cube est tourné de 90^o. Il s'agit d'un exemple de cas où l'ambiguïté perceptuelle ne peut pas être supprimée à partir de l'information contenue dans les images.

La figure 6.7 montre que I-ASIFT est robuste à de forts changements de points de vue en présence de motifs répétés. Ceci est dû à la stratégie proposée pour la reprojection des points depuis les images simulées, qui sélectionne automatiquement une grand nombre de groupes d'appariements cohérents avec une homographie locale, contrairement à ASIFT où la plupart des appariements viennent des mêmes images simulées. NNT+F et ACM+F ne donnent pas d'ensemble d'appariements corrects.

La figure 6.8 est un cas où ASIFT permet d'obtenir des appariements, mais uniquement sur les deux posters. ACM+H permet d'apparier les points de la moquette, mais ne couvre bien sur que ce plan. ACM+F ne permet pas d'obtenir ici des résultats satisfaisants, au contraire d'I-ASIFT qui permet une mise en correspondance sur les 3 plans visibles.

La figure 6.9 illustre une scène difficile où certains points mal mis en correspondance sont à la fois cohérents avec une homographie locale qui ne correspond pas au mouvement des caméras, et compatibles avec la géométrie épipolaire induite par ce mouvement car étant proches des lignes épipolaires.

6.5 Conclusion

La contribution de ce chapitre est de changer la méthode de mise en correspondance de ASIFT (i.e. la mise en correspondance au plus proche voisin) pour une approche plus sophistiquée qui regroupe les ensembles d'appariements consistant avec une homographie locale. Cette méthode n'est pas limitée aux plus proches voisins et fonctionne en présence de motifs répétés, qu'elle gère. L'algorithme produit est plus robuste que ASIFT, MSER, ou les détecteurs Harris/Hessien Affine aux

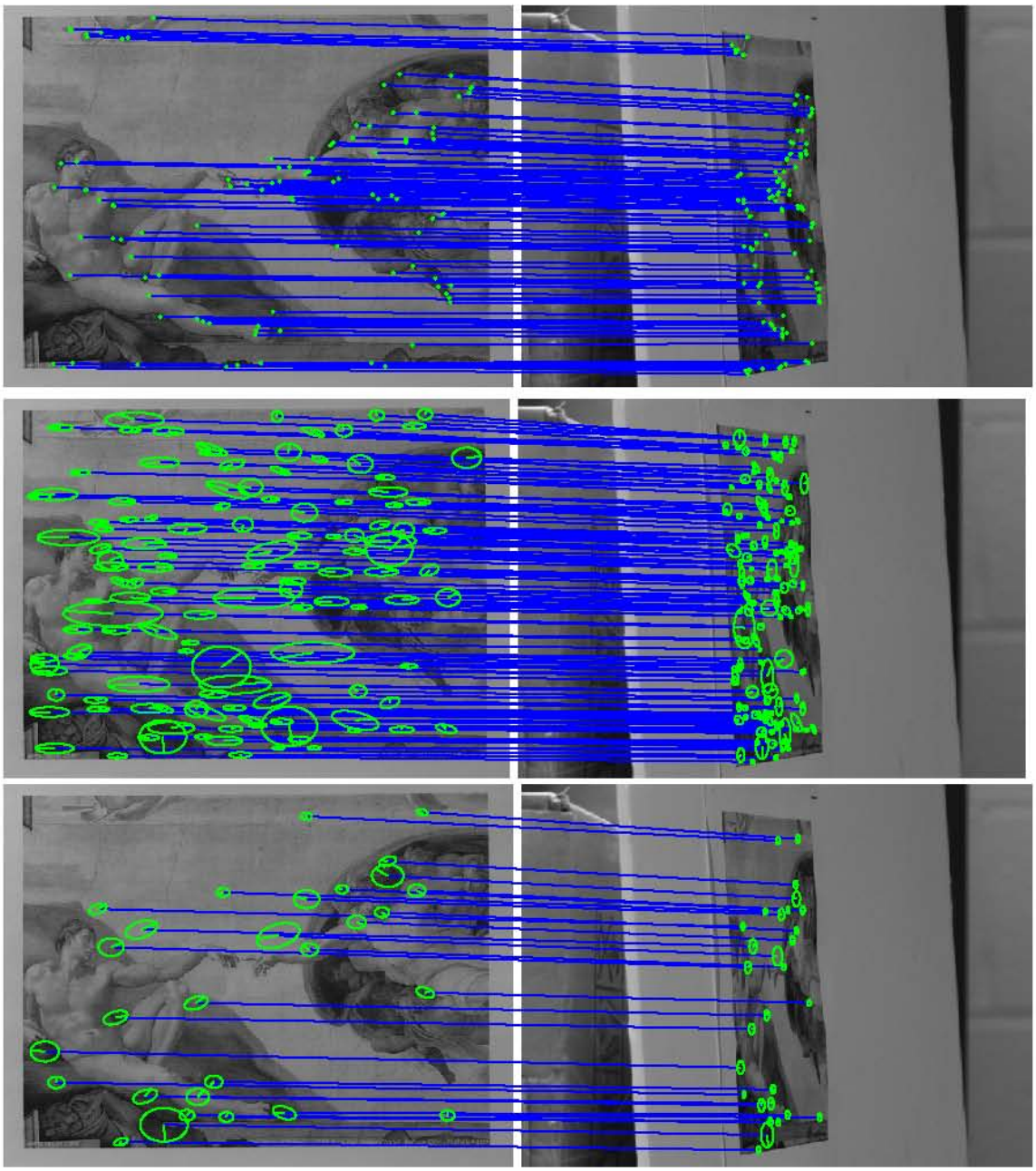

FIGURE 6.4 – *La création d'Adam* (d'après [MY09b, MY09a]). **Haut** : ASIFT. 100 appariements. **Milieu** : I-ASIFT+H. 124 appariements sont obtenus avec t échantillonné dans la gamme $\{1, \sqrt{2}, 2, 2\sqrt{2}, 4\}$ comme dans ASIFT. 49 groupes sont conservés. **Bas** : I-ASIFT+H. La gamme $\{1, \sqrt{2}, 2\}$ (que nous utilisons pour toutes les autres expériences dans ce chapitre avec I-ASIFT) donne quand-même 29 appariements (en 19 groupes).

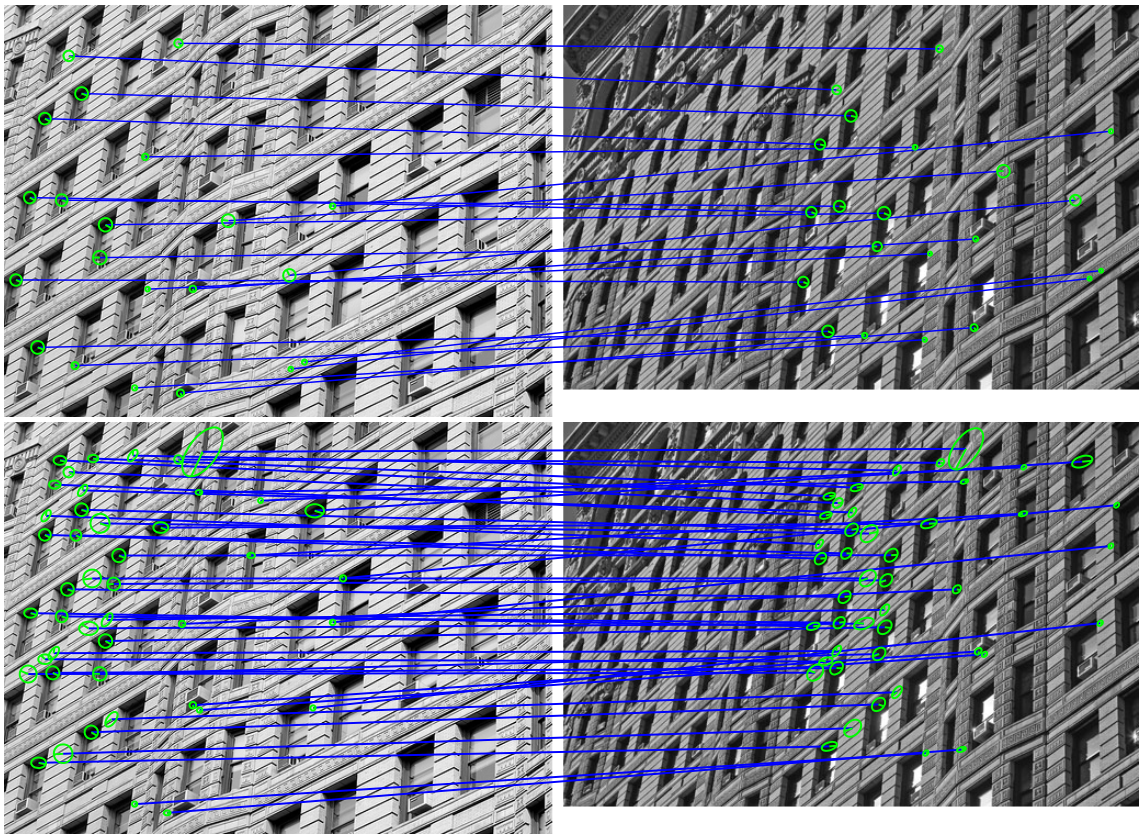

FIGURE 6.5 – *Flatiron Building*. **Haut** : ACM+H trouve 21 appariements, dont 13 ont leurs descripteurs plus proches voisins, 5 second plus proches voisins, les 3 restant étant entre 3ième et 7ième plus proches voisins. **Bas** : I-ASIFT+H trouve 44 appariements, dont 29 ont leurs descripteurs plus proches voisins, 7 second plus proches voisins, les 8 restant étant entre 3ième et 6ième plus proches voisins. 15 groupes sont conservés.

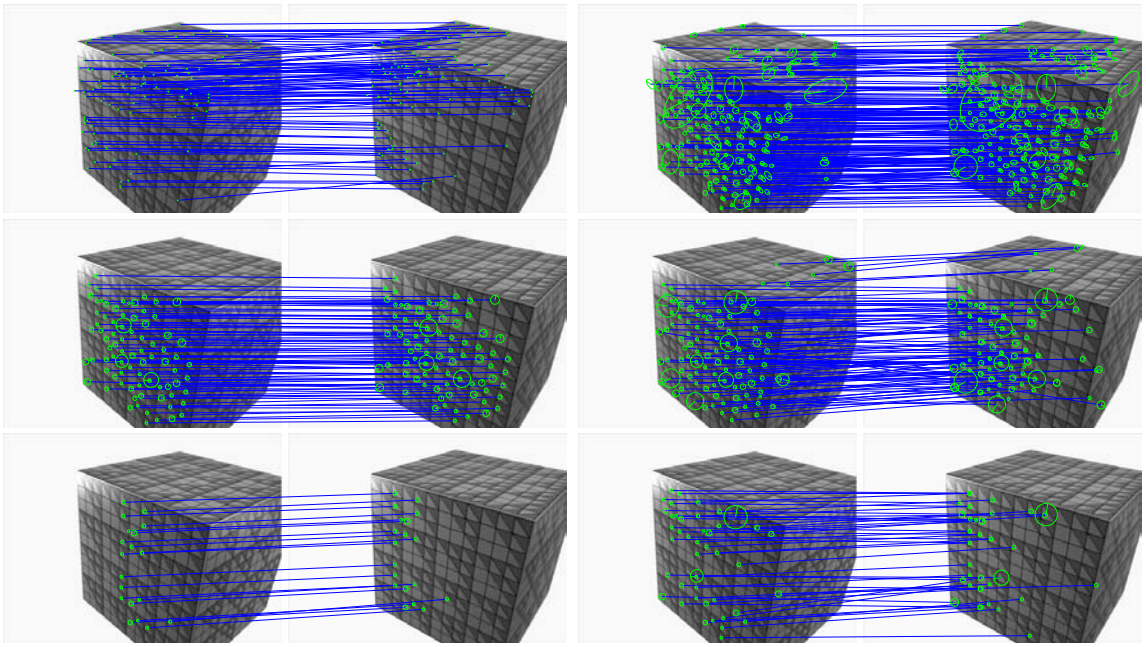

FIGURE 6.6 – *Cube Synthétique* . **En haut à gauche** : ASIFT. 101 appariements, presque la moitié d'entre eux ne sont pas corrects. De nombreux motifs en correspondance sont en effet décalés. **En haut à droite** : I-ASIFT+F. 192 appariements sont obtenus. Un examen minutieux montre qu'ils sont presque tous corrects. Seulement 102 parmi ceux-ci ont leurs descripteurs plus proches voisins, les autres appariements étant entre 2ième et 8ième plus proches voisins. 49 groupes sont conservés. **Au milieu à gauche** : ACM+H. 83 appariements (seulement 40% d'entre eux sont plus proches voisins), les motifs du plan "dominant" sont correctement appariés (à l'aide de la contrainte homographique). **Au milieu à droite** : ACM+F. 102 appariements (55% sont plus proches voisins). Des appariements aberrants sont visibles. C'est ici inévitable avec la méthode habituelle de mise en correspondance et la géométrie entre deux vues : dans cette expérience, de nombreux motifs répétés (corrects pour la contrainte photométrique), sont situés le long des droites épipolaires (et ainsi corrects pour la contrainte géométrique). **En bas à gauche** : NNT+H. 19 appariements, qui correspondent à des motifs décalés. **En bas à droite** : NNT+F. 42 appariements, la plupart aberrants.

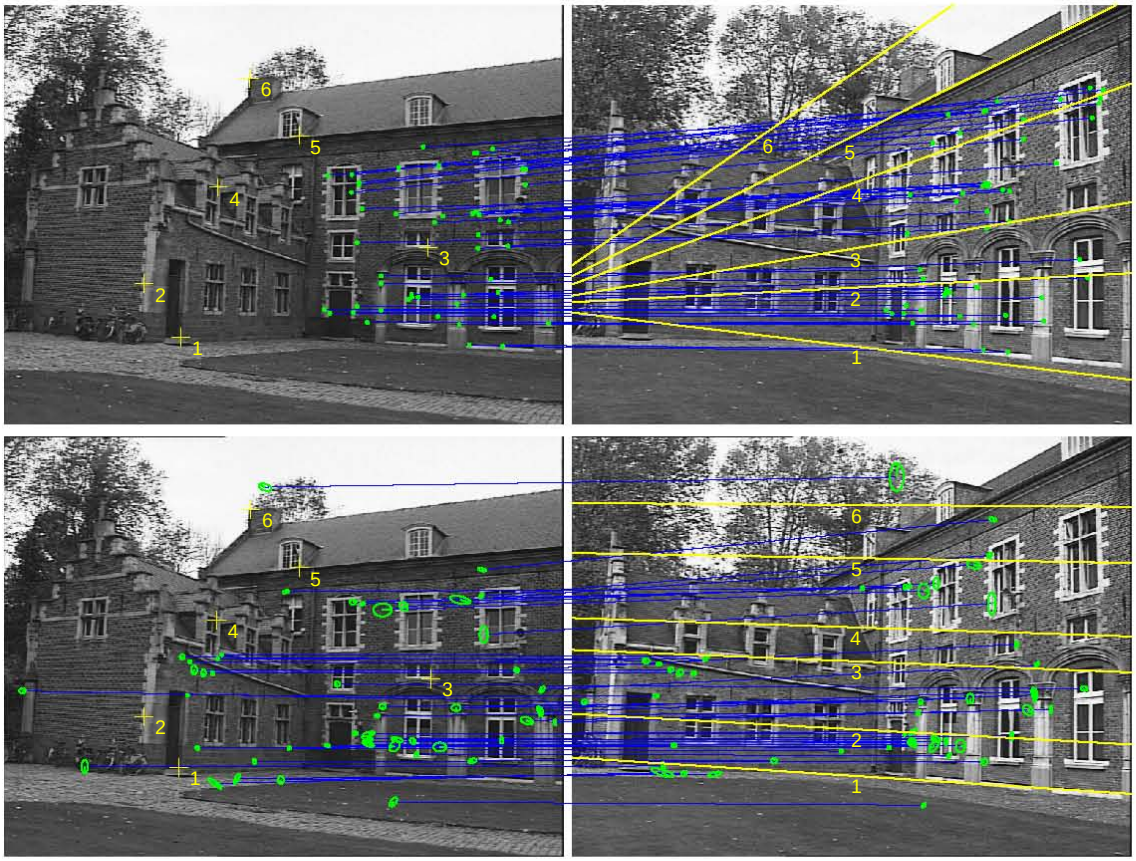

FIGURE 6.7 – *Château de Leuven* : deux images distantes de la séquence de M. Pollefeys. **Haut** : ASIFT. 94 appariements, dont seulement quelques-uns sont sur une façade différente. **Bas** : I-ASIFT+F. 118 appariements (24% ne sont pas plus proches voisins), distribués sur la totalité du bâtiment. Excepté 2, tous les autres sont corrects. La matrice fondamentale est alors estimée sur l'ensemble d'appariements obtenus. Les droites épipolaires (en jaune dans l'image de droite) correspondant à quelques points saisis à la main (dans l'image de gauche) prouve que I-ASIFT permet d'estimer le mouvement de la caméra de façon fiable. Les points associées aux points choisis à la main sont en effet à moins de 1 pixel de leur droite épipolaire correspondante. Au contraire, le mouvement de la caméra n'est pas restitué correctement à partir d'ASIFT. Remarquons que MSER donne 5 appariements, et les détecteurs Harris/Hessien Affine donnent 20-30 appariements principalement entre motifs répétés incorrectement associés. (code d'extraction de `featurespace.org`)

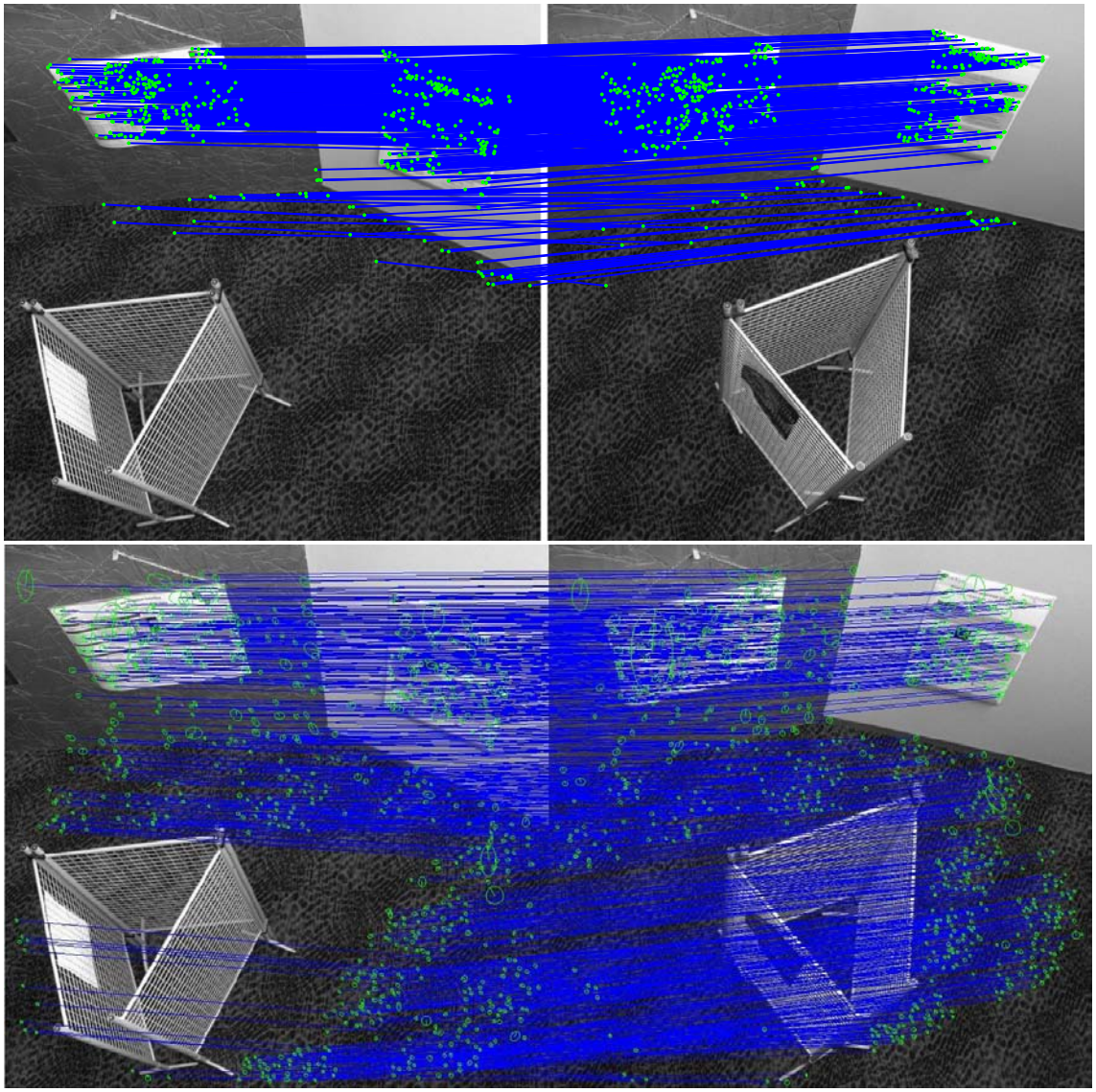

FIGURE 6.8 – *Loria* : En haut, avec ASIFT+F. En bas I-ASIFT+F. Cette fois-ci, les points sont correctement appariés dans toutes les zones de l'image (les deux posters et la moquette).

FIGURE 6.9 – *Bâtiment du second-cycle* : I-ASIFT permet ici de couvrir la scène d'appariements sur une grande profondeur. Il faut noter cependant que quelques points sont incorrectement mis en correspondance, en particulier ceux sur l'arbre à droite et ceux dans le ciel. Cela s'explique car ceux-ci sont cohérents avec une homographie locale, ainsi qu'avec le mouvement global donné par la géométrie épipolaire.

forts changements de points de vue, permettant ainsi une estimation fiable de la pose dans ces cas difficiles.

Conclusion et perspectives

1 Contributions de la thèse

Dans cette thèse, nous avons apporté un ensemble de contributions au problème de la mise en correspondance de points d'intérêt et de l'estimation du mouvement, en particulier en présence de motifs répétés. Pour cela, nous avons présenté deux méthodes, que nous résumons.

1.1 Mise en correspondance *a contrario* sous contraintes géométriques et photométriques

Nous avons étendu le modèle *a contrario* de mise en correspondance de points d'intérêt de Moisan et Stival [MS04b] pour prendre en compte de façon simultanée la ressemblance entre les descripteurs photométriques des points que l'on cherche à apparier [NSB07a, NSB10b, NSB10a]. Pour cela, nous utilisons un modèle *a contrario* qui détermine le nombre de fausses alarmes pour la recherche de mise en correspondance parmi des groupes de candidats 1-N, et nous incluons le score de ressemblance photométrique entre descripteurs circulaires de Rabin et al. [RDG09a]. Il est nécessaire d'équilibrer l'influence des contraintes géométriques et photométriques suivant l'application recherchée, les points les plus ressemblants correspondant souvent au mouvement apparent, mais non au mouvement réel de la caméra. Si cette opération est facile avec l'homographie, elle se révèle plus délicate avec la contrainte épipolaire. Afin d'évaluer le nombre de fausses alarmes, nous avons intégré une approche d'échantillonnage de type RANSAC, en lui incorporant des tests préemptifs pour rejeter sans évaluer leur significativité des groupes suffisamment peu rigides. L'évaluation de la significativité est réalisée en deux passes, une avec géométrie et photométrie mêlée, puis une en privilégiant la géométrie, ce qui permet d'améliorer la significativité du groupe trouvé.

Cela nous a permis de résoudre des situations difficiles en présence d'aliasing perceptuel.

1.2 Un algorithme ASIFT amélioré

Cet algorithme permet de réaliser la mise en correspondance sous de forts changements de point de vue même en présence de motifs répétés [NSB10c]. Par rapport, à l'algorithme ASIFT original, nous imposons que les appariements retenus parmi les couples d'images déformées soient significatifs au sens de notre modèle *a contrario* précédemment introduit, ce qui nous permet de bénéficier de sa puissance en présence d'aliasing perceptuel, sans faire diminuer la qualité d'extraction dans les autres situations. Disposer de groupes significatifs pour un même modèle a contrario nous permet en outre de disposer d'un ordre pour réaliser l'opération de fusion des appariements extraits dans chaque couple d'images déformées dans le couple initial. Nous n'utilisons pas d'étape d'accélération par analyse des images à une échelle réduite, et nous explorons aussi moins de transformations des images. Si l'absence de l'étape à basse résolution rend notre méthode coûteuse en temps de calcul, nous n'avons observé qu'un gain marginal à explorer des déformations supplémentaires.

2 Perspectives

Nos expérimentations nous ont permis de mettre en évidence des difficultés qui nous semblent être des problèmes ouverts pour les approches d'analyse de la structure et du mouvement. Les plus importants sont pour nous :

- Progresser dans les mesures de ressemblance photométrique, pour que deux objets semblables puissent être identifiés comme tels même dans des situations difficiles,
- Il est possible d'utiliser I-ASIFT dans une application d'estimation de mouvements à partir de vues éloignées, par exemple pour la construction de cartes 3D non structurées. Nous bénéficierons ainsi d'une meilleure précision de reconstruction grâce à la plus faible incertitude de triangulation donnée par des vues plus écartées. Pour valider ceci, il est nécessaire de le confronter à des mesures de type "vérité-terrain", et de vérifier que les points détectés par SIFT dans des vues simulées n'ont pas leurs positions biaisées.
- La prise en compte des motifs répétés pour améliorer la recherche d'appariements permet d'obtenir une couverture plus dense de certaines scènes par les points d'intérêt mis en correspondance. Lors d'une phase ultérieure de reconstruction, il est nécessaire de lever les ambiguïtés associées à la présence d'objets identiques à différents endroits de la scène pour les reconstruire séparément les uns des autres. Une piste est ouverte par Roberts et al. [RSSS11], qui combinent des contraintes d'apparence (les points d'intérêt détectés autour des zones répétées sont-ils aussi présents dans les autres vues ? Si oui, il s'agit de la même occurrence de l'objet, sinon d'une autre.) et séquentielles (horodatages des vues) pour guider la reconstruction.
- En particulier, pour ASIFT, comment éviter le calcul de toutes les simulations ? Est-il possible d'apprendre les déformations qui fourniraient les meilleures détections affines ? Dans un contexte de suivi où l'on dispose des directions des normales aux plans observés, c'est une extension raisonnable. Cela permettrait aussi de limiter le temps de calcul lié à l'exploration de toutes les déformations possibles.

Annexes

Annexe A

La géométrie projective pour l'analyse de la structure et du mouvement

A.1 Techniques de calcul de la matrice fondamentale

Soit un point assimilé à ses coordonnées pixel $\mathbf{m}_i = \begin{bmatrix} u_i \\ v_i \end{bmatrix}$ dans la première image correspondant au point $\mathbf{m}'_i = \begin{bmatrix} u'_i \\ v'_i \end{bmatrix}$ dans la deuxième image. Ils doivent satisfaire la contrainte épipolaire ${}^t\mathbf{m}'_i\mathbf{F}\mathbf{m}_i = 0$. On peut réécrire cette équation comme un système linéaire à 9 inconnues :

$$ {}^t\mathbf{u}_i\mathbf{f} = 0 $$

avec

$$ \mathbf{u}_i = {}^t\left[u_iu'_i, v_iu'_i, u'_i, u_iv'_i, v_iv'_i, v'_i, u_i, v_i, 1\right] $$
$$ \mathbf{f} = {}^t\left[F_{11}, F_{12}, F_{13}, F_{21}, F_{22}, F_{23}, F_{31}, F_{32}, F_{33}\right] $$

Si on dispose de n appariements, on obtient le système linéaire suivant :

$$ \mathbf{U}_n\mathbf{f} = \mathbf{0}, $$

avec

$$ \mathbf{U}_n = {}^t\left[\mathbf{u}_1, \cdots, \mathbf{u}_n\right] $$

A.1.1 Méthode des huit points

Proposé par [LH81], cet algorithme permet de déterminer la matrice fondamentale avec 8 points. L'équation $(x'\ y'\ 1)\ F\ {}^t(x\ y\ 1) = 0$ peut être réécrite comme une équation linéaire dépendant des 9 coefficients de F, vus comme un vecteur à 9 dimensions. La matrice F 3×3, est définie à un coefficient d'échelle près, 8 points sont donc suffisants pour déterminer une solution. [HZ00] a montré que la normalisation des coordonnées améliore de manière significative le conditionnement de la matrice $\mathbf{U}_n$. Zhang [Zha98] a écrit un article faisant un point complet sur les difficultés de son estimation.

Cependant, la matrice obtenue n'est pas de rang 2 dans le cas général : il est nécessaire de décomposer F en valeurs singulières et de remplacer la plus petite valeur propre par 0, pour forcer le rang de F.

A.1.2 Méthode des sept points

Bien qu'elle ait l'avantage de produire une solution unique, la méthode précédente est souvent critiquée ; en effet, elle n'impose le rang qu'à posteriori et surcontraint la solution en utilisant 8 points pour 7 degrés de liberté. Avec 7 points et ne nécessitant pas de normalisation, on lui préfère souvent le procédé suivant [TM93] :

Il est possible de déterminer F, en évitant le problème de rang, en considérant 7 appariements et en forçant la contrainte de rang à 2. Le système linéaire 7×9 obtenu avec 7 points de correspondances définit dans le cas général un espace de solutions de dimension 2. En prenant (F_1, F_2) une base de cet espace, la matrice fondamentale est obtenue comme solution de $F = \alpha F_1 + (1 - \alpha)F_2$ avec $\det F = 0$. Cette condition nécessite de déterminer les racines d'un polynôme réel de degré 3, qui a 1 ou 3 solutions réelles. Les matrices fondamentales obtenues par cette méthode définissent une géométrie épipolaire dont toutes les lignes épipolaires se coupent en un unique point, l'épipôle, le rang de F étant contraint à 2 dès l'étape de résolution.

Bibliographie

[AFE+09] W. Aguilar, Y. Frauel, F. Escolano, M. E. Martinez-Perez, A. Espinosa-Romero, and M. A. Lozano. A robust Graph Transformation Matching for non-rigid registration. *Image and Vision Computing*, 27 :897–910, 2009. 48

[AT02] M. E. Antone and S. J. Teller. Scalable extrinsic calibration of omni-directional image networks. *International Journal of Computer Vision*, 49(2-3) :143–174, 2002. 50, 59

[Aya89] N. Ayache. *Vision Stéréoscopique et Perception Multisensorielle : Application à la robotique mobile*. Inter-Editions (MASSON), 1989. 314 pages, in French. 9

[BL07] M. Brown and D. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1) :59–73, 2007. 19

[BTG06] H. Bay, T. Tuytelaars, and L. Van Gool. SURF : Speeded Up Robust Features. In *Proc. of the European Conference on Computer Vision (ECCV)*, volume I of *LNCS*, pages 404–417, 2006. 1

[Cap05] D.P. Capel. An effective bail-out test for ransac consensus scoring. In *Proc. of the British Machine Vision Conference (BMVC)*, 2005. 33

[CFR99] A. Censi, A. Fusiello, and V. Roberto. Image stabilization by features tracking. In *Proc. of International Conference on Image Analysis and Processing*, pages 665–667, 1999. 19

[CLM+08] F. Cao, J.L. Lisani, J.-M. Morel, P. Musé, and F. Sur. *A theory of shape identification*. Number 1948 in Lecture Notes in Mathematics. Springer, 2008. 2, 54

[CM02] O. Chum and J. Matas. Randomized RANSAC with $T_{d,d}$ test. In *Proc. of the British Machine Vision Conference (BMVC)*, September 2002. 33, 55

[CM05] O. Chum and J. Matas. Matching with PROSAC - Progressive Sample Consensus. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 220– 226, 2005. 30, 32, 33, 34

[CM08] O. Chum and J. Matas. Optimal randomized ransac. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30 :1472–1482, 2008. 33

[CMK03] O. Chum, J. Matas, and J. Kittler. Locally Optimized RANSAC. In *Deutsche Arbeitsgemeinschaft für Mustererkennung (DAGM) Symposium*, 2003. 33

[DA06] J. Domke and Y. Aloimonos. A Probabilistic Notion of Correspondence and the Epipolar Constraint. In *Symposium on 3D Data Processing, Visualization and Transmission (3D'PVT)*, 2006. 49

[DHS00] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000. 7

[DMM00] A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful Alignments. *International Journal of Computer Vision*, 40(1) :7–23, 2000. 2, 7, 9, 10, 11, 12, 18

[DMM08] A. Desolneux, L. Moisan, and J.-M. Morel. *From Gestalt theory to image analysis : a probabilistic approach*. Interdisciplinary applied mathematics. Springer, 2008. 2, 54

[DMSD06] H. Deng, E. N. Mortensen, L. Shapiro, and T. G. Dietterich. Reinforcement Matching Using Region Context. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2006. 47, 89

[DSTT03] F. Dellaert, S.M. Seitz, C. Thorpe, and S. Thrun. EM, MCMC, and chain flipping for structure from motion with unknown correspondence. *Machine Learning*, 50(1-2), 2003. 49, 50

[FB81] M. Fischler and R. Bolles. Random Sample Consensus : A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6) :381–395, 1981. 2, 30, 31

[FF87] M. A. Fischler and O. Firschein. *Intelligence : The Eye, the Brain and the Computer*. Addison-Wesley, 1987. 10

[FLP01] O. Faugeras, Q.-T. Luong, and T. Papadopoulou. *The Geometry of Multiple Images*. MIT Press, 2001. 2, 30

[FP02] D. A. Forsyth and J. Ponce. *Computer Vision : A Modern Approach*. Prentice Hall Professional Technical Reference, 2002. 8, 10

[GBI09] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2009. 29

[GBN09] P.F. Georgel, A. Bartoli, and N. Navab. Simultaneous in-plane motion estimation and point matching using geometric cues only. In *Motion and Video Computing, 2009. WMVC '09. Workshop on*, pages 1 –7, dec. 2009. 51

[GR96] S. Gold and A. Rangarajan. A Graduated Assignment Algorithm for Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(4) :377–388, april 1996. 47, 48

[GRL+98] Steven Gold, Anand Rangarajan, Chien-Ping Lu, Suguna Pappu, and Eric Mjolsness. New algorithms for 2d and 3d point matching : pose estimation and correspondence. *Pattern Recognition*, 31(8) :1019 – 1031, 1998. 48

[GS06] L. Goshen and I. Shimshoni. Balanced Exploration and Exploitation Model Search for Efficient Epipolar Geometry Estimation. In *Proc. of the European Conference on Computer Vision (ECCV)*, volume 3952, pages 151–164, 2006. 33

[HCH10] E. Hsiao, A. Collet, and M. Hebert. Making specific features less discriminative to improve point-based 3D object recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA, 2010. 46, 94, 97, 98

[Hou59] P. V. Hough. Machine analysis of bubble chamber pictures. In *International Conference on High Energy Accelerators and Instrumentation*, 1959. 31

[HS88] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Proc. of 4th Alvey Conference*, pages 147–151, Cambridge, août 1988. 21

[Hub81] P. J. Huber. *Robust Statistics*. Wiley, 1981. 31

[HZ00] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN : 0521623049, 2000. 2, 18, 30, 51, 64, 81, 113

[Inc09] Cloudburst Research Inc. Autostitch iphone, 2009. http://www.cloudburstresearch.com/autostitch/autostitch.html. 1

[JDS09] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *Conference on Computer Vision & Pattern Recognition*, jun 2009. 94

[Koe84] J. J. Koenderink. The structure of images. *Biological Cybernetics*, 50 :363–370, 1984. 22

[Kol04] Kolor. Autopano, 2004. http://www.kolor.com/. 1

[KvdG80] J. Krol and W. van der Grind. The double nail illusion : experiments on binocular vision with nails, needles, and pins. *Perception*, 9 :651–659, 1980. 87

[KZB04] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 345–457, 2004. 92

[LBC+07] S. Lehmann, A. P. Bradley, I. Vaughan L. Clarkson, J. Williams, and P. J. Kootsookos. Correspondence-free determination of the affine fundamental matrix. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1) :82–97, 2007. 50

[LF06] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9) :1465–1479, 2006. 62, 92, 93

[LH81] H. Longuet-Higgins. A computer program for reconstructing a scene from two projections. *Nature*, 293 :133–135, septembre 1981. 19, 53, 113

[Lin98] T. Lindeberg. Feature Detection with Automatic Scale Selection. *International Journal of Computer Vision*, 30(2) :79–116, 1998. 22

[LO06] H. Ling and K. Okada. Diffusion distance for histogram comparison. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 246–253, 2006. 27

[LO07] H. Ling and K. Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5) :840–853, 2007. 27, 28, 75

[Low04] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004. 1, 21, 24, 26, 27, 37

[MCUP04] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10) :761–767, 2004. 92

[MDR04] N. D. Molton, A. J. Davison, and I. D. Reid. Locally planar patch features for real-time structure from motion. In *Proc. of the British Machine Vision Conference (BMVC)*, 2004. 21, 93

[MGD07] A. Makadia, C. Geyer, and K. Daniilidis. Correspondence-free structure from motion. *International Journal of Computer Vision*, 75(3) :311–327, 2007. 51

[Moi03] L. Moisan. Modèles continus, numériques et statistiques pour l'analyse d'images. Mémoire d'habilitation à diriger des recherches, Université Paris 11, 2003. 9

[Mor77] H. Moravec. Towards automatic visual obstacle avoidance. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, page 584, August 1977. 21

[Mor80] H. Moravec. *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover*. PhD thesis, September 1980. 22

[MP07] P. Moreels and P. Perona. Evaluation of Features Detectors and Descriptors based on 3D Objects. *International Journal of Computer Vision*, 73(3) :263–284, 2007. 3, 23

[MS04a] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1) :63–86, 2004. 23, 62, 92

[MS04b] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *International Journal of Computer Vision*, 57(3) :201–218, 2004. 2, 4, 7, 13, 14, 15, 16, 17, 18, 32, 34, 37, 38, 42, 51, 52, 54, 58, 61, 64, 70, 75, 97, 98, 100, 109

[MSC+06] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.-M. Morel. An a contrario decision method for shape element recognition. *International Journal of Computer Vision*, 69(3) :295–315, 2006. 55

[MTS+06] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2) :43–72, 2006. 21, 23, 62, 92

[MY09a] J.-M. Morel and G. Yu. ASIFT. IPOL Workshop, 2009. http://www.ipol.im/pub/algo/my_affine_sift/ Consulted 4.26.2010. 95, 97, 98, 102

[MY09b] J.-M. Morel and G. Yu. ASIFT : A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2) :438–469, 2009. 3, 4, 29, 62, 91, 92, 93, 94, 95, 97, 101, 102

[Nis05] D. Nistér. Preemptive RANSAC for Live Structure and Motion Estimation. *Machine Vision and Applications*, 16(5) :321–329, December 2005. 33, 55

[NJD09] K. Ni, H. Jin, and F. Dellaert. Groupsac : Efficient consensus in the presence of groupings. In *Proc. of the International Conference of Computer Vision (ICCV)*, 2009. 33

[NSB07a] N. Noury, F. Sur, and M.-O. Berger. Fundamental matrix estimation without prior match. In *Proc. of the IEEE International Conference on Image Processing (ICIP)*, pages 513–516, 2007. 4, 109

[NSB07b] N. Noury, F. Sur, and M.-O. Berger. Modèles statistiques pour l'estimation de la matrice fondamentale. In *ORASIS, congrès francophone des jeunes chercheurs en vision par ordinateur*, 2007. 4, 37, 55

[NSB10a] N. Noury, F. Sur, and M.-O. Berger. Determining point correspondences between two views under geometric constraint and photometric consistency. Technical report, INRIA research report RR-7246, 2010. 4, 29, 109

[NSB10b] N. Noury, F. Sur, and M.-O. Berger. Modèle a contrario pour la mise en correspondance robuste sous contraintes épipolaires et photométriques. In *Actes de la conférence Reconnaissance de Formes et Intelligence Artificielle (RFIA)*, 2010. 4, 55, 109

[NSB10c] N. Noury, F. Sur, and Marie-Odile Berger. How to overcome perceptual aliasing in asift ? In *Proc. of the International Symposium on Visual Computing (ISVC), Las Vegas, NV, USA. 2010*, 2010. 4, 109

[OCLF09] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2009. 92

[PGV+04] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual Modeling with a Hand-Held Camera. *International Journal of Computer Vision*, 59(3) :207– 232, 09 2004. 19

[RC96] S. Roy and I. J. Cox. Motion without structure. volume 1, pages 728 –734 vol.1, aug. 1996. 50

[RDG08a] J. Rabin, J. Delon, and Y. Gousseau. A contrario matching of SIFT-like descriptors. In *Proc. of the International Conference on Pattern Recognition (ICPR)*, 2008. 27

[RDG08b] J. Rabin, J. Delon, and Y. Gousseau. Circular Earth Mover's Distance for the comparison of local features. In *Proc. of the International Conference on Pattern Recognition (ICPR)*, 2008. 27

[RDG09a] J. Rabin, J. Delon, and Y. Gousseau. A statistical approach to the matching of local features. *SIAM Journal of Imaging Sciences*, 2(3) :931–958, 2009. 27, 28, 45, 46, 52, 55, 56, 59, 75, 109

[RDG09b] J. Rabin, J. Delon, and Y. Gousseau. Transportation distances on the circle. Preprint arXiv :0906.5499, 2009. 28

[RDGM10] J. Rabin, J. Delon, Y. Gousseau, and L. Moisan. Mac-ransac : a robust algorithm for the recognition of multiple objects. In *Symposium on 3D Data Processing, Visualization and Transmission (3D'PVT)*, 2010. 7, 18, 32, 52, 62

[RFP08] R. Raguram, J.-M. Frahm, and M. Pollefeys. A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 500–513, Berlin, Heidelberg, 2008. Springer-Verlag. 33

[RHZ76] A. Rosenfeld, R. A. Hummel, and S. W. Zucker. Scene labeling by relaxation operations. *Systems, Man and Cybernetics, IEEE Transactions on*, 6(6) :420 –433, june 1976. 47

[RL87] P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987. 31

[RSSS11] R. Roberts, S. N. Sinha, R. Szeliski, and D. Steedly. Structure from motion for scenes with large duplicate structures. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, June 2011. 110

[RTG00] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2) :99–121, 2000. 27, 28

[Sap90] G. Saporta. *Probabilités, Analyse des Données et Statistique*. Technip, 1990. 8

[Sch96] Cordelia Schmid. *Appariements d'images par invariants locaux de niveaux de gris*. PhD thesis, Institut National Polytechnique de Grenoble, 1996. 1

[Sch99] C. Schmid. A structured probabilistic model for recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 2485, Los Alamitos, CA, USA, 1999. IEEE Computer Society. 89

[Sin64] R. Sinkhorn. A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. *Ann, Math. Statistics*, 35 :876–879, 1964. 48

[SM97] C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 :530–535, 1997. 22, 47

[SM06] R. Subbarao and P. Meer. Beyond ransac : User independent robust regression. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, page 101, Washington, DC, USA, 2006. IEEE Computer Society. 33

[SMJ08] D. Sidibe, P. Montesinos, and S. Janaqi. Matching local invariant features with contextual information : An experimental evaluation. *Electronic Letters on Computer Vision and Image Analysis*, 7(1) :26–39, 2008. 47

[Sna] Noah Snavely. Bundler : Structure from motion (sfm) for unordered image collections. http://phototour.cs.washington.edu/bundler. 1, 27

[SOL+10] E. Serradell, M. Ozuysal, V. Lepetit, P. Fua, and F. Moreno-Noguer. Combining Geometric and Appearance Priors for Robust Homography Estimation. In *Proceedings of the European Conference on Computer Vision*, 2010. 93

[SS00] G.P. Stein and A. Shashua. Model-based brightness constraints : on direct estimation of structure and motion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(9) :992 –1015, sep. 2000. 51

[SSS07] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 2007. 1

[Ste95] C. V. Stewart. Minpran : A new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10) :925–938, 1995. 14, 32

[SZ00] F. Schaffalitzky and A. Zisserman. Planar grouping for automatic detection of vanishing lines and points. *Image and Vision Computing*, 18(9) :647–658, 2000. 29

[SZ03a] F. Schaffalitzky and A. Zisserman. Automated location matching in movies. *Computer Vision and Image Understanding*, 92 :236–264, 2003. 94

[SZ03b] J. Sivic and A. Zisserman. Video google : A text retrieval approach to object matching in videos. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1470–1477, 2003. 19

[TC02] D. Tell and S. Carlsson. Combining Appereance and Topology for Wide Baseline Matching. In *Proc. of the European Conference on Computer Vision (ECCV)*, number 2350 in LNCS, pages 68–81, 2002. 47

[TFZ99] P. H.S. Torr, A. W. Fitzgibbon, and A. Zisserman. The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences. *International Journal of Computer Vision*, 32(1) :27–44, 1999. 2

[TG04] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *International Journal on Computer Vision*, 59(1) :61–85, 2004. 92

[TM93] P. Torr and D. Murray. Outlier Detection and Motion Segmentation. In P. S. Schenker, editor, *Sensor Fusion VI*, pages 432–443. SPIE volume 2059, 1993. 114

[TM97] P. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*, 24(3) :271–300, 1997. 2, 30

[TM02] B. Tordoff and D. W. Murray. Guided Sampling and Consensus for Motion Estimation. In *Proc. of the European Conference on Computer Vision (ECCV)*, volume 2350, pages 82 – 98, 2002. 32, 33, 34, 35

[TM05] B.J. Tordoff and D.W. Murray. Guided-MLESAC : Faster image transform estimation by using matching priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10) :1523–1535, 2005. 33, 35

[TM08] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors : a survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3) :177–280, 2008. 22

[Tor02] P. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *International Journal of Computer Vision*, 50(1) :35–61, 2002. 30

[TZ00] P. Torr and A. Zisserman. MLESAC : A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78 :138–156, 2000. 30, 32, 34

[VF08] A. Vedaldi and B. Fulkerson. VLFeat : An open and portable library of computer vision algorithms, 2008. http://www.vlfeat.org/ Consulted 4.26.2010. 24, 98

[vic00] vicon. Boujou., 2000. http://www.vicon.com/boujou. 1

[WB91] S.D. Whitehead and D.H. Ballard. Learning to perceive and act by trial and error. *Machine Learning*, 7(1) :45–83, 1991. 29

[WCL+08] Changchang Wu, B. Clipp, Xiaowei Li, J.-M. Frahm, and M. Pollefeys. 3d model matching with viewpoint-invariant patches (vip). In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008. 93

[ZDFL95] Z. Zhang, R. Deriche, O. D. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78(1-2) :87–119, 1995. 47

[Zha98] Z. Zhang. Determining the epipolar geometry and its uncertainty : a review. *International Journal of Computer Vision*, 27(2) :161–195, 1998. 2, 18, 30, 53, 54, 81, 113

[ZK06] W. Zhang and J. Kosecka. Generalized RANSAC Framework for Relaxed Correspondence Problems. In *Symposium on 3D Data Processing, Visualization and Transmission (3D'PVT)*, pages 854–860, 2006. 61

Résumé

L'analyse de la structure et du mouvement permet d'estimer la forme d'objets 3D et la position de la caméra à partir de photos ou de vidéos. Le plus souvent, elle est réalisée au moyen des étapes suivantes : 1) L'extraction de points d'intérêt, 2) La mise en correspondance des points d'intérêt entre les images à l'aide de descripteurs photométriques des voisinages de point, 3) Le filtrage des appariements produits à l'étape précédente afin de ne conserver que ceux compatibles avec une contrainte géométrique fixée, dont on peut alors calculer les paramètres. Cependant, la ressemblance photométrique seule utilisée en deuxième étape ne suffit pas quand plusieurs points ont la même apparence. Ensuite, la dernière étape est effectuée par un algorithme de filtrage robuste, Ransac, qui nécessite de fixer des seuils, ce qui se révèle être une opération délicate.

Le point de départ de ce travail est l'approche A Contrario Ransac de Moisan et Stival, qui permet de s'abstraire des seuils. Ensuite, notre première contribution a consisté en l'élaboration d'un modèle a contrario qui réalise la mise en correspondance à l'aide de critères photométrique et géométrique, ainsi que le filtrage robuste en une seule étape. Cette méthode permet de mettre en correspondance des scènes contenant des motifs répétés, ce qui n'est pas possible par l'approche habituelle. Notre seconde contribution étend ce résultat aux forts changements de point de vue, en améliorant la méthode ASift de Morel et Yu. Elle permet d'obtenir des correspondances plus nombreuses et plus densément réparties, dans des scènes difficiles contenant des motifs répétés observés sous des angles très différents.

Mots-clés : vision par ordinateur, analyse de la structure et du mouvement, méthode A Contrario, forts changements de point de vue.

Abstract

The analysis of structure from motion allows one to estimate the shape of 3D objects and the position of the camera from pictures or videos. It usually follows these three steps : 1) Extracting interest points, 2) Matching interest points using photometric descriptors computed on point neighborhoods, 3) Filtering previous matches so as to retain only those compatible with a geometric constraint, whose parameters can then be computed. However, for the second step, the photometric criterion is not enough on its own when several points are alike. As for the third step, it uses the Ransac robust filtering scheme, which requires setting thresholds, and that can be a difficult task.

This work is based on Moisan and Stival's A Contrario Ransac approach, which allows one to set thresholds automatically. After assessing that method, the first contribution was the elaboration an a contrario model, which simultaneously achieves robust filtering and matching through both geometric and photometric criteria. That method allows one to match scenes with repeated patterns, which is impossible with the usual approach. The second contribution extended that result to strong viewpoint changes, improving Morel and Yu's ASift method. The correspondences obtained are both more numerous and more densely distributed, in scenes containing many repeated patterns seen from very different viewpoints.

Keywords : computer vision, structure from motion analysis, A Contrario method, strong viewpoint changes