Compréhension de scènes urbaines par combinaison d'information 2D/3D

Marie-Anne Bauda

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (INP Toulouse)

Discipline ou spécialité :

Image, Information et Hypermédia

Présentée et soutenue par :

Mme MARIE-ANNE BAUDA

le lundi 13 juin 2016

Titre :

COMPREHENSION DE SCENES URBAINES PAR COMBINAISON D'INFORMATION 2D/3D

Ecole doctorale :

Mathématiques, Informatique, Télécommunications de Toulouse (MITT)

Unité de recherche :

Institut de Recherche en Informatique de Toulouse (I.R.I.T.)

Directeur(s) de Thèse :

M. VINCENT CHARVILLAT

MME SYLVIE CHAMBON

Rapporteurs :

M. ENGUERRAN GRANDCHAMP, UNIVERSITE ANTILLES GUYANE

M. EZZEDDINE ZAGROUBA, FACULTE DES SCIENCES DE TUNIS

Membre(s) du jury :

M. WILLIAM PUECH, UNIVERSITE DE MONTPELLIER, Président

Mme SYLVIE CHAMBON, INP TOULOUSE, Membre

M. PIERRE GURDJOS, INP TOULOUSE, Membre

M. VINCENT CHARVILLAT, INP TOULOUSE, Membre

À mes parents,

« La vraie découverte ne consiste pas à chercher de nouveaux paysages
mais à changer de regard. »
Marcel Proust

Remerciements

Je remercie en premier lieu Étienne Lamort de Gail, président de l'entreprise `Imajing`, de m'avoir offert l'opportunité de réaliser cette thèse et sans qui ce projet n'aurait pas eu lieu.

Je remercie mon directeur de thèse Vincent Charvillat de m'avoir accueillie au sein de l'équipe `Vortex`, IRIT/ENSEEIHT ainsi que pour les échanges intéressants et constructifs que nous avons pu avoir.

Un grand merci, à mes encadrants de m'avoir guidée sur le chemin de cette thèse en me permettant de bénéficier de leurs qualités scientifiques et humaines :

- Sylvie Chambon, pour ta constance et ton suivi bienveillant ;
- Pierre Gurdjos, pour nos échanges nécessaires à une « bonne construction » ;
- Mariana Spangenberg, pour ta vision éclairée à l'initialisation d'une « bonne trajectoire ».

Je tiens à remercier l'ensemble des membres du jury, pour leurs retours et questions pertinantes, ainsi que pour l'intérêt porté à mes travaux. Une attention particulière va vers les rapporteurs Ezzedinne Zagrouba et Enguerran Grandchamp pour avoir accepté d'évaluer mon manuscrit ainsi que pour leurs critiques. Je remercie William Puech, d'avoir accepté d'être le président du jury.

Je souhaite également exprimer mes remerciements à l'ensemble des membres de l'équipe `Vortex` et de l'équipe d'`Imajing`, pour la qualité des échanges scientifiques ou non-scientifiques, et pour les bons moments passés ensembles.

De plus, je tiens à exprimer ma plus grande gratitude à mes parents, Patricke Dupré et Alain Bauda, pour m'avoir permis de faire ces longues études. J'ai également une pensée affectueuse à l'égard de ma sœur Claire et de l'ensemble de ma famille (de sang ou de cœur), pour m'avoir soutenue dans les moments difficiles ou d'avoir été tout simplement présents.

Enfin, je remercie tous ceux, proches ou moins proches, qui ont contribué de près ou de loin à la réalisation de ce projet.

Résumé

Titre : Compréhension de scènes urbaines par combinaison d'information 2D/3D
Mots-clés : Segmentation sémantique, super-pixels, multi-vues, mesures de cohérence photométrique, planéité et homographie

Cette thèse traite du problème de segmentation sémantique d'une séquence d'images calibrées acquises dans un environnement urbain. Ce problème consiste, plus précisément, à partitionner chaque image en régions représentant les objets de la scène (façades, routes, etc.). Ainsi, à chaque région est associée une étiquette sémantique. Dans notre approche, l'étiquetage s'opère via des primitives visuelles de niveau intermédiaire appelés super-pixels, lesquels regroupent des pixels similaires au sens de différents critères proposés dans la littérature, qu'ils soient photométriques (s'appuyant sur les couleurs) ou géométriques (limitant la taille des super-pixels formés). Contrairement à l'état de l'art, où les travaux récents traitant le même problème s'appuient en entrée sur une sur-segmentation initiale sans la remettre en cause, notre idée est de proposer, dans un contexte multi-vues, une nouvelle approche de constructeur de super-pixels s'appuyant sur une analyse tridimensionnelle de la scène et, en particulier, de ses structures planes. Pour construire de « meilleurs » super-pixels, une mesure de planéité locale, qui quantifie à quel point la zone traitée de l'image correspond à une surface plane de la scène, est introduite. Cette mesure est évaluée à partir d'une rectification homographique entre deux images proches, induites par un plan candidat au support des points 3D associés à la zone traitée. Nous analysons l'apport de la mesure UQI (*Universal Quality Image*) et montrons qu'elle se compare favorablement aux autres métriques qui ont le potentiel de détecter des structures planes. On introduit ensuite un nouvel algorithme de construction de super-pixels, fondé sur l'algorithme SLIC (*Simple Linear Iterative Clustering*) dont le principe est de regrouper les plus proches voisins au sens d'une distance fusionnant similarités en couleur et en distance, et qui intègre cette mesure de planéité. Ainsi la sur-segmentation obtenue, couplée à la cohérence inter-images provenant de la validation de la contrainte de planéité locale de la scène, permet d'attribuer une étiquette à chaque entité et d'obtenir ainsi une segmentation sémantique qui partitionne l'image en objets plans.

Abstract

Title : Urban Scenes understanding by combining 2D/3D information
Key words : Semantic segmentation, superpixels, multi-view, photo-consistency measure, flatness and homography

This thesis deals with the semantic segmentation problem of a calibrated sequence of images acquired in an urban environment. The problem is, specifically, to partition each image into regions representing the objects in the scene such as facades, roads, etc. Thus, each region is associated with a semantic tag. In our approach, the labelling is done through mid-level visual features called super-pixels, which are groups of similar pixels within the meaning of some criteria proposed in research such as photometric criteria (based on colour) or geometrical criteria thus limiting the size of super-pixel formed. Unlike the state of the art, where recent work addressing the same problem are based on an initial over-segmentation input without calling it into question, our idea is to offer, in a multi-view environment, another super-pixel constructor approach based on a three-dimensional scene analysis and, in particular, an analysis of its planar structures. In order to construct "better" super-pixels, a local flatness measure is introduced which quantifies at which point the zone of the image in question corresponds to a planar surface of the scene. This measure is assessed from the homographic correction between two close images, induced by a candidate plan as support to the 3D points associated with the area concerned. We analyze the contribution of the UQI measure (Universal Image Quality) and demonstrate that it compares favorably with other metrics which have the potential to detect planar structures. Subsequently we introduce a new superpixel construction algorithm based on the SLIC (Simple Linear Iterative Clustering) algorithm whose principle is to group the nearest neighbors in terms of a distance merging similarities in colour and distance, and which includes this local planarity measure. Hence the over-segmentation obtained, coupled with the inter-image coherence as a result of the validation of the local flatness constraint related to the scene, allows assigning a label to each entity and obtaining in this way a semantic segmentation which divides the image into planar objects.

Table des matières

Table des figures

Liste des tableaux

Liste des Algorithmes

Introduction

Appréhender un environnement extérieur complexe à partir de données visuelles (photographies, vidéo etc.) est un problème informatique communément référencé sous le nom de « vision par ordinateur », dont l'intérêt aujourd'hui semble évident, ne serait-ce que pour permettre, par exemple, à l'ordinateur d'interagir avec cet environnement. Du gadget connecté à nos *smartphones* jusqu'à la voiture autonome et intelligente de demain, l'essor des objets *high-tech* dans notre quotidien a ainsi accéléré ce besoin de vision artificielle d'autant plus que, dans cette ère du tout numérique, il est maintenant facile de collecter de grandes bases de données visuelles et de disposer ainsi très rapidement de nombreuses informations redondantes sur l'environnement observé. Ainsi, les recherches dans le domaine de vision par ordinateur se multiplient, les enjeux économiques étant de toute évidence considérables. Aujourd'hui, de nombreux industriels sont directement impliqués dans ces défis scientifiques, tant dans les secteurs de la robotique, de l'automobile, de la cartographie que des réseaux sociaux de demain. C'est le cas de l'entreprise `Imajing`, qui propose un système de cartographie mobile compact (l'`imajbox`) conçu pour collecter des données à grande vitesse, et qui a collaboré au travail de cette thèse au sein d'un dispositif CIFRE (Conventions Industrielles de Formation par la REcherche).

La compréhension d'une scène complexe est une tâche visuelle que l'être humain peut effectuer de manière instantanée et sans effort. La tâche de vision par ordinateur qui consiste à analyser une scène réelle à partir de plusieurs photographies de celle-ci dans le but de reconnaître les différents éléments qui la constituent, en délimitant leurs contours par exemple, est sans aucun doute une des plus ambitieuses, et de ce fait une des plus difficiles. Comme le rappelle Szeliski dans son ouvrage [Szeliski 10], même si la vision par ordinateur a aujourd'hui acquis une grande maturité à la fois scientifique et technique à un point où il lui est maintenant possible de reconstruire des représentations géométriques et photométriques d'une scène complexe d'une façon extrêmement précise, à partir (en général d'un grand nombre) de photographies prises selon des points de vue différents, il est encore quasiment impossible de reconnaître et d'identifier tous ses différents constituants, alors qu'un enfant de jeune âge pourrait sans doute le faire. Naturellement, un être humain a la capacité de comprendre et d'interpréter une scène complexe et ambiguë comme celle de la figure 1, du fait que la perception visuelle chez lui est un sens automatique, fiable et rapide tout en mettant en œuvre des mécanismes très sophistiqués [Shokron 10]. La compréhension de ces mécanismes reste un problème ouvert [Desolneux 02] dans de nombreux domaines de recherche, qu'ils soient d'orientation neurologique ou psychologique. À l'heure actuelle, on peut dire qu'il n'existe pas de consensus chez les chercheurs de la communauté de la vision par ordinateur quant à la date à laquelle cette tâche visuelle pourrait être résolue. Récemment, les techniques dites d'apprentissage profond semblent apporter une réelle contribution [Badrinarayanan 15].

Dans cette thèse, nous nous intéressons à une formulation particulière du problème de compré-

FIGURE 1 – Les reptiles de Escher prenant de la hauteur.

hension automatique d'un environnement, que nous appellerons « segmentation sémantique » : il s'agit d'associer à chaque région d'une image une étiquette sémantique parmi celles décrivant l'ensemble des classes d'objets potentiellement présents dans cet environnement. En règle générale, un découpage de l'image en régions est obtenu via un processus de segmentation, en partitionnant initialement celle-ci en entités homogènes (les régions) telles que chacune d'elles décrive un certain aspect visuel des objets. Il est difficile d'établir une classification exhaustive des méthodes de segmentation, mais, dans la littérature, on distingue au moins trois types d'approches. D'une part, il existe des approches cherchant à regrouper des pixels selon une certaine *continuité* de l'information photométrique de l'image (niveau de gris, couleur etc.) pour former des régions homogènes ; les méthodes par lignes de partage des eaux (*watershed*) [Vincent 91] constituent un bon exemple. D'autre part, il existe des approches cherchant à séparer les pixels selon une certaine *discontinuité* de l'information photométrique, en délimitant des contours de régions hétérogènes. Les méthodes par contours actifs, de type *snakes*, sont les plus utilisées [Kass 88]. Enfin, il y les méthodes mixtes qui renforcent à la fois l'homogénéité intra-région et l'inhomogénéité inter-régions, ou bien celles qui s'appuient sur le principe de la classification : par exemple, en utilisant des approches non paramétriques de partitionnement de données multidimensionnelles comme le *mean-shift* [Comaniciu 02].

Un des modes opératoires de la segmentation sémantique récemment utilisé est d'explorer une troisième voie en effectuant une segmentation initiale de l'images en « super-pixels », qui sont des régions supposées délimiter visuellement des zones homogènes de l'environnement tridimensionnel. Cette segmentation initiale, que nous appellerons « sur-segmentation », peut combiner des informations de natures différentes plus riches que le niveau de gris ou la couleur comme des informations 3D (de la carte de disparité à la carte de profondeur en passant par le champ des normales), ou des informations contextuelles. L'intérêt de poser la (sur-)segmentation d'une image comme problème conjoint à celui de la classification de ses super-pixels, en créant donc des entités de niveau intermédiaire, est double : tout d'abord, on augmente les chances de délimiter plus précisément les contours des objets en introduisant des connaissances sur la scène, et

ensuite, on peut concevoir des algorithmes généralement moins coûteux en temps de calcul et de traitement. En règle générale, cette sur-segmentation initiale n'est pas remise en compte par l'étape de classification qui lui succède. Il est donc important d'intégrer dès le début l'intégralité des connaissances tridimensionnelles dont l'on peut disposer. Un exemple de sur-segmentation est donné dans la figure 2.

FIGURE 2 – Image originale et sur-segmentation en super-pixels (illustration extraite de [Hoiem 07]).

Dans le cadre des travaux effectués au cours de cette thèse, le problème de vision par ordinateur que nous abordons est celui de la *segmentation sémantique de séquences d images acquises dans un environnement urbain*. Il s'agit ici d'associer à chaque région d'une image une classe d'objets « urbains » prédéfinis (façade de bâtiment, route, ciel, végétation, voiture etc.). Une segmentation sémantique des images permettrait ainsi d'accéder numériquement aux informations d'objets réels, présents dans différents lieux physiques. Ces objets tels que la signalisation, l'équipement routier ou encore le mobilier urbain, sont aujourd'hui identifiés et positionnés manuellement. Les images géo-référencées de lieux extérieurs sont utilisées depuis peu, comme source d'information afin d'améliorer la cartographie. Cependant, le traitement automatique et l'exploitation de ce type de données restent encore limités.

Les scènes urbaines souvent associées dans la littérature à des représentations tridimensionnelles comportent de fortes « régularités structurelles » permettant de modéliser les murs plats des bâtiments, le fait qu'ils puissent être peu texturés, leurs arêtes saillantes etc. L'hypothèse de planéité par morceaux [Mičušík 10] concernant le modèle géométrique a été souvent utilisée afin de fortement contraindre le problème traité (reconstruction etc.), d'autant plus si l'on y rajoute l'hypothèse de Manhattan, c'est-à-dire la restriction à trois directions principales dominantes dans la géométrie du modèle. Malheureusement, un tel modèle simplifié est peu adapté aux cas réels de scènes urbaines « filmées sur le vif » qui peuvent contenir des objets aux caractéristiques hautement variables : objets rigides ou déformables, statiques ou en mouvement, réfléchissants ou dans l'ombre, occultés et/ou occultants. L'hypothèse de planéité par morceaux est dans ce cas insuffisante et doit être remplacée par une hypothèse de planéité partielle autorisant une scène urbaine à contenir à la fois des objets plans et non-plans.

L'approche que nous avons choisie, représentée dans la figure 3, consiste à utiliser un processus

de « sur-segmentation géométrique » qui regroupe les pixels similaires et cohérents avec les contours des objets, en entités de type super-pixel, à partir d'un ensemble de prises de vues, en accord avec la géométrie de la scène et deux cas de figures : « zone plane » ou « zone non-plane ». L'originalité de notre approche porte sur la construction de super-pixels qui s'appuie sur l'optimisation des nombreuses propriétés de l'ensemble des données disponibles (séquence d'images, reconstruction tridimensionnelle préliminaire éparse ou dense, pose des différentes caméras etc.) sous l'hypothèse d'une planéité partielle de l'environnement urbain. Ainsi, par ce travail, nous cherchons à démontrer l'utilité de combiner l'information photométrique et géométrique pour obtenir une segmentation plus robuste.

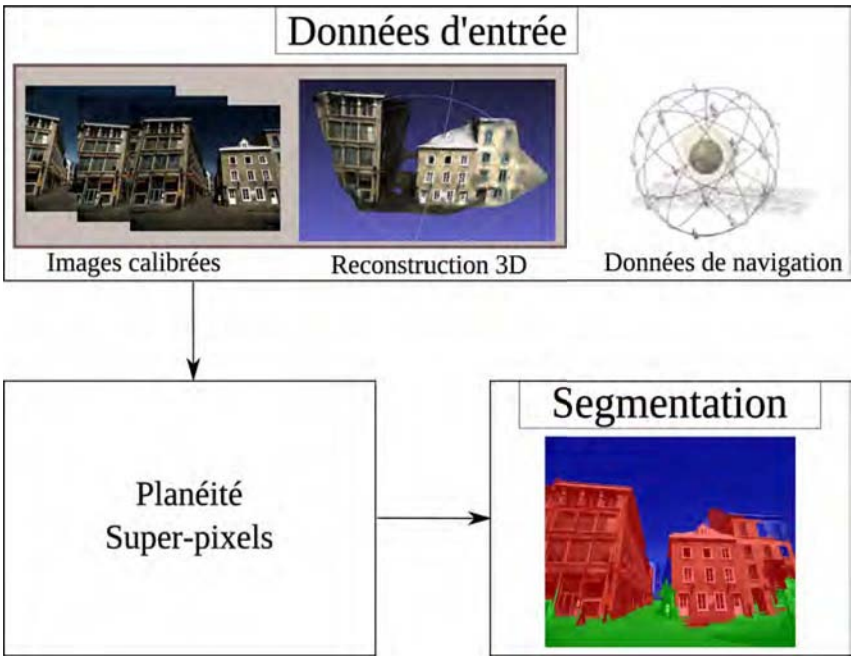


FIGURE 3 – Vue d'ensemble de l'approche proposée utilisant l'information photométrique (super-pixels) et géométrique (planéité) pour l'obtention d'une segmentation sémantique en plans.

Dans le cadre de cette thèse, nous ne considérons que les données multi-vues issues d'un capteur optique (appareil photographique ou caméra) en ne prenant pas en compte tout autre donnée complémentaire, par exemple comme celles collectées par la plateforme multi-capteurs `imajbox` (embarquée dans un véhicule en mouvement). Notre motivation est de fournir des algorithmes qui soient compatibles avec les données calculées par tout module de reconstruction tridimensionnelle à partir d'une séquence d'images.

Il est intéressant de noter que malgré les avancées effectuées dans le domaine de la compréhension de scènes, peu de travaux de segmentation de scènes d'extérieur utilisent l'information de cohérence photométrique provenant d'un ensemble de prises de vues d'une même scène. Bien que dans le cas d'une segmentation d'une image unique [Moore 09, Gould 08, Hoiem 05b], les approches récentes de sur-segmentation proposent de combiner l'information géométrique à l'information photométrique, cette fusion d'informations est encore peu utilisée dans les approches

multi-vues.

De manière plus détaillée, ce mémoire s'organise donc en quatre chapitres.

- Le chapitre 1 – Approches pour la compréhension visuelle de scènes urbaines – présente dans un premier temps les caractéristiques utiles à la compréhension générale, comme les points d'intérêt, les contours ou les région, la notion d'orientation tridimensionnelle ou encore le contexte. Puis nous présentons un état de l'art des familles d'approches de compréhension de scènes urbaines utilisant des informations géométriques et/ou sémantiques.
- Le chapitre 2 – Problème et données – traite du problème de compréhension automatique de scènes urbaines et en particulier de la segmentation sémantique en zones planes. Nous développons la méthodologie utilisée, les notions de vision par ordinateur indispensables à une compréhension du problème ainsi que les données disponibles et la les critères d'évaluation utilisés.
- Le chapitre 3 – Classification de zones planes – présente une analyse des mesures de cohérence photométrique inter-images et un protocole d'évaluation de ces mesures, appliqué à la classification de zones planes/non-planes.
- Le chapitre 4 – Applications à la segmentation en super-pixels – propose un état de l'art des approches de sur-segmentation en super-pixels, avec une description détaillée d'une des approches reconnue pour ses performances : *Simple Linear Iterative Clustering*, SLIC. L'approche que nous proposons est une extension de SLIC, qui intègre de l'information géométrique renforcée par un critère photométrique provenant de la comparaison inter-images. Enfin, le constructeur de super-pixels géométrique est intégré à une chaîne globale de segmentation sémantique en plans.

Chapitre 1

Approches pour la compréhension visuelle de scènes urbaines

Sommaire

1.1 Introduction

Dans ce travail de thèse, nous supposons avoir en entrée plusieurs image d'une même scène et nous souhaitons obtenir en sortie une segmentation sémantique de cette scène. La segmentation sémantique d'une scène consiste à attribuer à chaque pixel de l'image une étiquette correspondant à l'objet qu'il représente visuellement. Ce problème de segmentation sémantique est complexe mais aussi mal-posé [Mori 04, Hoiem 08] étant donné, entre autres, qu'il est possible de définir plusieurs niveaux d'étiquetage d'une même scène, comme illustré dans la figure 1.1. Cette ambiguïté dépend du dictionnaire de mots disponibles pour décrire une scène, ainsi que des objectifs de la tâche à accomplir [Levy-Schoen 68]. Par exemple, dans la figure 1.1, en (a) on cherche à reconnaître des objets composant la scène sans localisation précise, en (b) à détecter leur position [Alexe 10, Sturgess 07], en (c) à obtenir une segmentation [Dey 10] qui peut être sémantique comme en (d) si un découpage précis de chaque objet est recherché.

Notre objectif initial, vis-à-vis de la tâche segmentation sémantique, était de distinguer au minimum le sol, des bâtiments verticaux et le ciel. Dans ce contexte précis, cf. le chapitre 2 qui présente tous les détails sur les données utilisées, nous avons la possibilité de prendre en compte à la fois des aspects photométriques, géométriques et sémantiques. C'est la raison pour laquelle notre analyse de l'état de l'art s'oriente vers l'étude des approches faisant intervenir tous ces aspects. Plus précisément, nous citerons des approches de segmentation opérant aussi bien dans l'espace 3D d'une reconstruction préliminaire (approches essentiellement multi-vues) que des approches opérant directement dans l'espace 2D de l'image (approches généralement mono-vue ou multi-vues).

FIGURE 1.1 – Illustration des différentes segmentations sémantiques possibles d'une même scène (extrait de [Zhang 13]). (a) Catégorisation des objets d'intérêt sans localisation, (b) localisation des objets par boîtes enveloppantes, (c) localisation par extraction des contours ou segmentation, (d) segmentation sémantique de la scène où une étiquette est associée à chaque région de l'image.

La compréhension d'une scènes urbaine, qu'elle soit réalisée à partir d'une segmentation sémantique 3D d'une reconstruction préliminaire de la scène ou d'une segmentation sémantique 2D des images, constitue un sujet de recherche très actif. Les travaux de [Musialski 12], présentent une vue d'ensemble des méthodes de reconstruction de zones urbaines, en distinguant les approches suivant le type de données utilisées, comme par exemple, des données aériennes ou terrestres et des données provenant d'un capteur optique ou d'un capteur laser (*Light Detecting And Ranging scans* (LiDAR)). Actuellement, il existe de nombreux systèmes proposant une reconstruction 3D partielle, pseudo-réaliste, d'une ville dans sa globalité, comme ceux proposés

par *Google Earth* ou *Microsoft Virtual Earth*. En terme de reconstruction 3D de scènes urbaines, les travaux académiques les plus aboutis et les plus connus [Furukawa 09] extraient les axes principaux de la scène et s'appuient sur une hypothèse de planéité par morceaux. En parallèle, de nombreuses solutions faisant intervenir une étape d'apprentissage ont été proposées pour fournir une segmentation sémantique [Kim 12, Joulin 12, Lin 15]. Même si la notion d'apprentissage n'est pas utilisée dans ces travaux de thèse, car nous cherchons à mettre en place une approche descriptive, nous citerons de nombreux travaux faisant intervenir cet aspect, sans en donner les détails. En segmentation sémantique 3D, comme en segmentation sémantique, le cas particulier des scènes urbaines implique un certain nombre de particularités photométriques (manque de textures ou motifs répétitifs), géométriques (planéité des surfaces) et sémantiques (découpage en sol, bâtiment et ciel). Ces connaissances *a priori* permettent en partie de simplifier le problème de reconstruction 3D ou de segmentation, comme le fait d'avoir des structures régulières ou planes. En revanche, le manque de texture ou la présence de motifs répétitifs sont souvent des difficultés majeures pour ce type de scènes à analyser puisque cela rend, entre autres, la tâche de mise en correspondances de primitives homologues délicate.

En général, les travaux relatifs à la segmentation sémantique 3D d'une scène urbaine ont une formulation géométrique du problème et s'appuient sur une reconstruction préliminaire éparse ou dense de la scène réalisée à partir de plusieurs images. À l'opposé, les travaux relatifs à la segmentation sémantique 2D s'appuient sur l'extraction de caractéristiques dans une seule image avec l'ajout d'une étape d'apprentissage sur un ensemble de données détaillées (présentant des vérités terrain). La présentation de l'état de l'art s'articulera donc autour de ces deux aspects : l'utilisation de connaissances géométriques (reconstruction, planéité, etc.), éventuellement couplées à des notions de segmentation sémantique (super-pixels, apprentissage, etc.). Nous terminerons sur une description plus détaillée des approches les plus proches des travaux de cette thèse car ils intègrent des connaissances géométriques au sein d'une segmentation sémantique de la scène. Avant d'aborder ces aspects, il nous paraît important de rappeler certaines définitions ainsi que de fournir des références majeures sur les types de primitives ou d'éléments manipulés, à savoir, les contours, les points d'intérêt et la notion de sur-segmentation via la construction de super-pixels car ces éléments sont utilisés quelque soit le type d'analyse de la scène : géométrique ou sémantique.

1.2 Primitives usuelles pour l'analyse de scènes

Nous présentons dans cette section les primitives usuelles utilisées pour l'analyse de scènes : des contours aux régions en passant par la notion de points d'intérêt.

1.2.1 Détection de contours

Un contour est une frontière qui sépare deux objets/entités et qui se traduit par une variation d'intensité. Ainsi, une manière de définir un contour est de l'associer aux ruptures ou changements d'intensité dans l'image, par exemple en les faisant correspondre à un maximum local de la norme du gradient. Un des détecteurs de contour les plus couramment utilisés est le détecteur de [Canny 86] mais nous pouvons également citer les détecteurs utilisant des contours actifs [Kass 88] ou encore des courbes de niveaux (*level set*) [Chan 01]. Il existe également des approches s'appuyant sur la notion de lignes de partage des eaux [Vincent 91]. Pour les contributions plus récentes, nous pouvons citer l'approche par probabilité de contours introduite dans [Martin 04] et le détecteur de segments de droites, appelé *Line Segment Detector* (LSD) [Grompone von Gioi 12], qui utilise une approche dite *a contrario* permettant de contrôler le nombre de fausses détections. Ces techniques d'extraction de primitives linéaires sont utilisées dans la construction de primitives comme les super-pixels de [Duan 15] ou de la compréhension de scènes d'intérieur ou extérieur comme [Lezama 14, Witt 13, Bazin 12].

1.2.2 Notion de points d'intérêt

Un point d'intérêt est un point de l'image qui possède des caractéristiques particulières (forte variation d'intensité, texture etc.) qui permettent de le distinguer des autres points de l'image. L'idée est qu'un point d'intérêt correspond à la projection d'un point 3D particulier de la scène, qui a une forme « saillante » (coin, jonction, etc.). Ces caractéristiques du point d'intérêt doivent permettre la reconnaissance ou le suivi non ambigus en comparaison avec les autres entités de la scène. Ainsi, dans la plupart des cas, un point d'intérêt et son voisinage présentent des caractéristiques photométriques et géométriques discriminantes. Nous pouvons citer deux des détecteurs de points d'intérêt les plus connus : le détecteur de [Harris 88] et le détecteur SIFT, *Scale Invariant Feature Transform* [Lowe 04], robuste aux transformations affines et multi-échelle. Les travaux de [Mikolajczyk 04] donnent par ailleurs un état de l'art des détecteurs de points d'intérêt les plus connus.

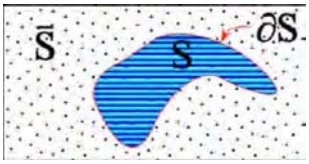
1.2.3 De la segmentation à la sur-segmentation (notion de super-pixels)

Formellement, si I désigne le domaine d'une image, la segmentation de l'image consiste à obtenir une partition de I en régions, notée $S = \{S_1, S_2, \ldots, S_n\}$, qui vérifie les propriétés de la figure 1.2.(a) où P est un prédicat permettant d'évaluer l'homogénéité de toute partie de I.

Historiquement, il existe deux grandes catégories de méthodes de segmentation qui utilisent deux types de primitives : celles s'appuyant sur les contours qui mettent en avant les discontinuités présentes dans l'image, et celles utilisant les régions de l'image qui favorisent le regroupement de pixels adjacents et similaires suivant certains critères de continuité dans l'image (similarité, proximité etc.). Le problème de segmentation binaire, comme le problème de classification binaire [Ren 07, Rubio 12], consiste à distinguer dans une image l'objet d'intérêt (souvent appelé

1. $\bigcup_i S_i = I$
2. $\bigcap_i S_i = \emptyset$
3. $P(S_i) = \text{vrai} \quad \forall i$
4. $P(S_i \cap S_j) = \text{faux} \quad \forall i, j \text{ avec } i \neq j.$

(a)



(b)

FIGURE 1.2 – Segmentation en n classes : En (a), la définition ; en (b), la cas $n = 2$ avec un objet S et son arrière-plan $\bar{S}$, où ∂S désigne le contour de la région.

avant-plan/*foreground*) de l'arrière-plan (*background*), comme le montre la figure 1.2.(b). La difficulté réside dans le fait de trouver soit un contour fermé (on recherche ∂S dans la figure 1.2.(b)) , soit la région de l'objet représenté (on recherche S, ou de façon équivalente $\bar{S}$, dans la figure 1.2.(b)). Il est délicat de donner un état de l'art des méthodes de segmentation en général et nous proposons donc un ensemble d'articles de référence pour le lecteur intéressé. Ainsi, la thèse de [Brun 96, partie 2] et l'article de [Dey 10] fournissent une description des méthodes classiques de segmentation. De plus, les travaux de [Szeliski 10, chapitre 5] donnent une présentation des techniques actuelles de segmentation en les liant aux applications existantes. Enfin, l'article de [Vantaram 12] présente une vue d'ensemble des nouvelles tendances de segmentation sur les images couleur.

Réaliser une segmentation est une tâche délicate, et de ce fait, de nombreuses approches récentes s'appuient sur une « sur-segmentation » préliminaire de l'image, c'est-à-dire une segmentation telle que le nombre de régions obtenues est supérieur au nombre de régions attendues. De façon spécifique, dans une segmentation sémantique, le nombre de régions obtenues est supérieur au nombre d'objets présents dans l'image. C'est ainsi que le terme de « super-pixel » est utilisé pour la première fois en 2003, dans [Ren 03]. Un super-pixel est une région fermée de l'image relatif à une segmentation de celle-ci de niveau intermédiaire, c'est-à-dire qui possède un support spatial plus cohérent, suivant un critère photométrique, mais moins régulier, suivant un critère géométrique, qu'une zone rectangulaire. Plus ou moins implicitement, dans la littérature du domaine, un super-pixel permet souvent de représenter un objet ou une partie d'un objet. Il existe autant de variations de constructeurs de super-pixels qu'il y a de définitions mathématiques pour les termes « structure locale » et « structure cohérente ». Il s'avère que les super-pixels sont souvent utilisés comme étape préliminaire à une segmentation sémantique car ils permettent aussi de réduire les temps de calcul dans les différentes phases du traitement [Achanta 12], tout en proposant une partition de l'image suffisamment régulière pour s'adapter aux contours des objets [Levinshtein 10]. Étant donné que les travaux de cette thèse portent, entre autres, sur la proposition d'une nouvelle approche de construction de super-pixels cohérents avec la géométrie de la scène, nous détaillerons les constructeurs existants dans la section 4.2.

1.3 Utilisation de connaissances géométriques

L'information d'intensité ou de couleur ainsi que la position relative dans l'image, associées à chaque pixel, ne permet pas d'inférer directement une information 3D. De plus, elle ne représente pas toujours une information 2D suffisante pour mettre en correspondance de manière fiable les pixels. Pour décrire une scène, il est donc utile d'extraire toutes les informations géométriques possibles, qu'elles soient 2D ou 3D.

L'information géométrique 2D, concerne la notion de contours, de points d'intérêt, de régions ou de super-pixels que nous avons abordés dans la section précédente. Les travaux proposés par [Chatfield 09], utilisent un descripteur de similarités [Shechtman 07], afin de mettre en correspondance, d'identifier et de classifier des images ayant des caractéristiques visuelles et géométriques proches. À partir d'une image contenant un objet requête, par exemple un cœur ou une pomme, l'algorithme récupère d'autres images représentant les mêmes objets ayant en particulier une forme commune et pouvant subir une déformation non-rigide.

L'information 3D peut être caractérisée par différents types de données : les paramètres de calibrage, la géométrie des points et lignes de fuite, la position des points 3D, l'orientation des normales des surfaces présentes dans la scène ainsi que les cartes de profondeur des images. Dans le reste de cette section, nous aborderons les approches qui comme [Chauve 10] utilisent essentiellement une information géométrique.

1.3.1 Utilisation simple d'une reconstruction partielle ou d'une carte de profondeur

Il existe deux approches « classiques » pour la reconstruction 3D d'une scène que nous décrirons plus en détails dans le chapitre suivant, cf. partie 2.5.3. Celles de type *Structure-from-Motion* (*Structure from Motion* (SfM)) sont utilisées lorsque la seule donnée en entrée disponible est une collection d'images (ordonnée ou non) éventuellement complétée par les paramètres internes de calibrage de la (des) caméra(s). Les approches SfMfournissent en sortie une reconstruction éparse de la scène (en général sous la forme d'un nuage de points 3D) ainsi que les différentes poses de la (des) caméra(s). Celles de type *Multi-View-Stereo* (*Multi View Stereo* (MVS)) sont utilisées lorsqu'il est possible de disposer, en plus de la collection d'images, des différentes poses de la (des) caméra(s). Les approches MVS fournissent en sortie une reconstruction dense de la scène sous la forme d'un nuage de points 3D ou d'un maillage 3D, texturés ou non. En général, les approches SfM et MVS peuvent être mises en séquence pour obtenir le maximum d'informations 3D à partir du minimum d'informations 2D.

Dans la littérature, l'approche de [Furukawa 09] est très souvent utilisée. Notons aussi que dans le cas d'un capteur actif avec lumière structurée comme la Kinect de Microsoft pour les scènes d'intérieur, ou directement d'un capteur de profondeur (laser etc.), des reconstructions denses peuvent être obtenues et utilisées pour la compréhension de scènes [Scharwächter 13, Erbs 11]. Les sources disponibles (passives et actives) peuvent être combinées pour renforcer la

robustesse de l'approche, comme dans [Miksik 15]. Ainsi, dans [Brostow 08], une approche de reconstruction dense de scènes d'extérieur, à partir de plusieurs vues et d'un nuage épars de points 3D obtenu par une approche du type SfM, est utilisée.

Contrairement à de nombreux travaux que nous étudierons par la suite, les auteurs n'exploitent pas l'hypothèse de planéité et exploitent uniquement la profondeur estimée pour chaque point de contrôle/point d'intérêt de la scène en manipulant directement les maillages 3D obtenus à partir d'une triangulation de Delaunay de ces points d'intérêt. Cela simplifie la mise en œuvre de l'approche, mais, ne permet pas de corriger correctement les défauts de reconstruction liés au problèmes des textures répétitives ou similaires entre deux plans de la scène avec des orientations différentes.

1.3.2 Planéité

Dans la littérature, pour analyser des scènes d'intérieur ou d'extérieur en milieu urbain, on utilise très fréquemment comme information *a priori* les hypothèses d'un univers de type Manhattan (respectivement d'Atlanta), c'est-à-dire le fait qu'il existe trois (respectivement cinq) directions dominantes dans l'espace 3D [Furukawa 09]. Cette hypothèse forte est particulièrement utile pour l'estimation de l'orientation des surfaces visibles [Coughlan 03]. Or, dans un contexte de scènes urbaines, on fait souvent l'hypothèse que la scène est plane par morceaux. En effet, ces représentations 3D sont stables, compactes et facilement manipulables. Le problème de construction de facettes tridimensionnelle est ancien [Zagrouba 94] en particulier dans le contexte de la stéréo-vision. Lorsqu'un nuage de points est disponible, l'estimation des plans peut être manuelle ou automatique, en utilisant par exemple une approche de type RANSAC [Fischler 81], dont les détails sont présentés dans l'annexe 4.7. Cette méthode itérative est robuste aux données aberrantes. Elle permet d'estimer les paramètres d'un modèle. De plus, elle a inspiré l'approche de *J-linkage* [Toldo 08, Fouhey 10], qui permet l'estimation de modèles multiples. Les approches peuvent se différencier suivant le fait que les plans soient estimés à partir d'un nuage de points 3D [Bartoli 07], ou à partir des droites détectées [Sinha 08, Mičušík 10] ou directement à partir de la carte de profondeur [Gallup 10].

Dans [Bartoli 07], l'algorithme de segmentation itérative en plans est une version modifiée de l'algorithme de RANSAC [Fischler 81]. La première contribution est de travailler avec des données recouvrantes, c'est-à-dire qu'un point peut appartenir à plusieurs plans dans le cas d'intersection de deux plans sécants. La seconde contribution est d'optimiser le problème de cohérence photométrique inter-image, à l'aide d'un critère combinant l'information géométrique provenant du nuage de points 3D avec l'information photométrique inter-image de l'ensemble d'images. Il s'agit de l'utilisation du critère de r-consistance, introduit par [Kutulakos 00], que nous détaillons au paragraphe 3.2.1.

L'approche proposée dans [Sinha 09], qui est dans la continuité des travaux de [Delage 07], utilise un nuage de points 3D combiné à un ensemble épars de droites détectées dans les images.

Cela permet d'obtenir une première segmentation en plans qui est améliorée en utilisant un modèle markovien faisant intervenir un critère photométrique classique mais surtout un critère géométrique exploitant les droites détectées ainsi que les points de fuites, dans la fonction potentielle utilisée.

L'approche de [Fouhey 10, Fouhey 13] permet également de travailler à partir d'une seule image afin d'en extraire les caractéristiques 3D pour détecter les plans principaux.

Les auteurs de [Habbecke 06, Bao 14] combinent l'estimation d'un nuage épars de points 3D par SfM à l'estimation de zones planes par morceaux. Alors que [Habbecke 06] travaille sur des images extérieures dans des condition MVS, [Bao 14] l'applique à des scènes intérieur en intégrant les orientations principales et en s'appuyant sur une configuration de Manhattan avec des orientations dominantes.

Dans le cas d'éléments naturels, i.e. non manufacturés, des éléments comme les arbres ou les animaux ne peuvent pas être représentés simplement par un ensemble de plans car leur représentation 3D nécessite une courbure de surface non nulle. C'est pourquoi, certains auteurs, comme [Gallup 10], considèrent un environnement moins contraint pouvant être constitué d'un ensemble de surfaces planes et non-planes. Dans ces travaux, la segmentation finale s'appuie à la fois sur un critère de cohérence photométrique inter-vues de [Birchfield 98], calculé via l'estimation des plans par RANSAC sur la carte de profondeur et sur une étape d'apprentissage supervisée s'appuyant sur les couleurs et textures de régions annotées manuellement. Les auteurs ont testé des approches de sur-segmentation en super-pixels [Felzenszwalb 04, Shi 00] mais ont choisi un découpage en grille régulière (16×16) car ils considèrent qu'il y a suffisamment d'information pour trouver le contour précis des objets, en combinant la cohérence photométrique et le terme de régularisation. L'un des avantages d'utiliser des rectangles est qu'ils contiennent tous la même quantité d'information, par leur régularité et leur densité.

L'approche proposée par [Sinha 14] présente un algorithme de stéréo-vision qui exploite une estimation locale des plans qui constituent la scène. L'intérêt de cette proposition réside essentiellement dans le fait que la méthode de mise en correspondance de pixels possède un temps d'exécution faible comparée à d'autres approches classiques de la littérature. L'estimation des plans s'effectue à partir d'une première mise en correspondance éparse suivie par une étape itérative pour faire varier localement l'orientation des plans et ainsi affiner l'estimation du plan et de la profondeur associée à chaque pixel.

Pour conclure sur cet aspect de planéité, il semble indispensable d'utiliser cette information *a priori* dans le contexte de cette étude. Différentes techniques sont envisageables pour intégrer cette contrainte, et nous verrons par la suite comment combiner cette hypothèse géométrique à une hypothèse photométrique.

1.3.3 Approche de co-segmentation

La co-segmentation est définie comme le problème conjoint de la partition de plusieurs images représentant le même objet ou le même type d'objet. Dans le cas de segmentation en deux classes (avant-plan et arrière-plan), la plupart des méthodes de co-segmentation proposées (à l'exception de [Joulin 12]) supposent une forte similarité des histogrammes de couleurs pour l'objet d'intérêt considéré comme avant-plan [Hochbaum 09, Bagon 08, Rother 06]. Ce type d'*a priori* est raisonnable lorsque l'on travaille sur un ensemble d'images, d'un même objet ou dans le cas de deux images acquises de manière proches temporellement (une séquence vidéo, par exemple). Ce type d'approche a été testé dans le contexte de scène urbaine [tos 07]. Les auteurs de [Rubio 12] proposent une méthode de co-segmentation non-supervisée et pour cela, une mise en correspondance de régions est utilisée. Enfin, dans le cas d'une segmentation à plus de deux classes, le problème a été abordé dans [Kim 12]. Nous pouvons également citer les recherches qui s'orientent également vers la co-segmentation de modèles 3D [Hu 12].

1.3.4 Utilisation d'une sur-segmentation en super-pixels

Plus récemment, l'information de profondeur est également utilisée dans les approches de segmentation, que ce soit pour une classification des points d'un nuage tridimensionnel [Tombari 07], ou pour l'estimation du flux optique [Nawaf 14, Vogel 13]. Dans ce cas particulier, il est exigé qu'un super-pixel ne contienne que des pixels avec une profondeur cohérente, cette notion de cohérence étant, la plupart du temps, relative au fait qu'on suppose que le super-pixel représente une surface plane. Ainsi, les profondeurs estimées doivent être correctes par rapport à cette hypothèse de planéité [Matsuo 13]. Les auteurs de [Bleyer 11] utilisent les hypothèses suivantes : un objet est compact et connexe dans l'espace 3D, et les parties visibles d'un objet sont similaires en apparence (s'il n'y a pas de réflexion spéculaire). Dans cette approche, les super-pixels sont utilisés pour contraindre la mise en correspondance. Dans [Bódis-Szomorú 14], que nous présentons en détails dans la partie 1.5.4, la contribution principale des auteurs est de combiner la reconstruction 3D éparse de SfM et les superpixels [Achanta 12] dans une approche *Markov Random Field* (MRF) multi-vues afin d'obtenir une approximation d'une reconstruction dense et plane par morceaux. L'article [Bódis-Szomorú 15] cherche à accélérer, simplifier et renforcer la fiabilité de la reconstruction 3D en s'appuyant sur un maillage de Delaunay 2D (utilisant différents types de points tels que les points de Harris, les extrémités des polygones représentant des contours binaires de l'image ou encore les sommets des polygones correspondant à la simplification des superpixels [Felzenszwalb 04]). Les points 3D du nuage éparse obtenus par l'approche de SfM, sont considérés comme des points de contrôle (GCPs, *Ground Control Points*) afin d'estimer l'orientation (pas uniquement fronto-parallèle) et la profondeur des facettes.

Bien que nos travaux portent sur la compréhension de séquence d'images de scènes urbaines, nous citons brièvement quelques articles s'intéressant à la segmentation vidéos et qui intègrent la cohérence inter-images dans les super-pixels [Grundmann 10] ou de manière plus générale dans

leur approche [Raza 13, Rubio 12, Floros 12].

1.4 Utilisation de connaissances sémantiques

Parmi les approches utilisant la connaissance d'information sémantique, nous pouvons distinguer le niveau d'apprentissage requis. Alors que certains travaux intègrent directement des informations provenant d'une vérité terrain [Liu 10, Saxena 09], d'autres choisissent la récupération automatique d'information à partir de grandes bases de données [Tighe 13] ou encore d'utiliser des données fournies par des utilisateurs [Sinha 08] ou des informations *a priori* [Xu 14, Moore 09].

Lors de la segmentation de scènes extérieures, il est intéressant d'inférer à partir de l'image sur les couches 3D de la scène. Plus précisément, la scène peut être décomposée en trois zones : le sol qui est horizontal (route, trottoir, etc.), les objets verticaux (bâtiments, arbres, véhicule ou piétons, etc.) et le ciel. Ceci a été réalisé entre autres par [Hoiem 07, Alvarez 12].

L'article de [Van Gool 13] met en avant l'utilité de combiner la robustesse des approches descendantes (*top-down*) en s'appuyant sur des règles conformes aux structures architecturales, utilisées pour la modélisation 3D, avec la flexibilité des approches ascendantes (*bottom-up*) des approches SfM. En effet, dans l'approche proposée, *Support Vector Machine* (SVM) est utilisé pour la classification de la scène en quatre classes : sans bâtiment, morceau de bâtiment, façade, rue. Les auteurs tentent de détecter le type d'architecture (haussmanienne, néo-classique, etc.) afin d'obtenir une fine description des façades (porte, fenêtre, mur, balcon, etc) pour une reconstruction 3D.

Récemment, l'article de [Badrinarayanan 15] propose une système d'apprentissage profond appelé SegNet utilisant un réseau à convolution entraîné pour étiqueter les pixels d'une image dans un ensemble de classes sémantiques. SegNet est composé d'une phase d'apprentissage des paramètres du modèle et d'un module de classification construit à partir d'une séquence composée de couches d'encodeurs et de couches de décodeur associés. Il est possible de tester l'approche avec la démonstration temps réel en ligne [1].

1.4.1 Utilisation de la silhouette ou de modèles de forme des objets

La silhouette d'un objet correspond à la projection 2D de son enveloppe visuelle. Elle est parfois appelée modèle de forme. Dans l'espace 2D, la silhouette peut-être combinée avec un modèle de segmentation qui est initialisé par un algorithme d'apprentissage non-supervisé, comme dans l'approche de co-segmentation proposée par [Dai 13] ou intégré dans un modèle de mise en relation des contours de l'objet [Toshev 10]. D'autre auteurs, comme [Wang 14] proposent, d'utiliser la cohérence des contours reprojetés d'un modèle 3D pour valider la fusion de deux nuages de points 3D.

1. SegNet démonstration en ligne `http://mi.eng.cam.ac.uk/projects/segnet/`

1.4.2 Position des objets les uns par rapport aux autres

Le travail de référence dans la prise en compte du contexte et entre autres de la position des objets les uns par rapport est celui de [Torralba 02]. Ces travaux indiquent que l'on peut toujours arriver à modéliser les relations spatiales, ainsi que le contexte particulier relatif à un objet précis, si on étudie les distributions des niveaux de gris dans l'image. Plus précisément, en combinant une analyse de Gabor à une modélisation probabiliste, et en utilisant une phase d'apprentissage, les auteurs sont capables de définir un modèle propre à chaque objet à identifier dans une scène.

Une autre technique bien connue dans la littérature est l'approche de [Hoiem 07] (nous la reprendrons en détails dans la section 1.5.1, dont le principe réside dans l'extraction d'un maximum d'informations photométriques afin de les utiliser pour une compréhension globale de la scène en s'appuyant sur une phase d'apprentissage pour l'estimation de l'arrangement spatial de ces objets. Ce principe est également repris dans les travaux de [Saxena 09].

1.4.3 Utilisation d'une sur-segmentation en super-pixels

Les approches présentées dans ce premier paragraphe s'apparentent au travail réalisé par les papiers cités dans la section 1.3.2 mais elles complètent ce type d'approches en ajoutant un apprentissage. Dans [Saxena 08, Saxena 09], l'objectif est d'estimer la profondeur avec une reconstruction 3D des principales structures de la scène, cf. figure 1.3. L'image est sur-segmentée avec l'approche de [Felzenszwalb 04] en considérant que chaque région est plane, afin d'inférer avec un modèle MRF sur la position et l'orientation (sans contrainte, toutes les directions sont possibles) de chaque surface 3D correspondante. L'obtention d'un maillage 3D permet de représenter les relations 3D entre les différentes parties de l'image. Contrairement à [Hoiem 05b] et [Delage 07], les auteurs ne font pas d'hypothèse sur la structure de la scène telle que les surfaces verticales se situent sur le sol horizontal.

Les auteurs de [Häne 13] combinent la reconstruction 3D dense et la segmentation sémantique. Ils s'appuient sur l'optimisation d'une fonction d'énergie représentant un volume implicite. Les paramètres du modèle sont estimés lors d'une phase d'apprentissage sur des critères géométriques et d'apparences (super-pixels de [Comaniciu 02]).

Nous venons de présenter quelques approches de compréhension de scènes urbaines utilisant différents type d'information (géométrique ou/et sémantique). Dans la partie qui suit nous détaillons les approches qui nous semblent les plus proches de nos travaux ou qui nous ont fortement inspirés.

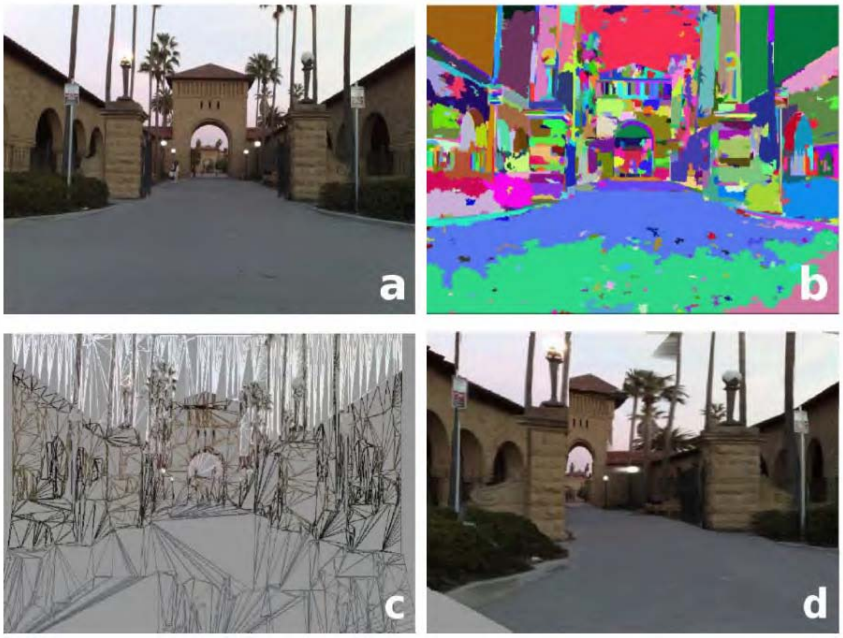


FIGURE 1.3 – Reconstruction 3D par super-pixels plan [Saxena 09] : (a) Image source, (b) sur-segmentation en super-pixels de [Felzenszwalb 04], (c) maillage à partir de l'hypothèse de planéité (orientation et position des plans), (d) vue de la scène depuis un nouveau point de vue.

1.5 Approches prenant en compte la planéité et une sur-segmentation en super-pixels

Nous avons vu dans les sections précédentes que certains travaux sont cités aussi bien pour utiliser des aspects géométriques que des aspects sémantiques. Nous allons à présent revenir en détails sur ces travaux car ils constituent l'ensemble des approches les plus proches de nos travaux et ont orienté cette thèse.

1.5.1 Approche *pop-up* mono-vue proposée par [Hoiem 05a]

Les travaux de Hoiem [Hoiem 05a, Hoiem 05b], illustrés dans la figure 1.4, ont pour objectif d'obtenir une classification en trois classes géométriques : le ciel, le sol et les objets verticaux. La classe des objets verticaux est elle-même divisée en sous-ensembles en fonction de l'orientation des surfaces. L'approche se décompose en quatre étapes :

1. Sur-segmentation multiple : l'approche de [Felzenszwalb 04] est utilisée afin de produire plusieurs sur-segmentations représentatives avec un jeu de paramètres qui font varier la taille et la forme des super-pixels.
2. Extraction des caractéristiques : la couleur, la texture, la position et la forme ainsi que la géométrie 3D (lignes et points de fuite) sont utilisées pour décrire les super-pixels et les régions de l'image.
3. Étiquetage des super-pixels : affectation d'un super-pixel à une des trois régions (sol, ciel,

objets verticaux) par calcul de la moyenne du maximum du log-vraisemblance sur les descripteurs des super-pixels d'une région.

4. Mise en perspective du modèle 3D : découpage et pliage des différentes régions verticales pour la génération d'un modèle 3D appelé *pop-up* par combinaison des informations géométriques et de l'estimation de la ligne d'horizon.

Dans une autre version de leurs travaux [Hoiem 08], les auteurs proposent de combiner des critères photométriques (couleur, texture) et géométriques ainsi que des informations *a priori* sur le contexte, comme la taille moyenne des personnes pour estimer, par exemple, la hauteur du système d'acquisition ou la distance entre la caméra et les objets. Enfin, dans la version la plus récente [Hoiem 11], un intérêt particulier est donné à l'interprétation des zones d'occultations, afin de renforcer la cohérence dans le découpage des objets lors de saut de profondeur.

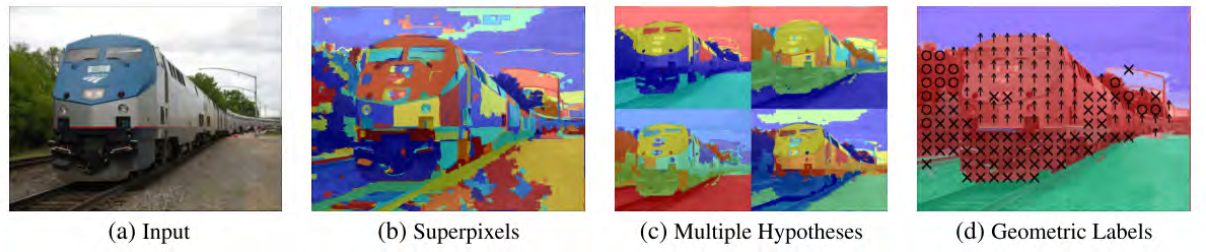


FIGURE 1.4 – Segmentation géométrique et sémantique par [Hoiem 05b] : (a) Image source, (b) Sur-segmentation selon [Felzenszwalb 04], (c) Exemples de regroupement de super-pixels en suivant différentes hypothèses, (d) classification en trois zones (ciel, sol, objets verticaux) où les zones verticales sont représentées par l'orientation des surfaces (ou → ou), les zones non-planes (représentée par ×) ou les zones poreuses ( ).

1.5.2 Approche par super-pixels plans proposée par [Delage 07]

L'approche de [Delage 07] propose une reconstruction 3D d'une scène d'intérieur en s'appuyant sur l'hypothèse d'une scène de Manhattan c'est-à-dire avec trois points de fuite orthogonaux deux à deux. La reconstruction nécessite deux étapes :

1. Segmentation [Felzenszwalb 04] de la scène où les régions sont considérées comme des zones planes ;

2. Estimation de l'orientation de ces zones supposées planes.

Ces deux étapes s'appuient sur une modélisation probabiliste, via des réseaux dynamiques bayésien et ds champ aléatoires de Markov (*Markov Random Fields* - MRF) afin de classifier les droites de l'image suivant les directions 3D dominantes, et de propager les étiquettes des droites aux pixels les plus proches afin d'estimer un masque pour les pixels appartenant au sol.

1.5.3 Super-pixels par balayage de plans [Mičušík 10]

Dans [Mičušík 10], à partir d'une sur-segmentation en super-pixels [Felzenszwalb 04] sur une image de référence, les auteurs émettent l'hypothèse que chaque super-pixel correspond à une zone 3D plane et ils utilisent la rétro-projection dans les images adjacentes pour estimer les super-pixels dans les autres images. Pour cela, ils introduisent la notion de balayage de plan qui consiste à faire varier les paramètres du plan pour maximiser la similarité inter-images. Comme illustré dans la figure 1.5, chaque région peut ainsi être projetée dans les images adjacentes via l'homographie induite par le plan de support (Π).

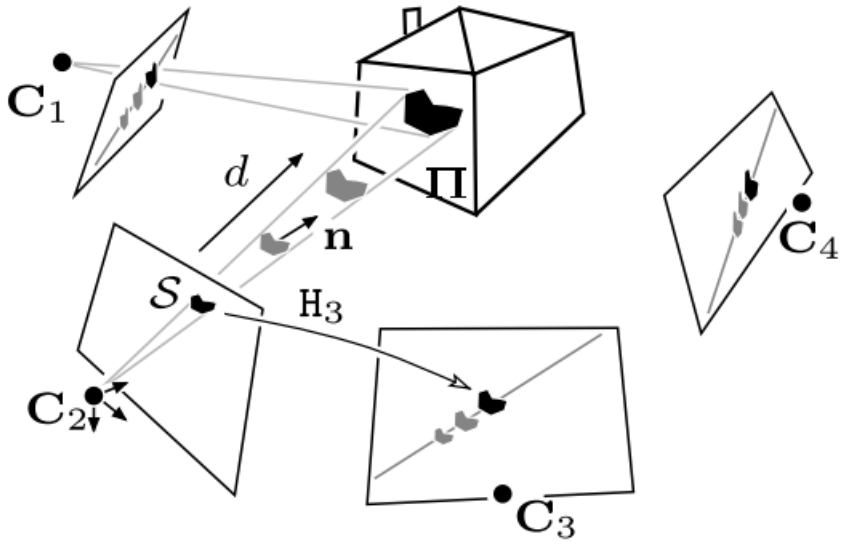


FIGURE 1.5 – Construction de super-pixel par balayage de plans [Mičušík 10] : Les centres optiques des caméras sont notés C_i. L'estimation des super-pixels $\mathcal{S}$ en noir, dans une image de référence (ici celle dont C_2 est le centre optique), correspondent à la zone plane $\Pi = [\vec{n}^\top d]^\top$, où $\vec{n}$ est la normale unitaire et d l'ordonnée à l'origine.

1.5.4 Approche d'approximation d'un modèle dense [Bódis-Szomorú 14]

L'approche [Bódis-Szomorú 14] d'approximation d'un modèle dense et plan par morceaux, illustrée dans la figure 1.6, est très proche de celle que nous avons proposée. L'algorithme proposé tente de répondre à trois problèmes de manière conjointe :

- ajuster les paramètres de chaque primitive plane ;
- segmenter la scène en mettant en relation les points 3D et les régions ;
- déterminer la visibilité de chaque région dans chaque image.

Les auteurs supposent que la scène est composée d'un ensemble de primitives planes et cherchent à générer, en utilisant l'approche de RANSAC, des hypothèses de plans qui expliquent et décrivent le nuage de points 3D obtenu par VisualSFM [Wu 11b]. Chaque image est au préalable décomposée en une collection de super-pixels avec l'approche [Achanta 12]. La solution au

problème global est obtenue en résolvant un problème d'optimisation d'une fonction d'énergie composée de trois termes :

- un terme unaire, correspondant au coût de l'association d'un label à un superpixel donné ;
- un terme d'appariement dans une vue donnée, qui pénalise deux super-pixels voisins appartenant à deux plans différents ;
- un terme d'appariement entre deux super-pixels d'une vue donnée, qui peut pénaliser les deux super-pixels qui appartiennent à différents plan s'ils partagent au moins un point 2D correspondant au même point 3D.

Cette application est plus stable aux changements d'échelles que les approches existantes de MVS.

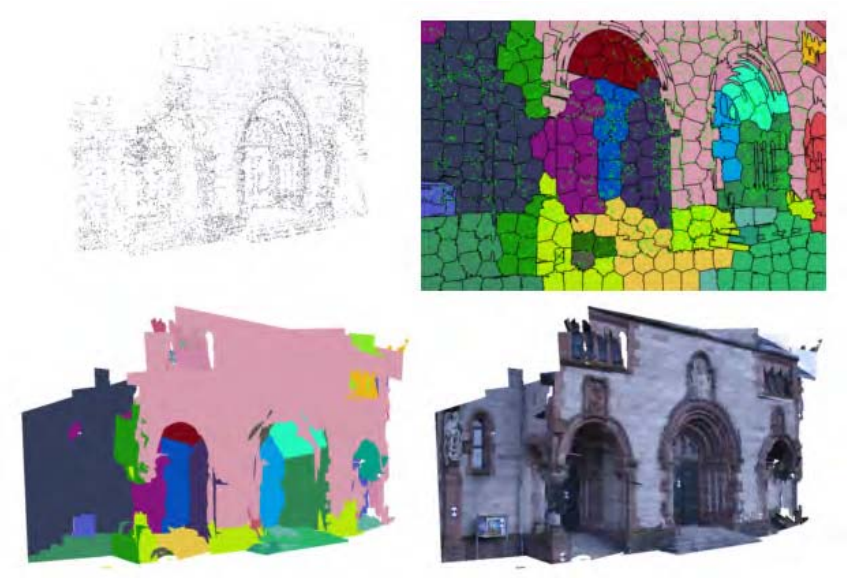


FIGURE 1.6 – Méthode de reconstruction dense combinant reconstruction éparse, planéité et superpixels [Bódis-Szomorú 14].

1.6 Synthèse sur les approches présentées

Ainsi, dans ce chapitre, nous avons abordé un ensemble de méthodes permettant d'analyser des scènes urbaines, aussi bien dans le domaine de la reconstruction 3D que dans celui de la segmentation sémantique. Nous avons mis en avant le fait que l'on peut utiliser une ou plusieurs images, des informations photométriques ou géométriques et en particulier en utilisant, voire en combinant la notion de planéité, ainsi que la notion de sur-segmentation ou de découpage en super-pixels.

Enfin, nous proposons dans le tableau 1.1 une vue synthétique des articles abordés dans cet état de l'art concernant la compréhension de scènes visuelles.

Approches	**Mono-vue**	**Multi-vues**	**Super-pixels**	**Planéité**	**Apprentissage**
[Tighe 13]	x		x		
[Delage 07]	x			x	
[Alvarez 12]	x				x
[Xu 14, Moore 09, Hoiem 05a]	x		x		x
[Fouhey 13, Liu 10, Saxena 09]	x		x	x	x
[Scharwächter 13, Grundmann 10]		x	x		
[Sinha 14, Lafarge 12, Chauve 10, Sinha 09, Bartoli 07, Toldo 08, Habbecke 06]		x		x	
[Badrinarayanan 15, Rubio 12, Brostow 08]		x			x
[Bódis-Szomorú 15, Bao 14, Bódis-Szomorú 14, Mičušík 10]		x	x	x	
[Raza 13]		x	x		x
[Gallup 10]		x		x	x

TABLE 1.1 – Synthèse des différentes approches de compréhension de scènes : en distinguant celles qui utilisent une vue unique ou du multi-vues, et intégrant soit des informations complémentaires comme les super-pixels, la planéité ou encore supervisées (profondeur, objets, etc.).

1.7 Conclusion

Dans le cadre de cette thèse, à partir de plusieurs vues d'une scène, nous cherchons à résoudre un problème de segmentation automatique non supervisée afin de classer chacune des régions obtenues vis-à-vis une description sémantique. Notre objectif n'est pas l'obtention d'une reconstruction tridimensionnelle dense de la scène, mais de segmenter l'image en régions cohérentes avec les structures représentées.

Les entités intermédiaires comme les super-pixels, fournissent un support spatial qui permet de faire des statistiques sur les pixels [Hoiem 05b], de réduire la complexité et le temps de calculs (en comparaison avec une analyse de la scène pixel à pixel) en renforçant la cohérence et les relations spatiales [Delage 07, Saxena 08]. Dans la littérature, la construction des super-pixels n'est pas contrainte par l'information 3D disponible. Or, parfois, en particulier dans le cas des scènes urbaines faiblement texturées ou présentant des textures répétitives où un critère photométrique seul n'est pas discriminant, malgré l'hypothèse sur le fait qu'un même super-pixel ne contient qu'un seul ou qu'une partie d'un seul et même objet, certains super-pixels contiennent plusieurs objets distincts, comme par exemple plusieurs morceaux de façades qui n'ont pas la même orientation. Nous proposons de réduire ce problème en intégrant directement la notion de planéité dans la construction des super-pixels. Plus précisément, nous souhaitons améliorer

l'estimation du support spatial des super-pixels en fusionnant l'information photométrique et géométrique disponible afin d'obtenir des super-pixels cohérents avec les contours des objets représentés et la géométrie de la scène.

Après la présentation du problème et des données traités, cf. chapitre 2, nous proposons dans le chapitre 3 un protocole d'évaluation des mesures de cohérence photométrique pour la classification de régions en zones planes ou non-planes. Ces mesures sont utilisées comme critère photométrique permettant d'infirmer ou de confirmer un *a priori* géométrique. Nous mettons en évidence la mesure la plus appropriée et celle-ci sera intégrée dans notre approche de construction de super-pixels, présentée dans le chapitre 4.

Chapitre 2

Problème et données

Sommaire

2.1 Introduction

Les avancées récentes dans le domaine de la reconstruction tridimensionnelle d'une scène observée à partir d'une collection d'images [Furukawa 09], fournissant en sortie un modèle mathématique défini par des quantités géométriques et photométriques, ouvrent de nombreuses possibilités d'applications nouvelles. En particulier, un tel modèle, augmenté des différentes poses de l'appareil d'acquisition et d'une certaine pré-interprétation géométrique voire sémantique, constitue une donnée d'entrée très informative du problème de compréhension automatique des images de la scène. Ainsi, de la même manière, nous cherchons à obtenir des régions ou des super-pixels cohérent(e)s sur plusieurs vues et ainsi décrire la scène avec des étiquettes ou classes adaptées.

Tout d'abord, nous introduisons les grandes lignes du problème considéré et le but recherché, à savoir une description sémantique de la scène. Ensuite, pour mettre en contexte nos travaux, nous présentons le système d'acquisition, l'`imajbox`, proposé par l'entreprise partenaire du projet : `Imajing`, ainsi que les données d'entrée impliquées dans la résolution du problème de segmentation sémantique de scènes urbaines. Ensuite, en s'appuyant sur les caractéristiques des données disponibles, la méthodologie utilisée est énoncée. Enfin, nous exposons l'évaluation mise en place avec les données de référence, les autres bases de données existantes et les critères utilisés.

2.2 Grandes lignes du problème considéré

(a) (b) (c)

FIGURE 2.1 – Exemple de segmentation sémantique sur une image de scène urbaine : (a) Image d'origine, (b) et (c) segmentation obtenue manuellement, appelée vérité terrain. Dans (b) chaque couleur correspond à une classe sémantique (sol, ciel, bâtiments, mobilier urbain), alors que dans (c) chaque couleur correspond à un objet. Par exemple, les façades appartenant à des groupes de bâtiments différents sont dissociées et la classe sol contient les objets route et trotoir.

Dans cette thèse, les scènes considérées sont des scènes urbaines, représentées par une collection d'images acquises depuis un véhicule en mouvement. Nous faisons l'hypothèse que les objets d'intérêt dans une telle scène sont associées à six classes principales d'objets : « sol », « ciel », « végétation », « construction urbaine »,« mobilier urbain » et « autre ». Le problème de compréhension d'images que nous cherchons à résoudre est celui d'une segmentation automatique

d'une image en régions, comme l'exemple de segmentation manuelle de la figure 2.1. Chaque pixel est associé à une région dont l'étiquette correspond à la classe de l'objet représenté.

Les données d'entrée correspondent aux sorties d'un module de « calcul de la structure et du mouvement » (traduction du terme anglo-saxon *Structure-and/from-Motion* / SfM) voire d'un module de « stéréoscopie multi-vues » (traduction du terme anglo-saxon *Multi-View Stereo* MVS). La donnée de sortie est un regroupement des pixels en régions, cohérentes avec les structures de la scène.

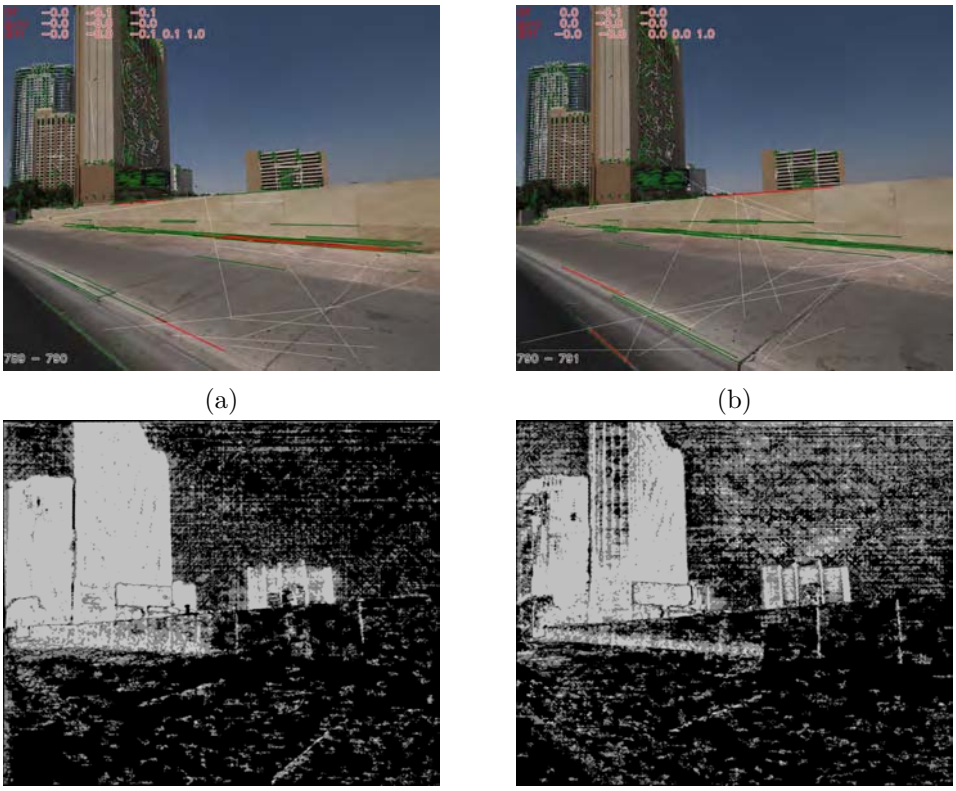
(a) (b)

FIGURE 2.2 – Reconstruction 3D éparse et dense, avec VisualSfm [Wu 11a] : ce résultat a été obtenu à partir d'une dizaine d'images d'une scène urbaine illustrée dans la figure 2.8.

Certaines données d'entrée peuvent notamment être acquises rapidement à l'aide du dispositif mobile `imajbox` que nous décrivons aussi. Pour les compléter, une reconstruction tridimensionnelle dense des points et des différentes poses de l'appareil peut être effectuée via des outils de vision par ordinateur (SfM+MVS) sous licence libre, largement utilisés dans le milieu de la recherche, comme VisualSfM [Wu 11a], voir figure 2.2. Ce programme combine la densification en appliquant *Patch-based Multi-view Stereo Software* (PMVS) [Furukawa 10] et l'optimisation du regroupement d'images avec *Clustering Views for Multi-view Stereo* (CMVS) ou encore CMPMVS [Jancosek 11]. De tels outils sont disponibles dans l'entreprise mais nous avons choisi ceux qui sont disponibles sous licence libre, évalués et utilisés dans le domaine de la recherche. Néanmoins, la présence de motifs répétitifs ou de grandes zones homogènes dans les images reste une difficulté lors de la reconstruction. La figure 2.3 montre les résultats obtenus avec les processus `Imajing` lors de la reconstruction éparse avec l'obtention d'une mise en correspondance de points d'intérêt et de la reconstruction dense avec l'estimation d'une carte de profondeur ou disparité [1]. Les données d'entrée disponibles sont décrites plus en détails dans la prochaine partie 2.4.

Notre approche est aussi fondée sur une représentation intermédiaire des images par des super-pixels, fournissant ainsi une sur-segmentation des images. Ce pré-traitement est essentiel

1. La disparité correspond à la norme du vecteur de déplacement d'un point d'une image à l'autre.



(a) (b)

(c) (d)

FIGURE 2.3 – Reconstruction 3D d'une scène à partir de deux images successives (fournie par `Imajing`) : Les images (a) et (b) représentent la mise en correspondance des points d'intérêt 2D obtenue lors de la reconstruction éparse. Les images (c) et (d) représentent les cartes de disparité obtenues lors de la reconstruction dense sur les mêmes images.

dans le sens où le problème de compréhension de la scène peut alors se formuler comme un problème de segmentation sémantique, c'est-à-dire un problème mixte de segmentation et de détection simultanées. Les travaux existants [Mori 04, Hoiem 05a, Gould 08] montrent l'importance et l'intérêt d'une telle approche, notamment concernant les temps de calculs. Dans notre cas, les constructeurs de super-pixels doivent préserver certains contours occultants des objets considérés car ce critère a une influence directe sur les performances finales [Hanbury 08]. Nous nous intéressons en particulier à l'intégration d'une information photométrique s'appuyant sur un *a priori* géométrique de planéité des surfaces dans la construction de super-pixels. Une vue d'ensemble de l'approche proposée est représentée dans la figure 2.4.

2.3 Approche hiérarchique

Une étape préliminaire à la segmentation sémantique est la définition et le choix d'une description sémantique adaptée à notre contexte, ainsi qu'au niveau de description que nous cherchons à obtenir. La description sémantique d'une image n'est pas intrinsèque à celle-ci. Elle dépend du but recherché et en particulier de ce que veut l'utilisateur. En effet, les travaux de [Boucher 05]

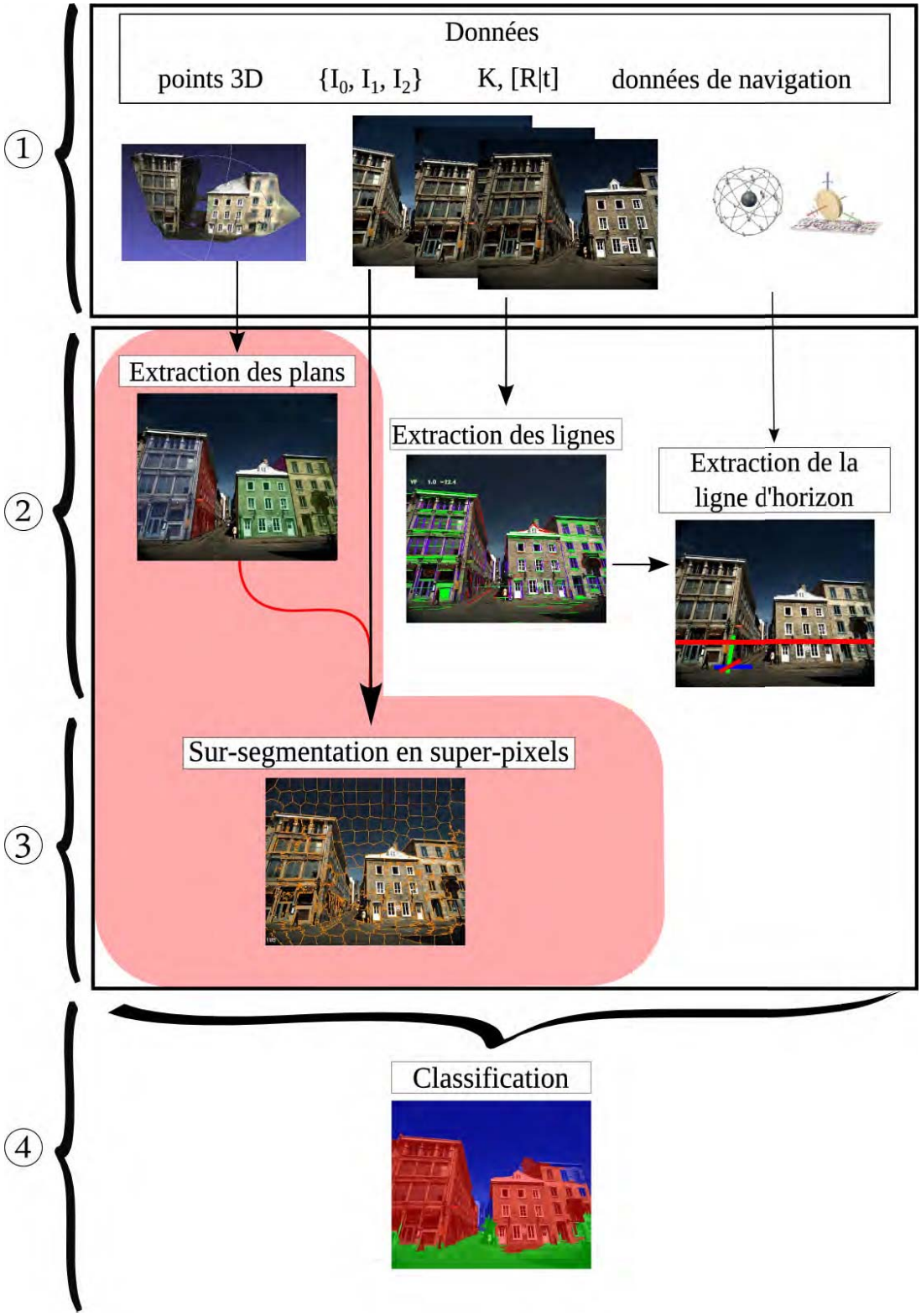


FIGURE 2.4 – Vue d'ensemble de l'approche proposée : ① les données d'entrée, ② extraction de l'information géométrique (plans, lignes et horizon), ③ sur-segmentation en super-pixels, ④ classification en trois classes (ciel, sol, objets verticaux). Nos contributions (→) portent sur l'intégration de l'hypothèse de planéité dans la construction de super-pixels.

proposent une approche pour extraire l'information sémantique d'une image à partir d'une approche de recherche d'image similaire. Ils mettent en avant la difficulté rencontrée en vision par ordinateur pour pallier le fossé sensoriel et sémantique. Le fossé sensoriel intervient au niveau de l'acquisition afin de comprendre la description numérique d'une image avant l'analyse bas-niveau de celle-ci. Le fossé sémantique, correspond au problème de mise en correspondance entre les caractéristiques bas-niveau extraites et les traitements haut-niveau souhaités. Comme le montre [Levy-Schoen 68], suivant la question posée à un observateur d'une image, le regard porté sur celle-ci sera variable en fonction de la requête, c'est pourquoi le choix d'une description sémantique adaptée est indispensable pour une bonne représentation.

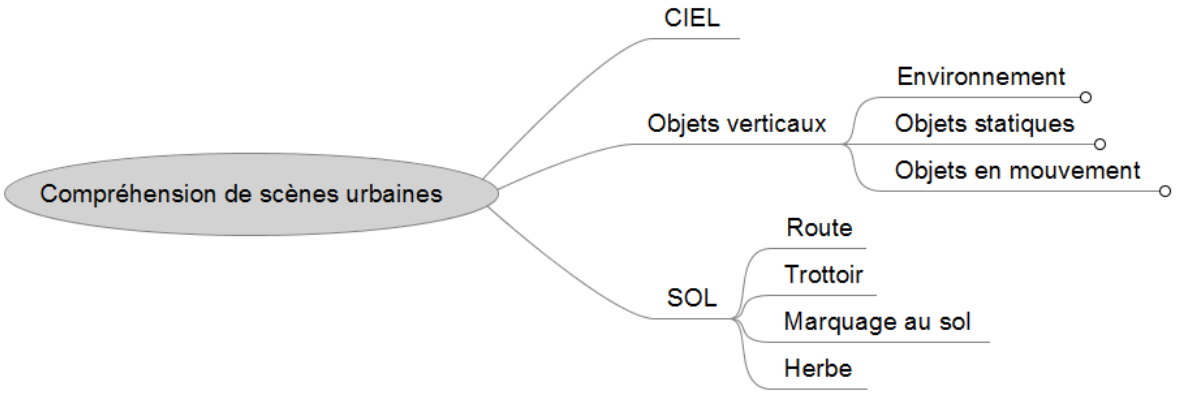


FIGURE 2.5 – Description sémantique hiérarchique adaptée au contexte de scènes urbaines.

Pour la description de scènes urbaines nous avons choisi d'utiliser une description sémantique hiérarchique, c'est-à-dire que pour chaque classe mère nous pourrons déterminer un ensemble de classes filles. Cela nous permettra de choisir le niveau de description, tout d'abord grossière afin d'aller vers une description plus raffinée adaptée au niveau de détails souhaités. Notre premier niveau de description sémantique correspond à trois classes : « sol », « ciel » et « objets », illustrées figure 2.5. La classe « objet » peut elle-même se décomposer en quatre sous-classes : « construction urbaine », « mobilier urbain », « végétation » et « autre », comme nous le montre la figure 2.17. Cette description est adaptée à notre contexte car elle permet de prendre en compte les principaux objets d'intérêt que nous souhaitons identifier. Par la suite, nous considérerons qu'un objet est vertical, s'il ne correspond ni à la classe « ciel » ni à la classe « sol » et qu'il existe une certaine relation spatiale entre l'objet et toute instance de la classe « sol ». Dans la classe des objets verticaux, il est possible de distinguer l'environnement comme les bâtiments et la végétation, des objets statiques ou en mouvement.

2.4 Données d'entrée et système d'acquisition imajbox

Dans un contexte applicatif, nous pouvons distinguer différentes possibilités sur les données disponibles et les pré-traitements réalisés, comme illustrés dans la figure 2.6. Plus précisément, nous pouvons considérés les sources d'information suivantes : capteur optique, capteur de profondeur, système de navigation qui peuvent être en interaction. De plus, grâce à des algorithmes de vision, il est possible de retrouver des informations de profondeur à partir d'acquisitions 2D.

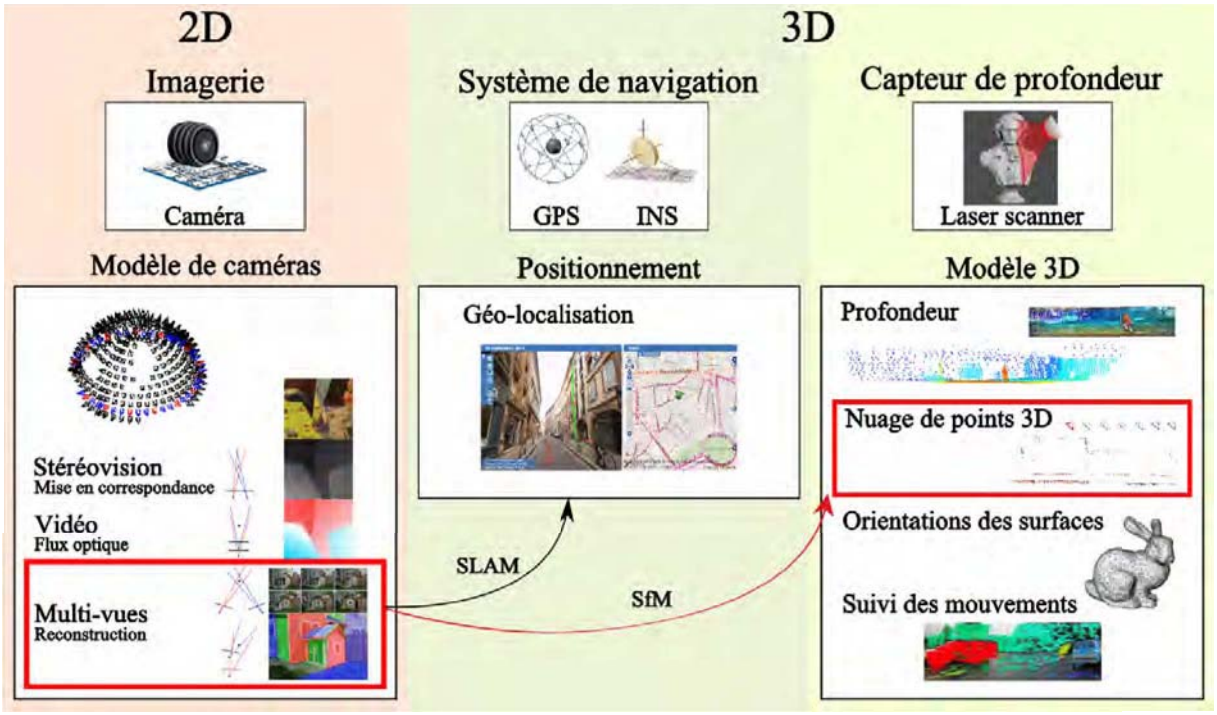


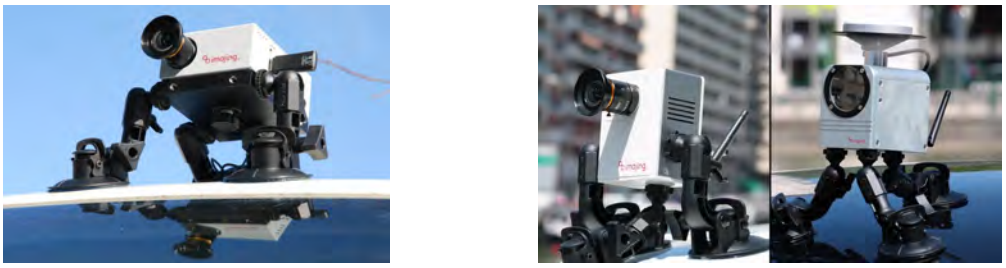
FIGURE 2.6 – Différentes sources de données d'entrée et de pré-traitements possibles : Dans ces travaux, nous nous appuyons sur les données d'`Imajing` : des images multi-vues avec un nuage de points 3D associé (cadres rouges dans le schéma).

Données d'entrée fournies par `Imajing`. Nous supposons que les données d'entrée dont nous disposons sont constituées :

- d'un nuage de points 3D, épars ou dense, correspondant à la reconstruction de la scène,
- des différentes poses et du calibrage intrinsèque de l'appareil d'acquisition,
- des mises en correspondance de points homologues entre paires de vues telles que la triangulation de deux points homologues coïncide avec un point 3D de la scène.

Les données sont acquises depuis un système en mouvement, à savoir l'`imajbox` qui est une solution de relevé portable de réseaux de transports depuis un véhicule mobile non-dédié (un train, une voiture, un bateau) proposée par la société `Imajing`. Les données disponibles font l'objet de certains pré-traitements utiles à une compréhension holistique de la scène.

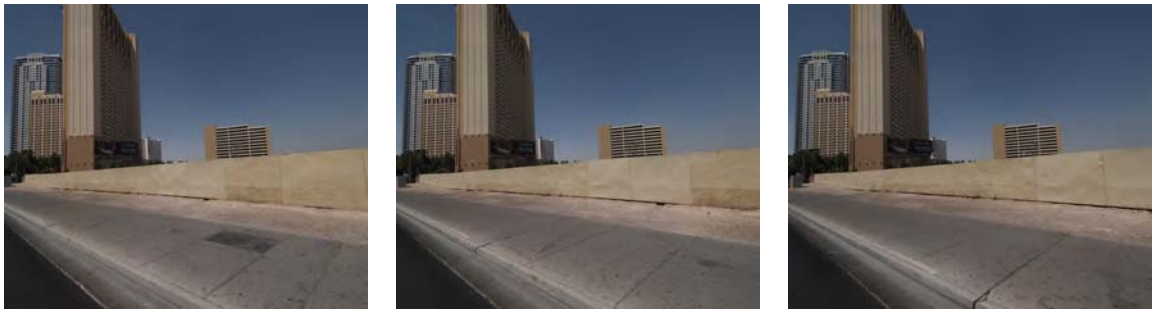
Système d'acquisition. L'`imajbox` présentée dans la figure 2.7 est le système d'acquisition propriétaire distribué par `Imajing`. Elle permet d'acquérir des séquences d'images géo-référencées, illustrées dans la figure 2.8. Elle est équipée d'un capteur CCD (*Charge-Coupled Device*) d'une fréquence maximale d'acquisition de 7 Hz et de taille 5Mpx (2050×2448), et d'un système de positionnement *Global Navigation Satellite System* (GNSS) composé d'une antenne *Global Position System* (GPS) et d'une centrale inertielle *Inertial Measurement Unit* (IMU) composée d'un gyroscope, d'un accéléromètre et d'une boussole.



(a) `imajbox 1` (b) `imajbox 2`

FIGURE 2.7 – Système mobile d'acquisition `imajbox` : ce système est utilisé dans le cadre de ces travaux. L'`imajbox 1` a été distribuée de 2009-2014. L'`imajbox 2` est disponible en 2015 en deux versions : compacte ou renforcée.

Ce système d'acquisition permet de collecter le long de la trajectoire du véhicule des images géo-référencées et orientées. Les données produites sont ensuite traitées par des algorithmes de navigation et de traitement d'images afin d'être visualisées, analysées et annotées par les utilisateurs avec le logiciel `imajview`, dont l'interface est visible dans la figure 2.9. Le service `imajnet` permet d'accéder aux données via internet. Cette chaîne de production, de traitement et de distribution permet de réaliser des inventaires d'infrastructures existantes et ainsi de gérer les équipements présents sur le terrain et visibles dans les images.



(a) (b) (c)

FIGURE 2.8 – Triplet d'images successives acquises par l'`imajbox` : en (b), il s'agit de l'image que nous appellerons centrale ou de référence, en (a) et (c) il s'agit des deux images dites adjacentes.

L'`imajbox` est un dispositif d'acquisition opportuniste. Les données acquises par ce système présentent des caractéristiques remarquables que nous détaillons dans ce qui suit.

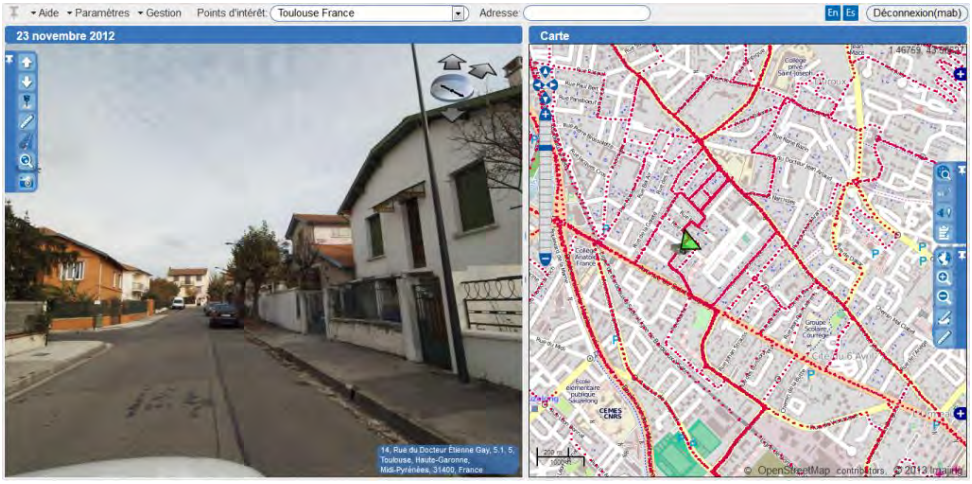


FIGURE 2.9 – Plateforme de web service `imajnet` : elle permet la gestion des projets avec, entre autres, la visualisation des séquences d'images et des traces de positionnement associées. Les points rouges représentent la position de chaque acquisition d'image et le triangle vert l'angle de vue de l'`imajbox`.

Caractéristiques des données étudiées. Les données traitées sont archivées dans la base de données `Imajing` qui présente deux particularités. La première est la variabilité (urbain, rural, route ou voix ferrée) et la diversité des données `Imajing` (2D, 3D, données inertielles et traces de navigation). Notons également que le nombre de données est conséquent, plus de 42 millions d'images disponibles, acquises depuis 2009.

Redondance d'information. Nous disposons d'informations redondantes dans l'espace et dans le temps comme illustré dans la figure 2.10. D'une part, la *redondance spatiale d information* provient du fait que les objets de la scène peuvent être visibles dans plusieurs images successives selon différents points de vue. D'autre part, lorsqu'une scène est relevée plusieurs fois, à différents moments dans l'année, cela fournit une *redondance temporelle*. Ainsi, dans le cas de multi-vues, l'information contenue dans une collection d'images représentant le même objet d'intérêt ou une même scène, selon différents points de vues, peut être utilisée afin de renforcer une cohérence photométrique.

Nous présentons dans ce qui suit des rappels de vision par ordinateur décrivant les fondements géométriques liés à la modélisation mathématique d'une (ou de plusieurs) prise(s) de vue et des relations inter-vues.

FIGURE 2.10 – Double redondance d'information spatiale et temporelle présente sur deux triplets d'images traités : Un objet d'intérêt, ici un panneau de signalisation encadré en rouge, est présent sur plusieurs images successives et sur plusieurs séquences.

2.5 Modélisation géométrique des prises de vue et rappels de vision par ordinateur

2.5.1 Modélisation géométrique d'une prise de vue

Nous invitons le lecteur à consulter l'ouvrage de [Hartley 04] pour plus de détails. Le modèle géométrique utilisé pour l'appareil photographique, désigné dans ce document par le terme *caméra*, est celui du sténopé et est illustré dans la figure 2.11. Ce dispositif optique simplifié dérive du principe de la chambre noire et est constitué d'un orifice de très faible diamètre représenté par un point appelé *centre optique*, qui fait office d'objectif. Dans ce document, une photographie sera désignée par le terme *image* lorsque cela ne prêtera pas à confusion avec le vocabulaire mathématique ou *vue*.

Nous utilisons les éléments suivants :

- Le plan de l'image est le plan d'équation $z = f$ sur lequel les points de l'espace se projette.
- Le repère de la scène permet de repérer les points de l'espace tridimensionnel et le repère de la caméra est centré sur le centre optique O, orienté vers le plan image et le coupe au niveau du point principal p ;
- Le repère pixélique de l'image et les coordonnées pixéliques permettent de positionner un point projeté dans le plan image.

D'un point de vue géométrique, la formation d'une photographie est décrite par la projection

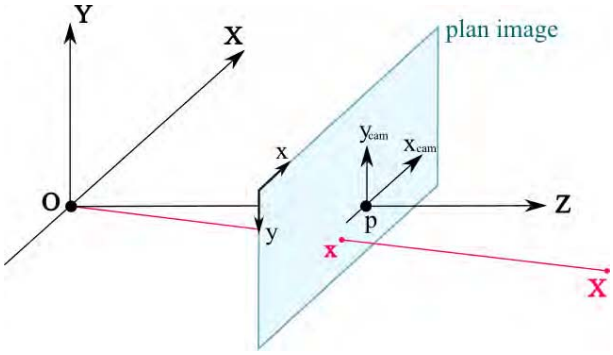


FIGURE 2.11 – Représentation du modèle sténopé : O est le centre optique et le point p est le point principal. Le centre optique correspond, ici, à l'origine de repère tridimensionnel. Le plan image est positionné en face du centre optique à une distance $OP = f$ correspondant à la distance focale. Le point 3D $\mathbf{X}$ se projette sur le plan image en $\mathbf{x}$.

centrale de l'espace 3D sur le plan 2D de l'image, à savoir par l'application qui, à tout point X de l'espace associe le point d'intersection de la droite, passant par le centre de la caméra C et X avec le plan de l'image.

Équation de projection centrale. En utilisant le formalisme de la géométrie projective, la projection centrale a pour équation en coordonnées homogènes :

$$\begin{pmatrix} \mathbf{x} \\ t \end{pmatrix} \sim \mathtt{P} \begin{pmatrix} \mathbf{X} \\ T \end{pmatrix} \tag{2.1}$$

où

- $\mathbf{X} = (X, Y, Z)^\top$ est le vecteur des coordonnées cartésiennes d'un point 3D pour $T = 1$,
- $\mathbf{x} = (x, y)^\top$ est le vecteur des coordonnées pixéliques de l'image de ce point pour $t = 1$,
- $\mathtt{P} \in \mathbb{R}_{3\times4}$ désigne la matrice de projection (en coordonnées homogènes),
- $\sim$ désigne l'égalité projective.

La matrice de projection se décompose sous la forme [Hartley 04, p. 156]

$$\mathtt{P} = \mathtt{K} \left[\mathtt{R} \mid \mathbf{T} \right] \tag{2.2}$$

où

- $\mathtt{K} \in \mathbb{R}_{3\times3}$ est une matrice triangulaire supérieure définissant les paramètres *intrinsèques* de l'appareil,
- $\mathtt{R} \in \mathbb{R}_{3\times3}$ est une matrice de rotation et $\mathbf{T} \in \mathbb{R}_3$ est un vecteur de translation définissant les paramètres *extrinsèques* de l'appareil.

On notera que, lorsque le repère de la scène coïncide avec le repère de la caméra, on a $\mathtt{R} = \mathtt{I}$ et $\mathbf{T} = \mathbf{0}$ dans (2.2).

La matrice des paramètres intrinsèques est appelée *matrice de calibrage* et est de la forme

$$\mathtt{K} = \begin{matrix} f & 0 & x_0 \\ 0 & f & y_0 \\ 0 & 0 & 1 \end{matrix} \tag{2.3}$$

Ses éléments sont

- la distance focale f (en pixels) et
- les coordonnées pixéliques (x_0, y_0) du point principal.

La matrice des paramètres extrinsèques $[\mathtt{R} \mid \mathbf{T}]$ est appelée *matrice de pose* de la caméra en rappelant que :

- le vecteur
$$\mathbf{C} = -\mathtt{R}^\top \mathbf{T}$$
contient les coordonnées cartésiennes du centre optique dans le repère 3D de la scène ;
- le vecteur $\mathtt{R}^\top (0, 0, 1)^\top$ représente la direction de l'axe optique.

Restriction de l'équation de projection centrale à un plan de la scène 3D. Soit (Π) un plan de la scène, sans perte de généralité, on se donne un repère de la scène tel que (Π) y ait pour équation $Z = 0$.

La restriction de la projection centrale au plan (Π) est une homographie de (Π) vers le plan de l'image (I), appelée perspectivité induite par (Π). Son équation s'écrit alors :

$$\begin{pmatrix} x \\ y \\ t \end{pmatrix} \sim \mathtt{KR} \overbrace{\begin{matrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{matrix}}^{\mathtt{H}} \mathbf{C} \begin{pmatrix} X \\ Y \\ T \end{pmatrix} \tag{2.4}$$

où $\mathtt{H} \in \mathbb{R}_{3\times 3}$ est la matrice de la perspectivité induite par (Π). Le point important à retenir ici est que cette application projective est bijective et donc que sa matrice $\mathtt{H}$ est inversible si l'axe optique ne se situe pas dans le plan (Π).

Point de fuite. Soit une direction de l'espace 3D non parallèle au plan de l'image, noté (I). Le *point de fuite* associé à la direction est le point où la droite passant par le centre optique C et de direction coupe (I).

Supposons que soit représentée par le vecteur $\mathbf{n} \in \mathbb{R}^3$, en utilisant le vocabulaire de la géométrie projective, on appelle *point à l'in ni* associé à la direction le point de vecteur de coordonnées homogènes $(\mathbf{n}^\top, T)^\top$ avec $T = 0$. De tels points obéissent aux lois de la projection centrale en se projetant via l'équation (2.1). Le résultat important est que le point de fuite associé à la direction est l'image par $\mathtt{P}$ du point à l'infini associé à cette même direction. On

montre très facilement que cette image ne dépend que de l'orientation (non de la position) de la caméra :

$$\begin{pmatrix}\mathbf{v}\\ t\end{pmatrix} \sim \mathtt{P}\begin{pmatrix}\mathbf{n}\\ 0\end{pmatrix} = \mathtt{K}\left[\mathtt{R} \mid \mathbf{T}\right]\begin{pmatrix}\mathbf{n}\\ 0\end{pmatrix} = \mathtt{KR}\mathbf{n} \tag{2.5}$$

où $\mathbf{v}$ est le vecteur des coordonnées pixéliques du point de fuite dans l'image pour $t = 1$.

Le lieu de tous les points à l'infini est un hyperplan remarquable de l'espace projectif (dans lequel est plongé l'espace affine considéré), appelé *hyperplan à l in ni.*

Ligne de fuite. Soit (Π) un plan de l'espace, non parallèle au plan de l'image (I) et ne contenant pas le centre optique C, et soit (Π_0) le plan parallèle à (Π) passant par C, en géométrie projective, la *ligne de fuite* du plan (Π) est la droite d'intersection du plan (Π_0) avec le plan de l'image (I). Une remarque importante est qu'aux directions parallèles à (Π) sont associées des points de fuite alignés dont le lieu est la ligne de fuite de (Π).

Soit une direction orthogonale à (Π) représentée par le vecteur $\mathbf{n} \in \mathbb{R}^3$ dans le repère de la caméra et soit $\mathbf{l}_\infty$ le vecteur des coordonnées homogènes de la ligne de fuite de (Π) dans l'image. On peut montrer le résultat important suivant :

$$\mathbf{l}_\infty \sim \mathtt{K}^{-\top}\mathbf{n}$$

En utilisant (2.5) avec $\mathtt{R} = \mathtt{I}$ et $\mathbf{T} = \mathbf{0}$, on en déduit que

$$\mathbf{l}_\infty \sim \mathtt{K}^{-\top}\mathtt{K}^{-1}\mathbf{v} \qquad \mathbf{v} \sim \mathtt{K}\mathtt{K}^{\top}\mathbf{l}_\infty \tag{2.6}$$

En d'autres termes, lorsque la caméra est calibrée ($\mathtt{K}$ est connue), la donnée de la ligne de fuite du plan (Π) est équivalente à celle de la donnée du point de fuite associé à la direction normale à (Π) .

Dans ce document, on appellera *ligne d horizon* la ligne de fuite associée au plan du sol.

2.5.2 Modélisation géométrique de deux prises de vue

On considère ici les images de deux caméras, dites gauche et droite, associées à deux matrices de projection notées $\mathtt{P}^g$ et $\mathtt{P}^d$. On notera par $\mathbf{x}^g = (x^g, y^g)^\top$ et $\mathbf{x}^d = (x^d, y^d)^\top$ les vecteurs des coordonnées pixéliques de deux points homologues dans les deux vues, c'est-à-dire images d'un même point 3D dont le vecteur des coordonnées cartésiennes est désigné par $\mathbf{X} = (X, Y, Z)^\top$ dans le repère de la scène.

Matrice fondamentale et géométrie épipolaire. La géométrie épipolaire traduit la relation projective (c'est-à-dire uniquement fondée sur des notions projectives comme l'incidence) entre deux caméras observant un même « objet » de la scène. Cette relation s'écrit, pour deux points homologues :

$$\left((\mathbf{x}^g)^\top, 1\right) \mathtt{F} \begin{pmatrix} \mathbf{x}^d \\ 1 \end{pmatrix} = 0 \tag{2.7}$$

où $\mathtt{F} \in \mathbb{R}^{3\times 3}$ est une matrice de rang 2, appelée *matrice fondamentale*.

Il existe donc une application projective, représentée par la matrice $\mathtt{F}$, qui associe à l'image d'un point dans la vue droite, de coordonnées cartésiennes (x^d, y^d), une *droite épipolaire* dans la vue gauche, dont le vecteur homogène est $\mathtt{F}(x^d, y^d, 1)^\top$: elle est l'image de la droite 3D passant par le centre optique de la caméra droite et le point 3D considéré. Dans le cas de la mise en correspondance algorithmique de points homologues, cette contrainte épipolaire est couramment utilisée. Il s'agit de contraindre la recherche du point 2D correspondant suivant la droite épipolaire correspondante.

En supposant que le repère de la scène coïncide avec le repère de la caméra gauche, les deux matrices de projections sont de la forme :

$$\mathtt{P}^g = \mathtt{K}^g \left[\mathtt{I}_{3\times 3} \mid \mathbf{0}_{3\times 1}\right] \text{ et } \mathtt{P}^d = \mathtt{K}^g \left[\mathtt{R} \mid \mathbf{T}\right] \tag{2.8}$$

La matrice fondamentale vérifie alors la décomposition suivante :

$$\mathtt{F} \sim (\mathtt{K}^d)^{-\top} [\mathbf{T}] \quad \mathtt{R} (\mathtt{K}^g)^{-1}. \tag{2.9}$$

où $[\mathbf{T}]$ est la matrice anti-symétrique associée au vecteur $\mathbf{T}$ tel que $\mathbf{T} \quad \mathbf{X} = [\mathbf{T}] \quad \mathbf{X}$, pour tout $\mathbf{X} \in \mathbb{R}^3$

L'équation (2.9) indique bien que la géométrie épipolaire est indépendante de la scène et donc seulement dépendante de l'orientation relative des deux caméras et du calibrage intrinsèque.

On appellera *matrice essentielle* la matrice fondamentale « calibrée » c'est-à-dire la matrice :

$$\mathtt{E} \sim (\mathtt{K}^d)^\top \mathtt{F} \mathtt{K}^g = [\mathbf{T}] \quad \mathtt{R}$$

Homographie induite par un plan 3D. Soit un plan (Π) de l'espace non plan de l'image (I) (et ne contenant pas le centre optique C), sans perte de généralité, on considère une nouvelle fois que (Π) a pour équation $Z = 0$ dans le repère de la scène. Les équations de projection d'un point situé dans (Π) s'écrivent pour les deux caméras :

$$\begin{pmatrix} x^g \\ y^g \\ 1 \end{pmatrix} \sim \mathtt{H}^g \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} x^d \\ y^d \\ 1 \end{pmatrix} \sim \mathtt{H}^d \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix}$$

où $\mathtt{H}^g \in \mathbb{R}^{3\times 3}$ et $\mathtt{H}^d \in \mathbb{R}^{3\times 3}$ sont les matrices des deux perspectivités induites par (Π).

Il existe donc une homographie inter-vues (de la caméra gauche vers la caméra droite) induite

par ce plan dont la matrice est :

$$\mathtt{H} \quad = \mathtt{H}^d(\mathtt{H}^g)^{-1}$$

2.5.3 Problème général de la reconstruction tridimensionnelle

Le problème que doit résoudre toute tâche de vision par ordinateur dont l'objet est de capturer intégralement ou en partie les données géométriques ou photométriques d'une scène, peut-être formulé comme dans [Hernández 10] « Étant donné une collection de photographies considérée comme donnée d'entrée, comment estimer la (l'ensemble des) forme(s)/modèle(s) tridimensionnelle(s) constituant la scène observée, qui générerait les mêmes photographies, sous les mêmes conditions de prises de vue, de propriétés de matériaux et d'éclairage. » Il est maintenant communément admis que l'obtention d'un tel modèle mathématique pour une scène pouvant être complexe et de taille très conséquente, photographiée par un grand ou petit nombre d'appareils, est le résultat du chaînage des deux modules de vision suivants :

- un module de « calcul de la structure et du mouvement » (que nous désignerons par le terme anglo-saxon *Structure-and/from-Motion*-SfM) prenant en entrée une collection non ordonnée de photographies, et fournissant une reconstruction éparse de la scène ;
- et un module de « stéréoscopie multi-vues » (que nous désignerons par le terme anglo-saxon *Multi-View Stereo*-MVS) prenant généralement en entrée les sortie du module SfM et fournissant en sortie un modèle géométrique dense, parfois maillé et très détaillé de la scène capturée, enrichi par des données décrivant son apparence photométrique de façon la plus réaliste possible.

Structure-from-Motion (SfM). Cette technique consiste à ordonner la collection de photographies puis à expliquer des trajectoires de primitives dites d'intérêts (points, droites etc.) au sein de la collection par un certain mouvement d'un (ou plusieurs) appareil(s) – dont sa trajectoire – et des caractéristiques géométriques tridimensionnelles de la scène. Un module de SfM fournit typiquement en sortie un nuage épars de points 3D ainsi que l'ensemble des poses d'un (ou des) appareil(s) et de leurs paramétrages intrinsèques associés. On parle de « *large scale SfM* » lorsque des milliers de photographies sont traitées.

Depuis quelques années, des capteurs spécialisés sont installés sur la plupart des dispositifs multimédia « grand-public » (smartphones, tablettes etc.) et sont en mesure de fournir, via un capteur GPS et/ou une centrale inertielle (IMU) composée d'un gyroscope, d'un accéléromètre et d'une boussole, respectivement le positionnement et l'orientation d'un dispositif et donc d'un appareil de prise de vues lorsqu'il est présent. Très récemment, de nouveaux capteurs, de type caméra *RGB-D* ou fonctionnant sur le principe du temps de vol (*Time of Flight*, TOF), sont disponibles sur les dispositifs multimédia et permettent d'associer à une photographie une information de profondeur qui restitue la troisième dimension. Il est évident que tous ces capteurs

peuvent être intégrés dans la chaîne de traitement des algorithmes classiques de vision par ordinateur pour la reconstruction 3D. Ainsi, ils peuvent fournir des informations *a priori* sur la pose de l'appareil et la scène initiale, ce qui facilite grandement la résolution du problème.

Multi-View Stereo (MVS). Ce terme désigne le problème qui consiste à générer un modèle géométrique de la scène, enrichi par des données décrivant son apparence photométrique, à partir d'une collection (non ordonnée) de photographies associées aux paramètres intrinsèques et extrinsèques de l'appareil (sous la forme des matrices de calibrage et de pose). Ce modèle géométrique doit être à la fois le plus complet et le plus précis possible, par exemple en terme de densité si le modèle est un nuage de points, et l'apparence doit être la plus réaliste possible. Une reconstruction 3D éparse de la scène peut compléter parfois les données d'entrée, comme toute information supplémentaire (ensemble de correspondances de primitives, de descripteurs visuels etc.) calculée par le module de SfM.

La chaîne de traitement de base d'un module de MVS peut être décrit schématiquement comme suit, dans une formulation proche de celle donnée dans [Hernández 10]. Dans une première étape, un nombre minimal de « cartes de photo-cohérence » est calculé décrivant à quel degré une partie de l'espace 3D est cohérent d'un point de vue photométrique pour un ensemble donné de vues/caméras. Dans la deuxième étape, le problème de la reconstruction 3D peut être assimilé à celui de la partition de l'espace 3D, en utilisant la photo-cohérence de ces cartes comme des contraintes, en vue d'obtenir des régions de « forme » et des régions « d'arrière-plan ». Dans l'étape finale, il s'agit de transférer des textures pour les régions de « forme » en déformant l'information de couleur associée à ces régions une fois reprojetées dans les vues considérées.

Les techniques de MVS diffèrent par les stratégies utilisées pour calculer les « cartes de photo-cohérence ». Les méthodes utilisant des facettes (*patches*), méthodes *patch-based*, génèrent des hypothèses de planéité globale ou locale du support 3D (comme dans [Furukawa 10, Xiong 15]). Cela permet de valider les meilleurs plans candidats, en testant la cohérence géométrique et photométrique entre vues adjacentes, en faisant l'hypothèse que la reprojection du voisinage d'un point de la facette correspond à un ensemble de pixels ayant des caractéristiques photométriques (intensité, couleur etc.) très similaires. Ces méthodes peuvent être divisées en deux catégories : celles qui utilisent des techniques de balayage en plans (plane sweeping) [Gallup 10] et celles utilisant des techniques de croissance à partir de germes (*seed and grow*), comme le logiciel PMVS [Furukawa 10].

CMPMVS. est un logiciel de reconstruction tridimensionnelle de scène s'appuyant sur un ensemble d'images rectifiées associé aux paramètres intrinsèques et extrinsèques de chaque caméra. Il permet de de construire un maillage texturé d'une scène rigide et visible dans les images, sachant que les objets en mouvement sont implicitement ignorés.

L'approche utilisée par CMPMVS peut se décomposer en deux étapes :

1. La reconstruction d'une carte de profondeurs [Jancosek 11] s'appuyant sur le balayage

en plans [Collins 96], permet d'estimer un ensemble de valeurs de profondeurs candidates pour chaque pixel de chaque prise de vue.

2. La reconstruction d'une maillage triangulaire [Jancosek 11] qui dans un premier temps fusionne l'ensemble des cartes de profondeurs afin d'obtenir un nuage de points 3D dense pouvant être vu comme un ensemble de tétrahèdres, puis, dans un second temps extrait la surface sous forme d'un maillage triangulaire en séparant l'espace, en zone "vide" ou "occupée", en résolvant un problème d'optimisation par graphe [Labatut 09].

Le point fort de CMPMVS est sa capacité à traiter les surfaces possédant un petit support, i.e. qui n'ont pas directement de support dans le nuage de points 3D, tout en fournissant une bonne qualité de représentation sur les autres surfaces.

2.6 Pré-traitements des données d'entrée

Dans notre approche, les données d'entrée sont pré-traitées afin d'extraire des informations de niveau intermédiaire (primitives géométriques 3D : plans, droites etc., régions 2D super-pixels etc.) facilitant la tâche à résoudre. Sachant que l'information de profondeur, de position et d'orientation des surfaces est implicitement ou explicitement encodée dans les données d'entrée, l'intégration d'une hypothèse sur la géométrie de la scène, telle que la planéité des surfaces, permet d'introduire de fortes contraintes entre images.

Nous supposons disposer d'une collection de vues-clé ordonnées temporellement, acquise par un capteur calibré installé sur un véhicule mobile, et d'un modèle (épars ou dense) de la scène. Ces données d'entrée doivent être pré-traitées afin d'extraire un niveau d'information intermédiaire. Nous décrivons ici, deux types d'informations géométriques utiles à la compréhension de scènes urbaines : la détection et l'estimation des points et lignes de fuite (dont la ligne d'horizon) dans les images ainsi que la détection et l'estimation des plans de la scène.

2.6.1 Points et lignes de fuite

Dans les différentes images, la géométrie de la scène est caractérisée en partie par le positionnement relatif des points de fuite associés à des directions connues de l'espace. Lorsque la caméra est calibrée, le vecteur des coordonnées homogènes d'un point de fuite et des coordonnées cartésiennes de la direction associée sont identiques à un facteur multiplicatif près. On comprend alors pourquoi la détection et l'estimation des points et lignes de fuite associés contribuent de façon déterminante à la compréhension de la scène. En particulier, ceci nous permet d'introduire les notions d'horizontalité et de verticalité, comme illustré dans la figure 2.12.

Tout d'abord, des primitives linéaires (segments de droites) sont extraites grâce au détecteur de droites LSD [Grompone von Gioi 12]. Un des grands avantages de ce détecteur est qu'il permet de contrôler le nombre de faux positifs. Les droites associées à ces primitives sont ensuite estimées puis associées à une direction de l'espace 3D. Au final, nous obtenons un ensemble

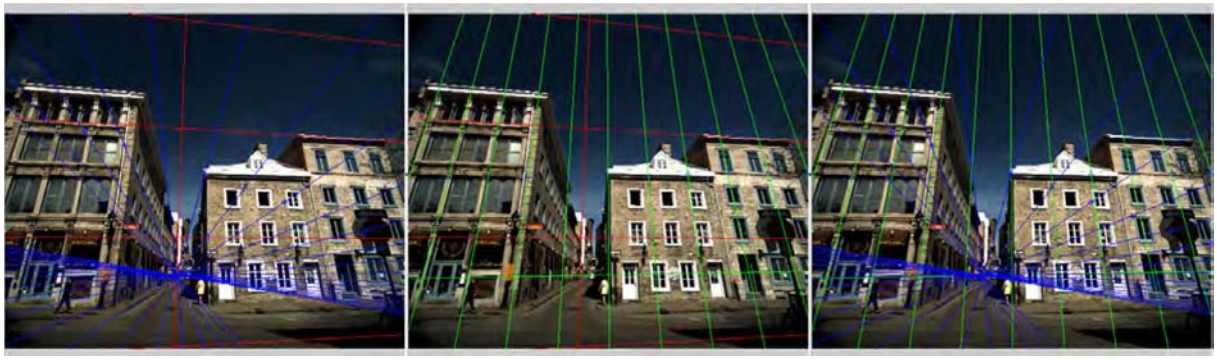


FIGURE 2.12 – Prolongement des droites d'une direction donnée jusqu'aux points de fuite associé (rouge, vert, bleu) : pour chaque image, nous considérons deux directions orthogonales.

de directions de l'espace et nous supposons qu'il contient les trois directions principales de l'espace, c'est-à-dire trois directions mutuellement orthogonales que l'on peut attacher aux axes principaux d'un repère 3D, cf. figure 2.13. Cela permet de renseigner sur la probable orientation des surfaces représentées dans l'image. Les travaux de [Tretyak 12] sont proches de ceux mis en place sur une image unique. Une première direction dominante est déterminée grâce à la méthode RANSAC [Fischler 81], détaillée dans l'annexe 4.7, en cherchant à la fois à maximiser le nombre de droites concourantes en un point de fuite (*inliers*) tout en minimisant la somme des distances des droites à ce point de fuite. La seconde direction est estimée de la même manière, en retirant les droites déjà utilisées pour la première estimation. Cependant le produit scalaire des vecteurs des deux directions doit être suffisamment petit pour garantir l'orthogonalité. Un simple produit vectoriel est utilisé afin de calculer la troisième direction afin de respecter l'hypothèse de Manhattan.

FIGURE 2.13 – Classification des lignes sur un triplet d'images : le repère monde a été estimé. Les lignes vertes sont associées à la première direction estimée, les violettes à la deuxième. Les lignes rouges n'ont pas été utilisées. Les trois directions estimées sont représentées sur un repère positionné devant la caméra.

Nous avons pu noter, de manière empirique, que l'un au moins des deux premiers points de fuite correspond très souvent à la direction verticale, colinéaire à la gravité.

Ligne d'horizon. La ligne d'horizon correspond à la droite de l'image qui est le lieu géométrique des points de fuite associés aux directions parallèles au plan du sol. Elle encode la géométrie affine des entités géométriques parallèles au plan du sol.

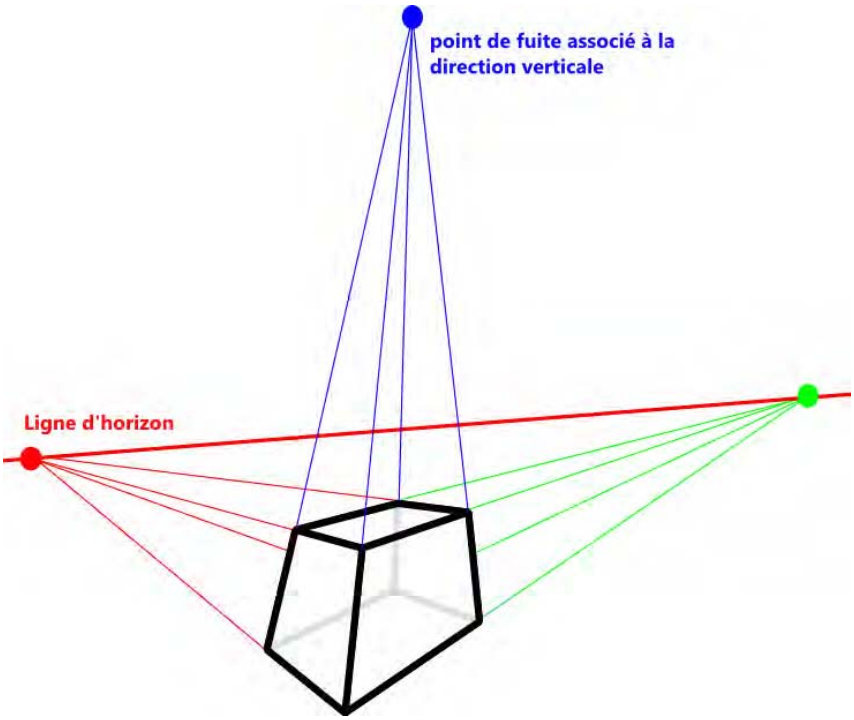


FIGURE 2.14 – Points de fuite et ligne d'horizon associée à la direction horizontale.

Lorsque le système d'acquisition est calibré, c'est-à-dire lorsque la matrice de calibrage $\mathtt{K}$ est connue (on notera que l'on peut toujours se ramener au cas où $\mathtt{K} = \mathtt{I}$), la ligne d'horizon encode la géométrie affine de l'espace 3D. En effet, l'équation (2.6) montre que la ligne de fuite d'un plan est en relation directe (dite de polarité) avec les paramètres intrinsèques de la caméra et le point de fuite associé à toute direction parallèle à la normale au plan. En particulier, l'équation (2.10) permet d'obtenir le point de fuite $\mathbf{v}$ associé à la direction verticale à partir de la ligne de fuite $\mathbf{l}_\infty$, comme illustré dans la figure 2.14 :

$$\mathbf{l}_\infty = (\mathtt{K}\mathtt{K}^\top)^{-1}\mathbf{v} \tag{2.10}$$

Inversement, connaissant la direction verticale, nous pouvons alors estimer la ligne d'horizon comme nous le montre la figure 2.12. Il est possible de rétro-projeter cette droite dans l'espace pour obtenir l'ensemble des plans parallèles à celui du sol. Ceci nous permet d'obtenir une connaissance *a priori* sur les notions de verticalité et d'horizontalité.

Ces deux orientations sont indispensables afin d'estimer les relations spatiales entre les objets présents dans la scène, sachant que les murs/façades des bâtiments sont construits verticalement c'est-à-dire perpendiculairement au plan du sol.

2.6.2 Plans et homographies

Dans notre contexte de compréhension de scènes urbaines, nous faisons l'hypothèse d'une scène partiellement plane par morceaux. Un des pré-traitements à réaliser est de détecter la présence d'entités planes dans le nuage de points représentant la scène et d'estimer les paramètres des plans associés. Ce problème, dit de segmentation en plans, prend une place importante dans la littérature [Bartoli 07, Sinha 08, Mičušík 09, Gallup 10] car les entités planes sont couramment représentées dans ce type de scène.

Segmentation en plans. Il existe schématiquement trois types d'approches de segmentation en plans :

- les approches purement géométriques,
- les approches purement photométriques,
- les approches conjointes.

Les approches purement géométriques, de type RANSAC [Fischler 81], ont été utilisées dans [Bartoli 07, Zhao 11], et étendues dans [Fouhey 13] avec la prise en compte de l'hypothèse de modèles multiples. Les approches purement photométriques sont plus orientées vers des algorithmes de mise en correspondance en stéréovision comme dans [Mičušík 09, Furukawa 09, Gallup 10]. La redondance d'informations disponibles dans nos données nous permet de nous orienter vers des approches conjointes comme la méthode de segmentation en plans de [Bartoli 07], qui combine l'approche RANSAC avec un critère de cohérence photométrique inter-images. Dans ces travaux, il est mis en parallèle le fait de travailler avec des ensembles de données disjointes où un point 3D appartient à un plan unique, avec le fait de travailler avec des données couvrantes où un point 3D peut appartenir à plusieurs plans. Ce cas particulier se produit à l'intersection de deux plans sécables dans l'image. L'avantage de travailler avec des données couvrantes est d'estimer de manière plus précise les plans et donc d'obtenir une meilleure estimation de l'intersection des plans. Cependant, dans le chapitre 3 nous ne combinons pas les approches géométrique et photométrique afin d'estimer les plans mais nous utilisons les paramètres des plans estimés afin de valider ou d'infirmer leur planéité tridimensionnelle à l'aide d'un critère photométrique.

Critères de recouvrement de plans. Des mesures permettent de comparer et quantifier le taux de recouvrement de deux ensembles, Π_i et Π_j. Dans le cas d'une mesure de similarité, si les deux ensembles de données sont proches alors la mesure est élevée, si les ensembles sont disjoints la mesure est faible. Deux exemples sont les mesures de Jaccard et de Dice. La mesure de dissimilarité de Jaccard (2.12) donne $J(i,j) = 1$ dans le cas $\Pi_i \cap \Pi_j = \emptyset$ et est utilisée dans [Toldo 10]. La mesure de similarité de Dice (2.11) est utilisée dans l'article [Bartoli 07] et vaut $D(i,j) = 0$ si $\Pi_i \cap \Pi_j = \emptyset$.

$$D(i,j) = \frac{2.\#(\Pi_i \cap \Pi_j)}{\#\Pi_i + \#\Pi_j} \tag{2.11}$$

$$J(i,j) = \frac{\#(\Pi_i \cup \Pi_j) - \#(\Pi_i \cap \Pi_j)}{\#(\Pi_i \cup \Pi_j)} \tag{2.12}$$

où $\#(A)$ est le cardinal (nombre d'éléments) de l'ensemble A.

Il est pertinent de chercher le meilleur support spatial associé à chaque plan. Dans [Bartoli 07], il est supposé que le support photométrique Π' est toujours inclus dans le support géométrique Π. Ceci se produit lorsque les points d'intérêt appariés utilisés se situent sur les contours, bordures des zones à délimiter.

Homographie inter-vues induite par un plan 3D. Un des avantages de l'hypothèse de planéité est qu'elle permet de transférer les textures des régions correspondant aux projections d'une entité plane d'une vue vers une autre, via l'homographie inter-vues induite par ce plan, cf. 2.5.2. Notre approche propose d'affiner le support de chaque plan. Pour cela, dans une zone supposée plane, l'information géométrique est analysée à partir d'un critère évaluant la « ressemblance photométrique » de la région respective dans la première image avec la région homologue dans la seconde image après transformation homographique adéquate. Un exemple d'une image de référence I et de l'image recalée $\tilde{I}$ est illustrée figure 2.15. Ce transfert sera utilisé dans des étapes ultérieures de notre approche. Une homographie inter-vues induite par un plan est complètement définie par la pose relative des appareils associée aux deux vues, le calibrage intrinsèque et les paramètres du plan dans le repère 3D du premier appareil. En d'autres termes, il est possible de calculer sa matrice à partir des données d'entrée et des paramètres du plan considéré.

Si on suppose que le repère de la scène coïncide avec le repère de la première caméra, les matrices de projection des deux caméras sont de la forme (2.8) et l'on montre que $\mathtt{H}$ admet la décomposition suivante

$$\mathtt{H} \sim \mathtt{K}^d(\mathtt{R} - \frac{1}{d}\mathbf{T}\mathbf{n}^\top)(\mathtt{K}^g)^{-1} \tag{2.13}$$

où $(\mathbf{n}^\top, d)^\top$ est le vecteur homogène du plan Π dans le repère euclidien de la caméra gauche.

Nous avons mis en place plusieurs méthodes d'estimation d'homographie, qui sont détaillées dans [Hartley 04, chap.13 p.326-338], à partir de :

- l'équation du plan $(aX + bY + cZ + d = 0)$ défini par trois points 3D et la matrice de projection $\mathtt{K}[\mathtt{R} \mid \mathbf{T}]$;
- trois points 2D mis en correspondance $x'_i \quad x_i, \forall i \in 1,2,3$, de la matrice fondamentale $\mathtt{F}$ et de l'épipole e ;
- une ligne et un point 2D mis en correspondance, de la matrice fondamentale $\mathtt{F}$ et de l'épipole e.

De manière empirique, nous avons remarqué que l'approche mettant en correspondance une

(a) Zone originale de I (b) Zone rectifiée de $I' : \tilde{I} = H(I')$

FIGURE 2.15 – Rectification d'image par l'homographie : Plus précisément, nous recherchons un critère photométrique dense entre l'image de référence I et l'image adjacente rectifiée par l'homographie $\tilde{I} = H(I)$ induite par le plan de support.

ligne et un point est moins sensible aux problèmes d'erreurs de paramétrage pour réaliser la rectification des zones planes de l'image adjacente sur les zones homologues de l'image de référence. Les deux images, image de référence 2.15a et image adjacente homographiée 2.15b, sont similaires sur la zone plane utilisée et différentes au niveau des objets situés en retrait de celui-ci. Ainsi, il est possible, par comparaison, d'extraire une information photométrique dense cohérente à la géométrie de la scène et utile à l'intégration de l'information géométrique.

2.7 Méthodologie utilisée

Les données à traiter sont riches et variables. Les données inertielles semblent non fondamentales dans notre approche car elles peuvent être substituées par des approches de vision. Par exemple, le processus d'estimation de la ligne d'horizon peut être effectué par vision pure. Nous pouvons noter que la connaissance de l'orientation des surfaces facilite la compréhension de la scène. Pour cela, la densification des nuages de points 3D épars semble nécessaire.

Nous travaillons avec des entités de niveau intermédiaire possédant un support spatial fermé (triangle ou super-pixel). De plus, nous avons choisi d'utiliser une hypothèse de planéité par morceaux afin d'extraire une information locale. L'estimation des plans s'effectue sur le nuage de points 3D.

Le calcul des homographies permet, d'une part, une visualisation virtuelle de l'image adjacente du point de vue de l'image de référence et, d'autre part, de comparer ces deux images, l'image de référence et l'image adjacente recalée, avec un critère de cohérence photométrique inter-image, comme les *Image Quality Assessment* (IQA).

Pour chaque zone, l'hypothèse de planéité est validée ou infirmée en fonction de la valeur de l'IQA choisie. L'approche de sur-segmentation que nous proposons intègre les informations de planéité et de cohérence photométrique inter-image, afin d'obtenir des entités cohérentes avec la géométrie de la scène.

En combinant l'information, en vue de l'intégrer au niveau de la construction des super-pixels, nous cherchons à minimiser les risques de fausse sur-segmentation afin de moins influencer la classification sémantique. Dans la suite, nous présentons la stratégie d'évaluation mise en place afin de comparer nos travaux à l'existant.

2.8 Évaluation

L'évaluation permet de qualifier, quantifier la qualité ou valider des résultats obtenus après l'application d'un algorithme donné. Elle dépend de trois caractéristiques principales : des paramètres de l'algorithme utilisé, de la base de connaissances exploitée et de l'algorithme lui-même. Dans [Rosenberger 06], l'auteur présente la problématique de l'évaluation d'algorithmes de traitement d'images. Cette évaluation peut se concevoir à différents niveaux : qualitatif (bon, moyen, mauvais), quantitatif (à l'aide d'un taux ou d'une valeur), fonctionnel (répond t-il à l'objectif ?) et robuste (le résultat est-il fiable en présence de données aberrantes ?). Par conséquent, il existe plusieurs méthodologies :

- l'évaluation des performances d'un critère caractéristique optimisé,
- l'évaluation par adéquation qui ne permet qu'une validation fonctionnelle de l'algorithme,
- l'évaluation par diagnostic qui permet de comparer une segmentation de référence, appelée vérité terrain, avec les résultats obtenus et ainsi d'évaluer la performance de l'algorithme.

La littérature sur le sujet distingue trois types de critères d'évaluation :

- les critères supervisés utilisés lorsque nous disposons d'une vérité terrain exhaustive ;
- les critères non supervisés choisis sans aucun *a priori* et à partir des informations intrinsèques ;
- les critères hybrides qui combinent une vérité terrain qui peut être partielle avec des informations intrinsèques.

Dans ces travaux, nous nous intéressons en particulier aux méthodes d'évaluation utilisant des critères supervisés car nous avons réalisé des vérités terrain, éléments de référence, pour nous permettre de quantifier nos résultats en différenciant les critères d'évaluation pour les approches de classification, de celles utilisées pour les approches de segmentation.

2.8.1 Corpus

Le corpus d'images représentatives des variabilités de la base de données `Imajing` présente un ensemble d'images de référence, dont un sous-ensemble est représenté dans la figure 2.16. Pour chaque image de ce corpus nous réalisons manuellement une vérité terrain, image segmentée sémantiquement par un agent, qui servira de référence lors de l'évaluation.

(a) Espagne (b) Andorre (c) Canada

(d) Toulouse lieu n°1 (e) Toulouse lieu n°1 (f) Los Angeles

(g) Los Angeles (h) Paris lieu n°1 (i) Paris lieu n°2

FIGURE 2.16 – Corpus d'image en zones urbaines denses : Images avec (a,e,h) et sans (c,g,i) véhicule en mouvement, avec (b,d) et sans (f,i) végétation, (d,e) images d'un même lieu en hiver et au printemps.

Plusieurs outils de segmentation manuelle sont disponibles dans l'état de l'art. Par exemple, *Graphic Annotation Tool* (GAT) [2] est un outil d'annotation manuelle. Ce logiciel utilise la notion d'arbre binaire. L'interface web LabelMe [3] propose une grande base de données segmentées manuellement par l'utilisateur. Ces données sont de plus en plus fiables car après avoir été

2. `http://upseek.upc.edu/gat/`
3. `http://labelme.csail.mit.edu/Release3.0/index.php`

annotées, elles sont validées par des experts. *Object Labeling Tool* (OLT) [4] est un outil MATLAB adaptatif au contexte qui, en plus d'attribuer à chaque zone de l'image un nom d'objet, permet de gérer l'organisation spatiale de la scène, en attribuant une position relative aux objets en fonction de leur proximité à la caméra.

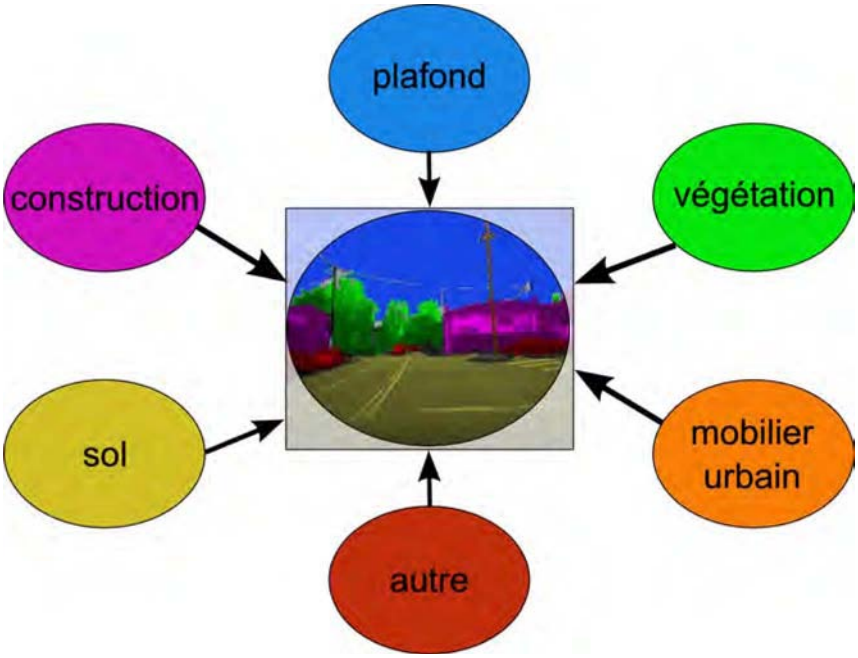


FIGURE 2.17 – Exemple de vérité terrain manuelle pour une segmentation sémantique sur une image de scène urbaine : chaque couleur représente une famille d'objets.

Nous avons choisi d'utiliser ce dernier outil, pour sa simplicité d'adaptation afin d'obtenir les segmentations manuelles souhaitées. Ce programme OLT permet de segmenter manuellement en polygones les objets présents dans l'image et d'annoter chaque polygone avec le nom de notre choix. L'ensemble des objets segmentés est modifiable. Nous pouvons modifier les objets, ajouter ou supprimer un objet ou une partie d'objet, le déplacer vers l'avant-plan ou l'arrière-plan relativement à leur proximité.

Nous avons apporté deux améliorations par rapport à OLT, illustrées dans la figure 2.18. La première est l'introduction d'une étape éventuelle de sur-segmentation qui permet d'obtenir une résolution de notre vérité terrain au niveau du pixel. La sur-segmentation mise en place est celle proposée dans [Felzenszwalb 04]. La seconde amélioration est de pouvoir choisir le nom, l'étiquette de l'objet segmenté à partir de la liste des classes proposées dans la description sémantique et hiérarchique présentée précédemment, dans la figure 2.5.

Ces opérations ont été réalisées sur les images du corpus afin d'obtenir une vérité terrain sur un ensemble représentatif, comme nous pouvons le voir sur la figure 2.19.

4. `http://www.cs.illinois.edu/homes/dhoiem/`

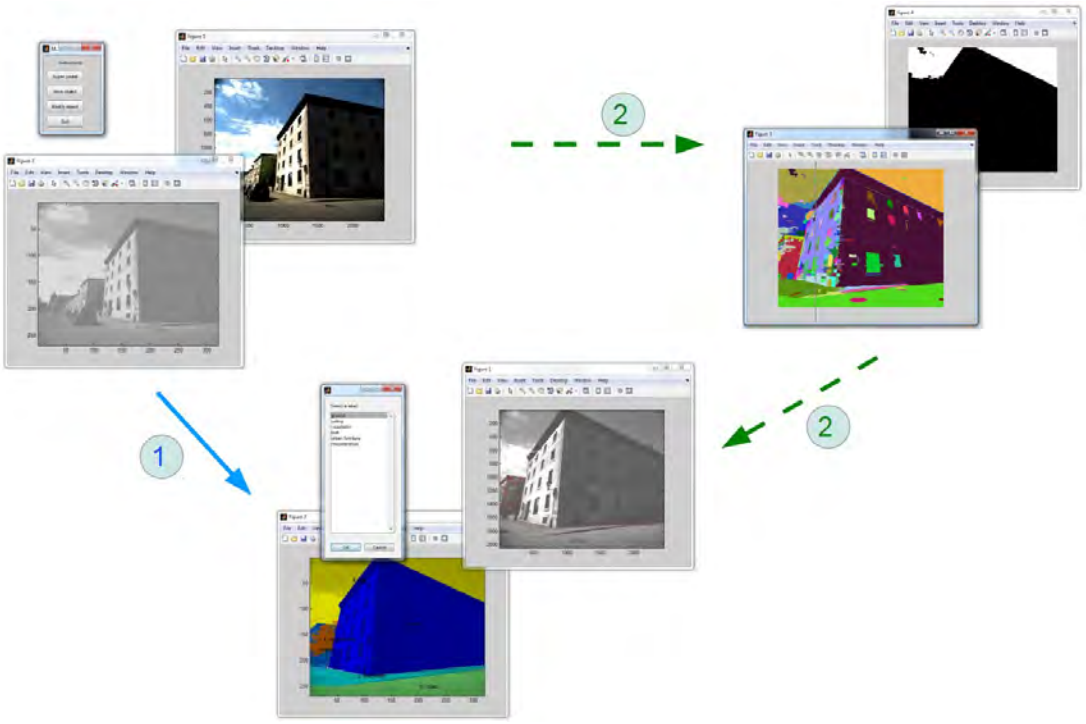


FIGURE 2.18 – Interface utilisateur pour la segmentation manuelle : OLT amélioré. **En haut à gauche :** Image originale en couleur et masque de segmentation en niveau de gris accompagné du menu des actions possibles. **En bas :** résultat de la segmentation avec proposition de la description sémantique lors de l'ajout d'un objet. **En haut à droite :** sur-segmentation proposée afin d'obtenir une précision au niveau du pixel.

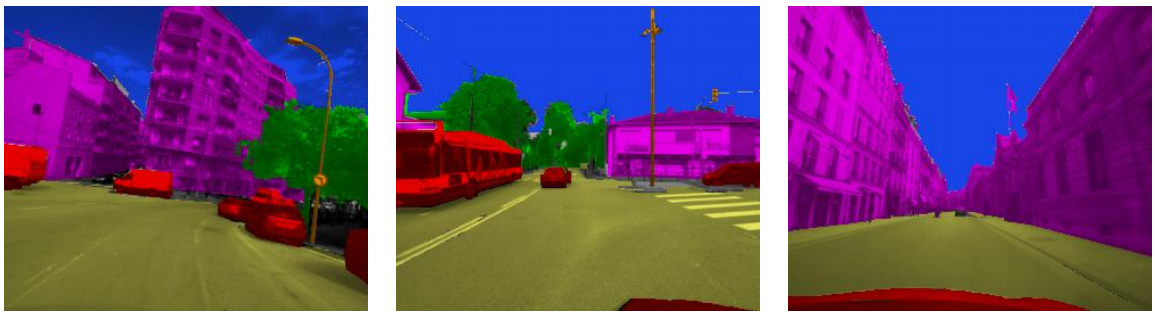


FIGURE 2.19 – Segmentation manuelle effectuée sur trois images de référence du corpus : les éléments appartenant à la même classe sémantique présentée dans la figure 2.17 sont représentés de la même couleur : plafond (bleu), construction (violet), mobilier urbain (orange), végétation (vert), sol (jaune) et autre (rouge).

2.8.2 Bases de données

La variabilité des données disponibles dans les bases de données existantes donne une idée des enjeux, en vision par ordinateur, des approches multi-vues pour la compréhension de scènes urbaines. Le site internet de *Yet Another Computer Vision Index To Datasets* (YACVID) [5] fournit une liste des bases de données existantes, parmi les plus utilisées. Entre autres, nous

5. `http://riemenschneider.hayko.at/vision/dataset/`

pouvons citer :

- **VGG-Oxford -** Les données de *Visual Geometry Group*[6] de l'université d'Oxford sont un ensemble d'images associées à de l'information géométrique 2D (points d'intérêt et segments de droite mis en correspondance) et 3D (pose des caméras, points 3D, segments de droite). Nous avons utilisé en particulier les données Valbonne church (15 images - 512×768), Merton College III (3 images - 1024×768) et Wadham college (5 images - 1024×768). Pour les deux derniers ensemble, nous avons construit la vérité terrain associée à une image dite de référence.
- **ZuBuD -** L'école polytechnique fédérale, ou ETH-Zürich gère les données de la ZuBuD [Shao 03], pour *The Zurich Building Dataset*. Cette base de données regroupe 5 vues de 201 bâtiments de Zürich, soit 1005 images de résolution 640×480. Chaque acquisition diffère du point de vue mais il y a peu de variation d'illumination car les images d'un bâtiment sont prises au même moment de la journée.
- **Daimler -** La base de données *The Daimler Urban Segmentation Dataset*[7] fournit des séquences vidéo en trafic urbain. Elle contient 500 paires d'images en niveau de gris stéréoscopiques rectifiées d'une résolution de 1024×440. Parmi les 500 images, une image toute les 10 est annotée manuellement au niveau pixel en 5 classes sémantiques : sol, bâtiment, véhicule, piéton et ciel. Les cartes de disparité associées sont fournies. Elles ne sont cependant pas obtenues manuellement mais en utilisant une mise en correspondance semi-globale.
- **Kitti -** La plateforme d'acquisition est équipée de 4 caméras haute résolution 1242×375, d'un laser scanner Velodyne et d'une solution de navigation. Le corpus de référence KITTI [8] contient 389 paires d'images stéréoscopiques acquises à haute fréquence. Cette base de données est la deuxième référence pour l'évaluation en vue de la conception de la voiture autonome [Geiger 12]. De nombreux domaines de la recherches sont étudiés : mise en correspondance stéréoscopique, flux optique, odométrie, reconnaissance et suivi d'objets d'intérêt (piéton, voiture, route).

Pour chacune de ces cinq bases de données une image représentative est illustrée dans la figure 2.20. Les données d'Oxford et de ZuBuD présentant des images multi-vues de scènes urbaines sont proches des images d'`Imajing` car elle ont des distances inter-image correspondant à une grande *baseline*. Il nous paraît important de citer les deux bases de données proposées en partenariat avec les constructeurs automobiles en vue de la voiture autonome : Daimler (Mercedes-Benz) et Kitti (Toyota Technological Institute at Chicago). Toutefois, nous n'utiliserons pas ces deux dernières bases de données car les acquisitions réalisées ne nous permettent pas d'estimer une reconstruction 3D éparse à partir d'approches telles que SfM ou MVS. De plus, le système d'`Imajing` ne bénéficie pas de ce type de données, à savoir un nuage de points

6. `http://www.robots.ox.ac.uk/~vgg/data/`
7. `http://www.6d-vision.com/scene-labeling`
8. `http://www.cvlibs.net/datasets/kitti/`

3D structurés, obtenu avec un laser. Il ne s'agit donc pas du même contexte applicatif.

(a) Daimler (1024 × 440)

(b) Kitti (1242 × 375)

(c) ZuBuD (640 × 480)

(d) Oxford - Merton College III (1024 × 768)

(e) `Imajing` (2448 × 2050)

FIGURE 2.20 – Exemples d'images des cinq bases de données Daimler, Kitti, Oxford, ZuBuD et `Imajing` : les caractéristiques sont synthétisées dans le tableau 2.1. Dans nos évaluations seules les bases (c) à (e) ont été testées. La base (c) a été évaluée qualitativement (puisque nous ne disposons pas de vérité terrain) alors que les bases (d) et (e) ont pu être évaluées quantitativement.

Le tableau 2.1 propose une comparaison des bases de données existantes avec les données Imajing. Les vérités terrain effectuées sur les données de référence permettent d'évaluer les algorithmes testés suivant différents critères présentés dans ce qui suit.

2.8.3 Critères d'évaluation de la classification

Pour évaluer la performance d'un classifieur, nous avons choisi d'utiliser les résultats fournis par les matrices de confusion, ainsi que par deux méthodes applicables aux classificateurs binaires, la courbe de Précision-Rappel (PR) et la courbe *Receiver Operator Characteristic* (ROC) qui fournissent deux types d'informations complémentaires. Des détails sur les liens entre ces deux courbes sont données dans [Davis 06].

	Daimler	Kitti	ZuBuD	Oxford	Imajing
Images stéréoscopiques	x	x			
Flux vidéo	x	x			
Multi-vues			x	x	x
GNSS/IMU		x			x
Laser		x			
Profondeur	x				
Vérité terrain	x	x			

TABLE 2.1 – Comparaison des bases de données utilisées dans les approches de compréhension de scènes urbaines.

Matrice de confusion pour la classification binaire. Nous considérons tout d'abord une classification des pixels en deux classes : l'une, dite « classe positive » et notée V_1, correspond à l'avant-plan (là où se trouve l'objet recherché) ; l'autre, dite « classe négative » et notée V_0, correspond à l'arrière-plan.

Par la suite, en se référant à la classe V_1, on appellera

- vrais positifs, *True Positives* (TP) : les pixels étiquetés à raison dans cette classe,
- vrais négatifs, *True Negatives* (TN) : les pixels étiquetés à raison dans l'autre classe,
- faux positifs, *False Positives* (FP) : les pixels étiquetés à tort dans cette classe,
- faux négatifs, *False Negatives* (FN) : les pixels étiquetés à tort dans l'autre classe.

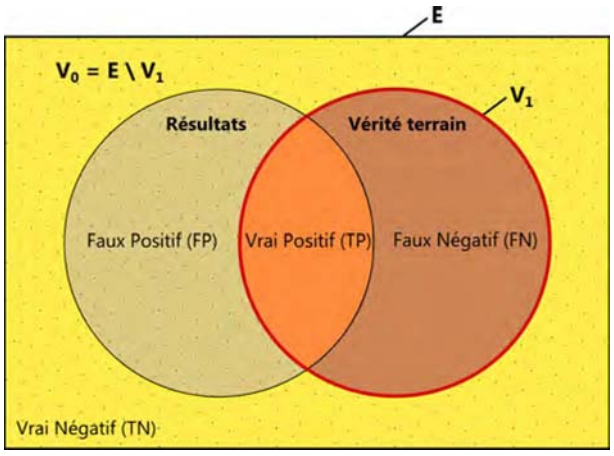


FIGURE 2.21 – Représentation graphique de la vérité terrain et du résultat d'un classifieur binaire. La « classe positive » (avant-plan), notée V_1, correspond au cercle de droite. La « classe négative » (arrière-plan), notée V_0, correspond au rectangle englobant privé du cercle V_1.

Dans le cas de deux classes, et dans ce cas uniquement, les vrais positifs de V_0 sont les vrais négatifs de V_1 et les les faux positifs de V_0 sont les faux négatifs de V_1. Toutes ces différentes quantités, calculées à partir du résultat obtenu par une classification que l'on compare à une

Résultat \ Vérité terrain	Étiquette Positive	Étiquette Négative
Décision Positive	**Taux Vrais Positifs / Rappel (R)**	**Taux Faux Positifs (FPR)**
Décision Négative	**Taux Faux Négatifs (FNR)**	**Taux Vrais Négatifs / Rappel Négatif (NR)**

TABLE 2.2 – Matrice de confusion pour l'évaluation des performances d'un classifieur binaire.

carte de référence (vérité terrain), peuvent être représentées graphiquement comme dans la figure 2.21.

Divers critères d'évaluation peuvent être définis à partir des quantités que nous venons d'introduire. Deux mesures communément utilisées sont la précision et le rappel.

- La précision, *precision* (Pr), mesure la capacité du classifieur à ne pas faire de mauvais étiquetages. On définit la précision comme la proportion d'étiquetages faits pour une classe qui sont corrects, c'est-à-dire comme le nombre de vrais positifs divisé par le nombre total de pixels étiquetés dans cette classe par le classifieur :

$$Pr = \frac{TP}{TP + FP} \tag{2.14}$$

- Le rappel, *recall* (R), mesure la capacité du classifieur à bien étiqueter tous les éléments d'une classe. On définit le rappel comme la proportion des étiquetages corrects dans cette classe, c'est-à-dire comme le nombre de vrais positifs divisé par le nombre total de pixels étiquetés dans cette classe par la vérité terrain :

$$R = \frac{TP}{TP + FN} \tag{2.15}$$

On définit aussi de façon similaire :

- la précision négative, *Negative Precision* (NPr)

$$NPr = \frac{TN}{TN + FN} \tag{2.16}$$

- le rappel négatif, *Negative Recall* (NR)

$$NR = \frac{TN}{TN + FP} \tag{2.17}$$

- le taux de faux positifs, *False Positive Rate* (FPR), tel que $NR + FPR = 1$

$$FPR = \frac{FP}{FP + TN} \tag{2.18}$$

- le taux de faux négatifs, *False Negative Rate* (FNR), tel que $R + FNR = 1$

$$FNR = \frac{FN}{TP + FN} \tag{2.19}$$

Dans le cas de deux classes, il est à noter que la précision négative est la précision de la classe négative et le rappel négatif est le rappel de cette même classe. On cherche à déterminer si un pixel étiqueté dans une certaine classe correspond à la vérité-terrain. La matrice de confusion synthétise la qualité d'une telle classification en calculant le rappel des deux classes (soient le rappel et le rappel négatif) et le taux de faux positifs des deux classes (soient le taux de faux positifs et le taux de faux négatifs) en les reportant à l'intérieur d'une matrice carrée 2×2. Les pixels correctement classés (rappel) se retrouvent sur la diagonale de la matrice. Dans le cas contraire, ceux-ci se retrouveront dans les autres éléments de la matrice, cf. tableau 2.2.

À titre d'exemple, pour la classification en zones planes/non-planes, qui sera abordée dans le chapitre 3, le cas d'une zone plane est considéré positivement (1) et le cas d'une zone non-plane négativement (0).

Matrice de confusion pour la classification multi-classes. Dans le cas d'une classification multiple, la matrice de confusion est étendue et adaptée au nombre de classes souhaitées. Le tableau à double entrée fournit alors pour chaque catégorie obtenue sa répartition suivant les classes de la vérité terrain. L'évaluation peut se ramener au cas binaire en considérant tour à tour chaque classe par rapport au reste. Nous pouvons cependant, noter la matrice de confusion générale comme le tableau 2.3.

Rés. \ VT	V_1	V_2	V_3
R_1	$\frac{\#(R_1 \cap V_1)}{\#V_1}$	$\frac{\#(R_1 \cap V_2)}{\#V_2}$	$\frac{\#(R_1 \cap V_3)}{\#V_3}$
R_2	$\frac{\#(R_2 \cap V_1)}{\#V_1}$	$\frac{\#(R_2 \cap V_2)}{\#V_2}$	$\frac{\#(R_2 \cap V_3)}{\#V_3}$
R_3	$\frac{\#(R_3 \cap V_1)}{\#V_1}$	$\frac{\#(R_3 \cap V_2)}{\#V_2}$	$\frac{\#(R_3 \cap V_3)}{\#V_3}$
R_0	$\frac{\#(R_0 \cap V_1)}{\#V_1}$	$\frac{\#(R_0 \cap V_2)}{\#V_2}$	$\frac{\#(R_0 \cap V_3)}{\#V_3}$

TABLE 2.3 – Matrice de confusion pour l'évaluation des performances d'une classification multi-classes : VT correspond aux éléments de la Vérité Terrain (V_1, V_2, V_3) et Rés. correspond aux résultats obtenus (R_1, R_2, R_3, R_0) avec R_0 la classe de rejet.

Courbe PR. Cette courbe représente la précision Pr en fonction du rappel R. Elle donne une bonne indication sur la qualité de la méthode, en particulier sur la pertinence du retour d'information par rapport à une requête. Une approche avec une bonne performance correspond à une courbe qui se situe dans le quart supérieur droit où le rappel et la précision sont maximisés.

Cependant, cette représentation est sensible à la distribution des deux classes, ce qui est un problème important. La courbe ROC tente de pallier ce problème.

Courbe ROC. L'analyse des performances par la courbe ROC s'est révélée très efficace lorsque le rapport entre les deux classes sont mal connus ou méconnus [Fawcett 06]. Il existe différentes représentations de celle-ci. Nous avons choisi la représentation graphique avec le taux de vrais positifs TPR (= rappel R) en fonction du taux de faux positifs FPR. Dans ce cas, une approche avec de bonne performance obtient une courbe qui se situe dans le quart supérieur gauche lorsque le rappel est maximisé et le taux FPR est minimal.

L'aire sous la courbe, notée AUC, *Area Under the Curve*, est un indicateur global de la qualité de la classification. Un classificateur donnant des résultats satisfaisants aura une AUC égal à 1. Pour la courbe ROC, un autre indicateur est le taux d'erreurs équivalentes EER, *Equal Error Rate*, il donne la valeur pour laquelle la probabilité d'obtenir des faux positifs est égale à la probabilité d'obtenir des faux négatifs. Cette valeur correspond à l'intersection de la courbe ROC et de la seconde diagonale $y = x$.

Autres critères. Le critère *Rand Index* [Rand 71], originalement introduit pour l'évaluation générale d'une classification a été décliné sous plusieurs variantes [Unnikrishnan 05, Arbeláez 09], pour, entre autres, prendre en compte plusieurs vérités terrain. Il n'est donc pas adapté à notre problème.

2.8.4 Critères d'évaluation de la segmentation

Les travaux de [Neubert 12] présentent une analyse des définitions des métriques utilisées lors de l'évaluation d'une sur-segmentation en super-pixel, comme l'erreur de sous-segmentation ou le recouvrement de contours que nous reprenons dans la suite.

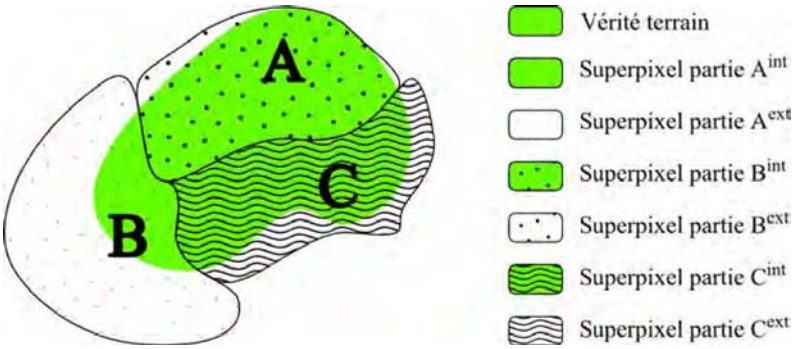


FIGURE 2.22 – Illustration de l'erreur de sous-segmentation lors d'une sur-segmentation en super-pixels : l'élément de vérité terrain G (vert) est recouvert par trois super-pixels (A,B,C) qui peuvent se situer au-delà des contours de la vérité terrain.

Erreur de sous-segmentation. L'erreur de sous-segmentation mesure l'adhérence des contours d'une sur-segmentation aux différentes zones de vérité terrain. Pour chaque région de la vérité terrain g_i, l'ensemble de super-pixels qui la recouvre $\{s_j | s_j \cap g_i \cup \emptyset\}$, permet d'évaluer quantitativement la juxtaposition et la cohérence du résultat obtenu avec la vérité terrain. En notant $|.|$ le nombre de pixels contenu dans une zone, M le nombre d'éléments de vérité terrain et B le pourcentage de recouvrement minimal entre un super-pixel et une zone de vérité terrain pour être pris en compte, l'erreur est définie par :

$$ErrSS_1 = \frac{1}{N}\Big[\sum_{i=1}^{M}\Big(\sum_{s_j\ s_j\ \ g_i>B} \#s_j\Big) - N\Big] \tag{2.20}$$

Les valeurs limites sont dans le cas d'un unique super-pixel, avec $n = 1$, $ErrSS = M - 1$, et dans le cas d'une infinité de super-pixels, où on obtient $\lim_{N\ \ \infty} ErrSS = 0$.

Cependant, si $B = 0$ avec cette première erreur, il arrive qu'une pénalité importante soit attribuée lorsqu'un super-pixel de taille importante ne se situe que sur une petite zone d'une région de la vérité terrain. Afin de pallier ce problème, deux variantes de l'erreur de sous-segmentation ont donc été proposées dans la littérature. Dans [Achanta 12], les auteurs ne prennent en compte que les super-pixels dont la surface d'intersection avec la zone de vérité terrain est supérieure à 5% de la taille de g_i, i.e. $B = 5$. Dans l'approche de [Neubert 12], les auteurs proposent de considérer chaque super-pixel $s_j = s_j^{int} \cup s_j^{ext}$ comme l'union d'une partie intérieure et d'une partie extérieure à g_i et de prendre en compte seulement la plus petite partie dans l'erreur.

$$ErrSS_2 = \frac{1}{N}\Big[\sum_{i=1}^{M}\Big(\sum_{s_j\ s_j\ \ g_i=} min(s_j^{int}, s_j^{ext})\Big)\Big] \tag{2.21}$$

Pour l'exemple de la figure 2.22, $ErrSS_1 = \dfrac{\#A^{ext} + \#B^{ext} + \#C^{ext}}{\#G}$ car le super-pixel B induit une forte pénalité dans le calcul de l'erreur, car seul une petite partie B^{int} recouvre l'élément de vérité terrain G, alors que $ErrSS_2 = \dfrac{\#A^{ext} + \#B^{int} + \#C^{ext}}{\#G}$.

Recouvrement de contours. Ce critère correspond à la fraction des contours de la vérité terrain se trouvant dans un voisinage de distance inférieure à la distance d du contour d'un super-pixel. L'équation est celle du rappel 2.8.3.

2.9 Conclusion

Les données récoltées pour la compréhension de scènes urbaines sont riches et variables. L'image seule contient de nombreuses caractéristiques photométriques, mais l'information extraite est généralement enrichie soit par de nouvelles vues, soit par des données d'autres capteurs (position, profondeur,...). En effet, la connaissance du positionnement et de l'orientation des sur-

faces permet une meilleure compréhension et une meilleure appréhension de la scène visuelle.

Tout d'abord, dans nos travaux nous travaillons avec des images multi-vues afin d'obtenir une segmentation sémantique d'une image de référence. Nous cherchons à intégrer lors de la construction d'une sur-segmentation en super-pixels l'information géométrique extraite. C'est pourquoi, dans le chapitre qui suit, nous cherchons à mettre en avant le lien entre l'information photométrique et l'*a priori* géométrique de planéité. Le critère photométrique utilisé est la cohérence photométrique (IQA) entre deux images. Une image de référence est comparée à une seconde image, correspondant à une vue virtuelle dite recalée, estimée via l'homographie induite par le plan de support. Le protocole proposé évalue les mesures pour la classification d'une région en zone plane ou non-plane.

Enfin, après avoir mis en évidence dans le chapitre 3, le type de mesures le plus approprié pour la distinction entre une zone plane et une zone non-plane, le chapitre 4 présente l'intégration de la mesure se similarité et de l'information de planéité dans la construction de super-pixels afin d'obtenir une segmentation sémantique en zones planes.

Chapitre 3

Classification de zones planes

Sommaire

3.1 Introduction

Comme nous avons pu l'illustrer dans les deux chapitres précédents, dans le contexte des scènes urbaines, l'hypothèse d'une scène plane par morceau est raisonnable et cohérent avec la géométrie de celle-ci. Ainsi, il nous paraît naturel de travailler sur la mis en œuvre d'un algorithme permettant de différencier les zones planes des zones non-planes afin de pouvoir distinguer, par exemple, la route des façades ou deux façades ayant des orientations différentes (mais parfois la même texture).

Comme nous l'avons vu dans l'état de l'art, un super-pixel est une entité de niveau intermédiaire plus cohérente photométriquement mais moins régulière qu'une boîte englobante ou un *patch*. Pour une résolution donnée des super-pixels, en particulier pour un faible nombre d'entités, il arrive que des régions ne soient pas cohérentes avec les contours des objets représentés dans l'image car le critère photométrique utilisé n'est pas suffisamment discriminant. Par exemple, sur la figure 3.1, les super-pixels SP_1 et SP_2, sont cohérents avec la géométrie de la scène car les pixels appartenant à une même région ont la même orientation, alors que SP_3 contient deux parties de deux façades avec des orientations perpendiculaires l'une par rapport à l'autre. Lors de l'étape de classification, un super-pixel non-plan tel que SP_3, sera difficile à classifier car les pixels qu'il contient appartiennent à deux entités sémantiques différentes.

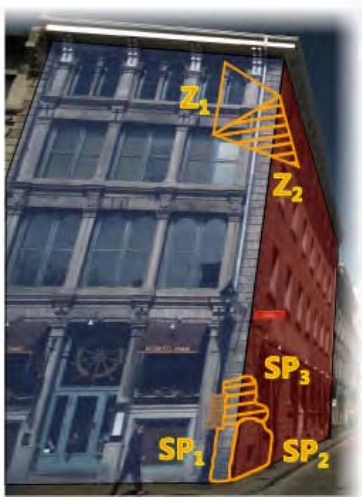


FIGURE 3.1 – Problème des super-pixels ou régions triangulaires contenant des surfaces avec différentes orientations : les régions SP_1, SP_2 et z_1 correspondent bien à des zones planes représentant la même entité alors que SP_3 et z_2 se retrouvent sur deux façades orthogonales et contiennent donc deux entités distinctes.

Dans un premier temps, en supposant que nous travaillons avec des images calibrées, nous proposons d'appliquer une triangulation de Delaunay sur un nuage de points 3D épars. Connaissant les correspondances de points 2D de ces points 3D, l'utilisation des zones triangulaires appariées, nous permet d'introduire un protocole simple pour qualifier la propriété d'une mesure de similarité ou dissimilarité détecter les erreurs d'appariement liées au fait que les zones appariées ne sont pas planes. En effet, cela est simplifié car nous pouvons calculer directement et de manière

unique le plan 3D associé à chaque zone triangulaire [Hartley 04], grâce à l'estimation facilitées de l'homographie induite par le plan de support.

Plus précisément, dans ce chapitre, nous cherchons à démontrer le lien entre la cohérence photométrique et l'estimation des orientations lors de la reconstruction 3D. En étudiant des triangles représentés sur au moins deux images, comme z_1 et z_2 de la figure 3.1, nous mettons en œuvre un processus permettant de distinguer les triangles représentant une surface 3D réellement plane comme z_1, de ceux correspondant à une zone non-plane comme z_2. L'approche de classification proposée, s'appuie sur un critère de similarité ou dissimilarité photométrique entre une image de référence et une image adjacente recalée via l'homographie induite par le plan de support. L'analyse de la similarité ou dissimilarité photométrique à partir d'un *a priori* géométrique permet de valider ou d'infirmer l'hypothèse de planéité. La figure 3.2 illustre les cas d'une zone plane et d'une zone non-plane.

(a) Zone plane (b) Zone non-plane

FIGURE 3.2 – Principe permettant de distinguer une zone plane d'une zone non-plane : Nous travaillons sur deux régions d'intérêt de l'image de référence I. Nous illustrons la manière de mettre en évidence la non-planéité en observant ce qui se passe si nous supposons que la zone n'est pas plane et doit donc être divisée en deux parties, chaque partie contenant une zone plane. Pour cela, le point q_λ parcourt le segment $[q_1 q_2]$. Dans le cas non-plan, l'intersection des deux plans $\pi_1 \cap \pi_2$ est notée $q_{\lambda^\star}$ et délimite le contour entre les deux surfaces que nous cherchons à distinguer.

Idéalement, une mesure de similarité ou dissimilarité adaptée à ce que nous souhaitons mettre en évidence : la non-planéité, doit avoir un comportement assez évident. En effet, la similarité (dissimilarité) doit être plus élevée (faible) dans le cas d'une surface plane que pour une surface non-plane car dans le cas d'une zone plane, le plan de support est un support 3D réel. Ainsi, l'homographie induite par ce plan est correcte et la similarité (dissimilarité) entre l'image de référence et l'image adjacente recalée sur l'image de référence est importante (faible). Dans le cas d'une zone non-plane, le plan de support ne correspond pas à un plan 3D réel. L'homographie induite par ce type de plan estime une mauvaise image recalée, peu similaire à l'image de référence et donc, nous attendons que cette mesure de similarité (dissimilarité) fournissent un score plus faible (plus important) que dans le cas plan.

Ainsi, ce chapitre présente les travaux publiés dans [Bauda 15b]. Étant donnée des images multi-vue calibrées, nous allons démontrer qu'à partir d'un *a priori* géométrique, nous pouvons trouver un critère photométrique permettant de distinguer les zones planes des zones non-planes. Notre contribution porte sur l'élaboration d'un protocole d'évaluation permettant de quantifier la qualité de la classification en zones planes et en zones non-planes et surtout de mettre en évidence les mesures de similarité et de dissimilarité les plus qualifiées pour le problème que nous nous posons : l'identification des zones planes et non-planes. Nous présentons dans un premier temps, les mesures de similarité et de dissimilarité que nous avons étudiées puis nous les analysons sur des exemples simples. Nous exposons ensuite, en détails, le protocole d'évaluation de ces mesures mis en place pour la classification des zones triangulaires en régions planes et non-planes. Cela permettra d'affiner le support de chaque plan, en combinant l'information géométrique et photométrique disponible. L'estimation des plans permet de calculer la transformation homographique correspondante afin d'utiliser un critère photométrique inter-image dense. Enfin, les résultats obtenus sont analysés et interprétés.

3.2 Mesures de cohérence photométrique inter-images

Les mesures de cohérence photométrique (mesures de similarité ou dissimilarité) également appelées IQA. Dans cette section, nous rappelons que nous souhaitons comparer l'ensemble des niveaux de gris d'une région plane donnée avec l'ensemble des niveaux de gris d'une région plane candidate pour être sa correspondante, recalée suivant l'homographie estimée à partir des correspondances de points. Nous souhaitons que cette mesure soit faible (similarité)/forte (dissimilarité) lorsque la région considérée au départ n'est pas plane. Ainsi, ces mesures doivent permettent de quantifier la similarité ou la dissimilarité de deux régions $\mathbf{z}$ et $\tilde{\mathbf{z}}$. Dans chacune de ces deux régions il y a N pixels. Voici les notations utilisées :

- Nous notons v_i, respectivement $\tilde{v}_i$, la valeur d'intensité ou de luminosité dans l'espace de couleur CIELab d'un pixel q_i, respectivement $\tilde{q}_i$, pris dans la région $\mathbf{z}$, respectivement $\tilde{\mathbf{z}}$.
- De plus, nous notons v, respectivement $\tilde{v}$, le vecteur linéarisé contenant l'ensemble des scalaires v_i, respectivement $\tilde{v}_i$, pris dans un voisinage de q_i, respectivement $\tilde{q}_i$. La définition de ce voisinage dépend de la mesure définie.

Nous distinguons les mesures s'appuyant sur la distance Euclidienne de celles utilisant le calcul du cosinus de l'angle, c'est-à-dire le produit scalaire (formés par les deux vecteurs de données v et $\tilde{v}$). Les représentantes les plus significatives et les plus connues de ces deux types de mesures sont respectivement la mesure *Mean Square Error* (MSE) et la mesure *Structure SImilarity Measure* (SSIM). Elles ont été largement analysées et comparées, notamment dans les articles [Wang 09, Palubinskas 14].

3.2.1 Mesures s'appuyant sur la distance euclidienne

Les trois mesures que nous détaillons utilisent donc la distance Euclidienne entre les valeurs d'intensité de chaque pixel des régions comparées, cf. figure 3.3. Ces mesures quantifient la dissimilarité entre les deux vecteurs v et $\tilde{v}$. L'intervalle de valeurs de ces mesures est $[O, M^2]$, où M est la valeur maximale de luminosité atteinte et dépend du type d'encodage de l'image (caractère non signé, réel).

MSE est la mesure la plus connue de cette classe. Une simple différence entre les valeurs d'intensité des pixels de chaque image est considérée. Elle est définie par :

$$\mathrm{MSE}(\mathbf{z}, \tilde{\mathbf{z}}) = \frac{1}{N} \sum_{i=1}^{N} (v_i - \tilde{v}_i)^2 \tag{3.1}$$

***Mean Square Error* sur un voisinage r ($\mathbf{MSE}_r$)** est une extension de MSE sur une fenêtre de taille $(2r+1) \times (2r+1)$ centrée sur un pixel donné q_i.

$$\mathrm{MSE}_r(\mathbf{z}, \tilde{\mathbf{z}}) = \frac{1}{N} \sum_{i=1}^{N} [\frac{1}{(2r)^2} \sum_{j/\ q_i - q_j \atop r} (v_j - \tilde{v}_j)^2] \tag{3.2}$$

Cette mesure permet de prendre en compte le voisinage du pixel courant en effectuant la moyenne des variations sur la fenêtre considérée.

R-cohérence (r-coherence ($\mathbf{RC}_r$)) est utilisée dans [Bartoli 07]. Cette mesure quantifie la différence pixélique entre le pixel $q_i \in \mathbf{z}$ et le pixel le plus similaire dans le r-voisinage de $q'_i \in \mathbf{z}'$. Elle minimise la différence sur un voisinage pour compenser la distorsion due à l'estimation de l'image recalée.

$$\mathrm{RC}_r(\mathbf{z}, \tilde{\mathbf{z}}) = \frac{1}{N} \sum_{i=1}^{N} \Big(\min_{j/(q_i - q_j)^2 < r^2} |v_i - \tilde{v}_j| \Big)^2 \tag{3.3}$$

Analyse des trois mesures présentées. Les trois mesures présentées précédemment s'appuient sur la distance Euclidienne en effectuant une simple différence entre les valeurs d'intensité de l'image de référence $\mathbf{z}$ et de l'image adjacente $\tilde{\mathbf{z}}$. La variante MSE_r moyenne la variabilité sur un voisinage et RC_r optimise cette variabilité sur le voisinage. Cependant, la différence terme à terme n'est pas toujours représentative des variabilités inter-image, c'est pourquoi les mesures présentées dans la suite s'appuient sur des données statistiques calculées sur une fenêtre centrée sur le pixel d'intérêt.

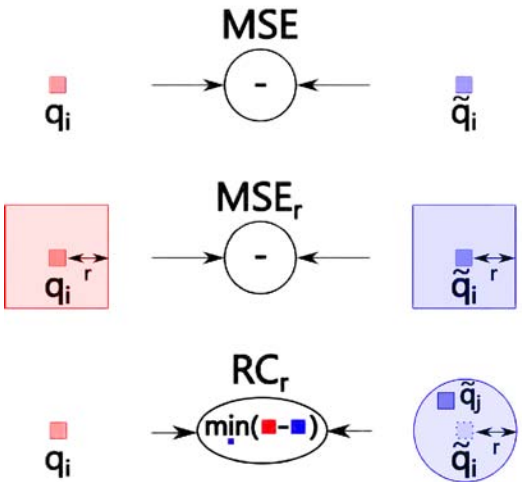


FIGURE 3.3 – Mesures de dissimilarité s'appuyant sur la distance Euclidienne : elles sont appliquées à tous les pixels q_i de l'image de référence et $\tilde{q}_i$ de l'image recalée. Il s'agit d'une représentation graphique des trois mesures de dissimilarité pour un seul pixel ($N = 1$) : MSE, MSE_r et RC_r. MSE correspond à une simple différence terme à terme des valeurs d'intensité des pixels. MSE_r est une extension de MSE en prenant en compte le voisinage. RC_r correspond à l'optimisation de la variation étant donnée un voisinage r dans l'image adjacente, atteinte pour le pixel $\tilde{q}_j$.

3.2.2 Mesures s'appuyant sur le produit scalaire

Il s'agit de mesures de similarité qui s'appuient sur le calcul du produit scalaire entre les deux vecteurs v et $\tilde{v}$, contenant $(2n+1)^2$ pixels. Ces vecteurs peuvent être vus ici, comme des variables aléatoires. Une représentation graphique des mesures est donnée figure 3.4. Nous définissons également les termes suivants :

- $\mu_{\mathbf{z}}$ (respectivement $\mu_{\tilde{\mathbf{z}}}$) la moyenne pondérée par w_i des v_i (respectivement $\tilde{v}_i$) sur la région $\mathbf{z}$ (respectivement $\tilde{\mathbf{z}}$)

$$\mu_{\mathbf{z}} = \sum_{i=1}^{(2n+1)^2} w_i v_i; \tag{3.4}$$

- $\sigma_{\mathbf{z}}$ (respectivement $\sigma_{\tilde{\mathbf{z}}}$) l'écart-type pondéré par w_i de $\mathbf{z}$ (respectivement $\tilde{\mathbf{z}}$)

$$\sigma_{\mathbf{z}} = \Big(\sum_{i=1}^{(2n+1)^2} w_i (v_i - \mu_{\mathbf{z}}) \Big)^{\frac{1}{2}}; \tag{3.5}$$

- $\sigma_{\mathbf{z}\tilde{\mathbf{z}}}$ la covariance de $\mathbf{z}$ et de $\tilde{\mathbf{z}}$ pondérée par w_i

$$\sigma_{\mathbf{z}\mathbf{z}'} = \sum_{i=1}^{(2n+1)^2} w_i (v_i - \mu_{\mathbf{z}})(v'_i - \mu_{\mathbf{z}'}). \tag{3.6}$$

SSIM est proposée dans [Wang 04]. Elle est définie par le produit de trois termes :

$$\text{SSIM}(\mathbf{z}, \tilde{\mathbf{z}}) = l(\mathbf{z}, \tilde{\mathbf{z}}) \quad c(\mathbf{z}, \tilde{\mathbf{z}}) \quad s(\mathbf{z}, \tilde{\mathbf{z}}) \tag{3.7}$$

où :

- $l(\mathbf{z}, \tilde{\mathbf{z}})$, la luminosité, mesure la similarité des valeurs moyennes.
- $c(\mathbf{z}, \tilde{\mathbf{z}})$, le contraste, quantifie la similarité des écart-types de chaque zone.
- $s(\mathbf{z}, \tilde{\mathbf{z}})$, la structure, évalue la similarité inter-image.

Chacun de ces termes est défini par :

$$l(\mathbf{z}, \tilde{\mathbf{z}}) = \frac{2\mu_{\mathbf{z}}\mu_{\tilde{\mathbf{z}}} +}{\mu_{\mathbf{z}}^2 + \mu_{\tilde{\mathbf{z}}}^2 +} \tag{3.8}$$

$$c(\mathbf{z}, \tilde{\mathbf{z}}) = \frac{2\sigma_{\mathbf{z}}\sigma_{\tilde{\mathbf{z}}} +}{\sigma_{\mathbf{z}}^2 + \sigma_{\tilde{\mathbf{z}}}^2 +} \tag{3.9}$$

$$s(\mathbf{z}, \tilde{\mathbf{z}}) = \frac{\sigma_{\mathbf{z}\tilde{\mathbf{z}}} +}{\sigma_{\mathbf{z}}\sigma_{\tilde{\mathbf{z}}} +} \tag{3.10}$$

D'une part, les poids Gaussiens w_i sont introduits afin de donner plus d'importance au pixel central. D'autre part, les constantes , et permettent d'éviter la division par zéro, cas particulier où les zones sont homogènes et proches de la valeur nulle (zone noire), $\sigma_{\mathbf{z}} = 0$ et $\mu_{\mathbf{z}} = 0$. La mesure SSIM est symétrique, bornée entre $[-1, 1]$ et atteint son maximum lorsque les deux zones sont similaires i.e. $\mathbf{z} = \tilde{\mathbf{z}}$.

Le coefficient de corrélation est calculé en divisant la covariance des deux variables par le produit de leur écart-type. En particulier et après réécriture, lorsque $= 0$, le terme de structure $s(\mathbf{z}, \tilde{\mathbf{z}})$ correspond à la mesure *Zero mean Normalised Cross-Correlation* (ZNCC) [Aschwanden 92].

UQI est proposée par [Wang 02]. Elle correspond à un cas particulier de SSIM avec $= = = 0$ et avec une pondération uniforme à la place de la pondération gaussienne. Cela signifie que tous les pixels de la fenêtre glissante ont la même importance. Si $\mu_z = \mu_{\tilde{z}}$ ou $\sigma_z = \sigma_{\tilde{z}}$ alors $UQI = 1$. Après simplification d'écriture, UQI s'écrit :

$$\text{UQI}(\mathbf{z}, \tilde{\mathbf{z}}) = \frac{4\sigma_{\mathbf{z}\tilde{\mathbf{z}}}\mu_{\mathbf{z}}\mu_{\tilde{\mathbf{z}}}}{(\sigma_{\mathbf{z}}^2 + \sigma_{\tilde{\mathbf{z}}}^2)[\mu_{\mathbf{z}}^2 + \mu_{\tilde{\mathbf{z}}}^2]} \tag{3.11}$$

***Universal Quality Image* sur un voisinage r (RUQI)** combine les avantages de l'analyse statistique de SSIM et de l'optimisation de la similarité sur le r-voisinage de la r-cohérence. Cette mesure combine deux types de voisinages : celui pris en compte dans la recherche de l'optimisation de la similarité dans $\tilde{z}$, noté r et celui de la taille de la fenêtre utilisée pour les statistiques de premier ordre sur les deux images, noté n. Tout comme le critère de RC_r, cette mesure renvoie un seul terme qui est la valeur de similarité inter-image optimale obtenue dans le voisinage défini par r. Or, la valeur de r semble également renseigner sur la déformation entre les deux images.

$$\mathrm{RUQI}(\mathbf{z},\tilde{\mathbf{z}}) = \frac{1}{N}\sum_{i=1}^{N}\Big(\max_{j/(q_i-q_j)^2<r^2}(\mathrm{UQI}(\xi_i,\tilde{\xi_j}))\Big) \tag{3.12}$$

où ξ_i, respectivement $\tilde{\xi_j}$, est défini par une fenêtre glissante sur $\mathbf{z}$, respectivement $\tilde{\mathbf{z}}$, centrée sur q_i, respectivement q_j.

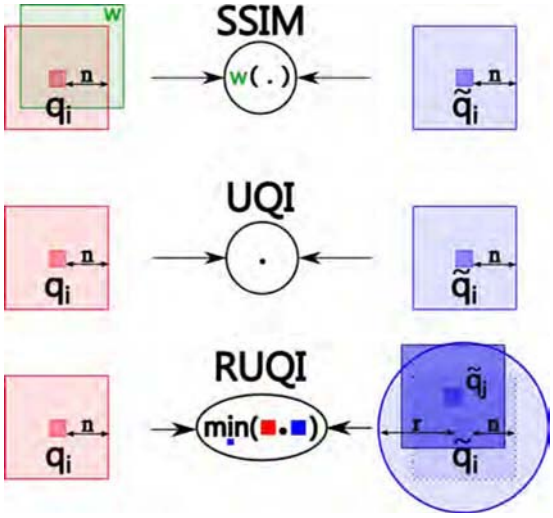


FIGURE 3.4 – Mesures de similarité s'appuyant sur le produit scalaire : elles sont appliquées à tous les pixels q_i de l'image de référence et $\tilde{q}_i$ de l'image recalée. Il s'agit d'une représentation graphique des trois mesures de dissimilarité pour un seul pixel ($N = 1$) : SSIM, UQI et RUQI. Ces trois mesures prennent en compte les statistiques des valeurs d'intensité sur une fenêtre de taille $(2n + 1)$ centrée sur le pixel d'intérêt q_i. UQI est un cas particulier de SSIM où la pondération est uniforme et les constantes sont nulles. RUQI combine les avantages des données statistiques et de l'optimisation sur le voisinage r.

Analyse des trois mesures présentées. Elles permettent de prendre en compte des éléments statistiques sur la distribution des données.

Afin de mieux comprendre les caractéristiques des deux types de mesures, nous proposons une analyse préliminaire de leur sensibilité respective à différentes perturbations.

3.3 Analyses préliminaires de toutes les mesures présentées

Tout d'abord, nous pouvons remarquer que chacune des mesures définie précédemment est symétrique par rapport à z et $\tilde{z}$. Ainsi, nous avons $IQA(z,\tilde{z}) = IQA(\tilde{z},z)$ et contrairement aux méthodes de flux optique qui fournissent l'orientation du déplacement, nous ne disposons que d'une similarité/dissimilarité inter-image non-signée.

Pour les mesures utilisant la distance euclidienne, r est la taille de la marge de la zone d'intérêt. Lorsque $r = 0$, il faut noter que $\mathrm{RC}_0 = \mathrm{MSE}_0 = \mathrm{MSE}$. La mesure MSE compare l'information pixel-à-pixel pendant que MSE_r prend également en compte le voisinage. RC_r optimise la diffé-

rence sur un voisinage r. Les mesures s'appuyant sur le produit scalaire prennent en compte la distribution statistique des valeurs dans chaque image z et $\tilde{z}$ sur une fenêtre de corrélation de taille n. Pour SSIM, des poids gaussiens sont utilisés pour donner plus d'importance au pixel central de la fenêtre. La mesure UQI est un cas particulier de SSIM, avec pondération uniforme sur les pixels de la fenêtre et non réajustée pour les valeurs de moyennes et d'écart-type proche de zéro. La mesure RUQI, que nous avons proposée, fusionne les avantages de UQI et de RC_r, en combinant l'information de la distribution statistique des données à l'optimisation de la similarité sur un r-voisinage.

À présent, nous nous intéressons à trois difficultés ou configurations ayant un impact significatif sur la qualité de la quantification de la similarité/dissimilarité photométrique inter-image :

- le bruit qui va induire des erreurs de recalage et donc une similarité plus difficile à établir ;
- le déplacement des objets dans les images qui va généré, entre autres, des occultations et ainsi compromettre la qualité du recalage ;
- la résolution de l'image qui a un impact sur la taille de la région à étudier, de manière à obtenir une information suffisamment riche pour la décrire et la distinguer des autres régions.

Nous allons donc étudier le comportement des mesures face à ces trois configurations.

3.3.1 Comportement des mesures face à un bruit et/ou un déplacement

Les mesures présentées précédemment sont analysées sur quatre cas simples, présentés dans la figure 3.5. Les images traitées ici sont des cas particuliers dans le sens où nous avons tenter de simuler de la manière la plus simple les deux situations que nous souhaitons étudiées. Nous avons donc généré deux images d'origine avec de nombreuses régions homogènes (non-texturées) en niveau de gris comprenant, pour la première image, image gauche en (c), des zones blanches (1 ou 255) et noires (0) et, pour la seconde, image gauche en (a), nous avons ajouté deux zones de niveaux de gris intermédiaires. Pour la génération des images adjacentes, nous avons ajouté un bruit blanc (bruit gaussien centré et réduit) noté $\mathcal{N}(0,1)$ et/ou une translation T. Ainsi, nous avons généré les images adjacentes suivantes, images de droite dans la figure 3.5 :

- (a) Image homologue : $z' = z$,
- (b) Image homologue bruitée : $z' = z + \mathcal{N}(0,1)$,
- (c) Image avec un déplacement global (une translation suivant l'axe des ordonnées) : $z' = T(z)$,
- (d) Image avec un déplacement global (une translation suivant l'axe des ordonnées) et bruitée : $z' = T(z) + \mathcal{N}(0,1)$.

Dans la figure 3.6, nous présentons les résultats visuels du calcul de la similarité/dissimilarité pixel à pixel, dans chacun des 4 cas et pour les 6 mesures présentées. Nous remarquons :

- **Pour le cas (a) :** $z' = z$ – Comme nous pouvions nous y attendre, toutes les mesures sauf RUQI, ont le comportement attendu une faible/forte réponse de dissimilarité/simi-

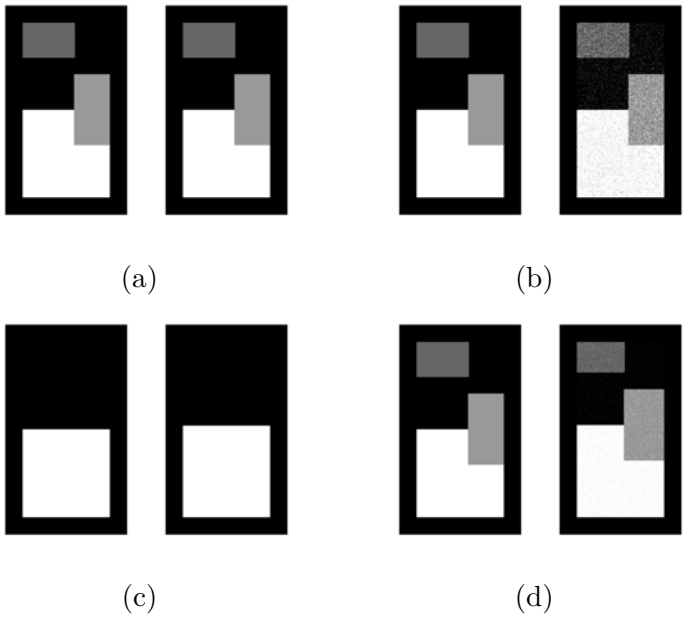


FIGURE 3.5 – Couples d'images synthétiques utilisés pour étudier le comportement des mesures de similarité et de dissimilarité : (a) l'image de référence et l'image adjacente sont similaires, (b) l'image adjacente est bruitée avec un bruit blanc. (c) l'image adjacente est translatée verticalement, (d) l'image adjacente est bruitée avec un bruit blanc et translatée verticalement.

larité, égale à la valeur minimale/maximale. Dans une zone parfaitement homogène, la variance est nulle, et, dans ce cas, le dénominateur de RUQI est nul. Ce cas est filtré par l'implémentation que nous avons utilisée, c'est-à-dire que l'on estime que cette zone n'est pas fiable et on attribue la similarité la plus basse.

- **Pour le cas (b) :** $z' = z + \mathcal{N}(0, 0.1)$ – Toutes les mesures ont une faible/forte réponse de dissimilarité/similarité, sauf UQI et RUQI où la réponse est élevée seulement sur les zones où les variations d'intensité sont non nulles, i.e. sur les contours. Cela est dû au fait qu'en dehors de ces zones de variations générées par la discontinuité au niveau du contour, les niveaux de gris sont homogènes et donc, comme pour le cas (a), la similarité sera minimale pour ces pixels, et, par conséquence, plus élevée sur la zone d'influence relative à la taille de r autour du contour.
- **Pour le cas (c) :** $z' = T(z)$ – Toutes les mesures donnent une dissimilarité/similarité élevée/faible sur la zone du déplacement, bien que pour SSIM, la variation du signal soit plus faible que pour les autres.
- **Pour le cas (d) :** $z' = T(z) + \mathcal{N}(0, 0.01)$ – Toutes les mesures mettent en avant les zones de contours, s auf SSIM pour qui la variation de la réponse est très faible. Nous remarquons également que la mesure UQI est moins performante lorsque le déplacement et le bruit sont appliqués conjointement.

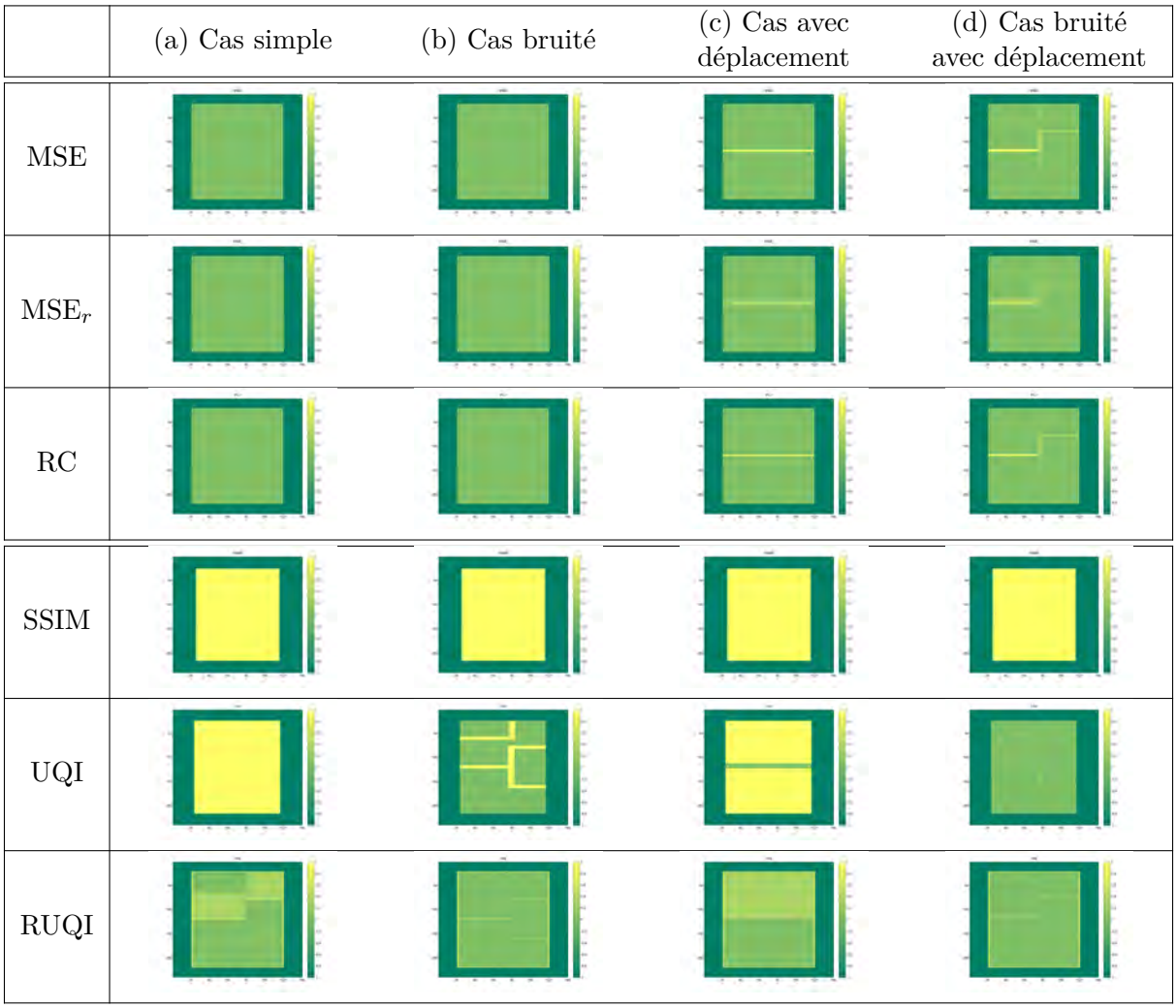


FIGURE 3.6 – Comportement des mesures de similarité et de dissimilarité face au bruit et à la présence de déplacement : les couples utilisés pour les cas (a) à (d) sont présentés dans la figure 3.5. Les trois premières mesures sont des mesures de dissimilarité, par opposition aux trois suivants qui sont des mesures de similarité (c'est la raison pour laquelle on devrait observer des comportements inversés). Plus le pixel est vert, plus la dissimilarité, respectivement la similarité, entre l'intensité du pixel dans l'image d'origine et l'intensité du pixel dans l'image adjacente est faible, respectivement forte, et inversement pour le jaune.

Alors que les mesures s'appuyant sur la différence euclidienne ont un comportement proche, les mesures s'appuyant sur le produit scalaire ont des comportements variables et différents. En effet, la mesures SSIM indique une forte similarité quelle que soit la perturbation appliquée à l'image adjacente $\tilde{z}$. La mesure UQI présente une forte dissimilarité sur les contours. La mesure RUQI a l'avantage de ne pas fournir de « fausse » similarité dans les zones homogènes et préfère ne pas se prononcer. C'est un comportement particulier qui est avantageux dans certains cas, en particulier dans le cas de scènes urbaines où l'on peut être en présence de zones homogènes étendues et qui ne comportent pas d'information directement exploitable, comme le ciel. En revanche, ce comportement peut également être un problème car cela a tendance à renforcer

la similarité dans un cas bruité ou avec un déplacement par rapport à un cas simple mais peu texturé.

3.3.2 Comportement des mesures suivant di érentes résolutions

Les mesures RC_r et RUQI font intervenir un paramètre r permettant de rechercher le meilleur pixel correspondant dans un certain voisinage relatif à r. Plus la valeur de r est importante plus on augmente les chances de trouver un pixel correspondant proche de celui recherché. Cela signifie aussi que la mesure est moins discriminante puisqu'elle va chercher en quelque sorte à « forcer » ou renforcer la similarité même dans un cas où deux pixels ne se correspondent pas. Pour ces mesures, un second paramètre est utilisé : n qui indique la taille du voisinage à prendre en compte pour faire le calcul de la similarité/dissimilarité. Dans l'analyse du comportement des mesures, il est équivalent d'augmenter la valeur du paramètre n pour une résolution donnée ou de réduire la résolution de l'image pour une valeur de n donnée.

Pour simplifier l'étude, nous nous intéressons ici, en particulier aux mesures de référence pour chacune des deux classes : MSE et SSIM. Nous avons étudié trois résolutions différentes : 100%, 40% et 20% où l'interpolation de Lanczos [Getreuer 11] est utilisée pour la génération des images à basse résolution. Le test que nous avons réalisé consiste à étudier ce qui se passe lorsque l'on étudie un triangle non-plan. Dans ce cas, on cherche à estimer λ, paramètre relatif à l'orientation de la droite qui permet de séparer ce triangle non-plan en deux sous-triangles plan, cf. la figure 3.2 en (b), où λ correspond au seul point où le recalage sera juste sur chaque sous-triangles. En ce point, la valeur de similarité ou dissimilarité sera maximale, respectivement minimale, en comparaison avec toutes les autres valeurs de λ. Ce que nous observons dans la figure 3.7, c'est que plus la résolution de l'image est faible plus la similarité est globalement forte, et donc, moins la valeur de similarité/dissimilarité obtenue en λ se distinguera des autres. En effet, pour une taille de fenêtre de corrélation donnée, lorsque la résolution diminue, la précision diminue nous pouvons illustré ce comportement en notant que, dans cette figure, les courbes (rouges) qui correspondent au cas où la résolution est de 20% sont moins discriminantes que celles correspondant à une résolution de 100% (bleu), c'est-à-dire que le pic obtenu en λ est moins saillant. De plus, pour MSE, nous remarquons également que les valeurs ont diminué de moitié entre la résolution à 20% et à 100%, et, de même, pour SSIM elles sont passées de 0.3 donc peu similaire à plus similaire pour une valeur de 0.5.

3.3.3 Synthèse sur les mesures de cohérence photométrique

En conclusion de ces analyses préliminaires, nous avons remarqué que la mesure UQI montre un comportement particulier au niveau des contours puisque c'est uniquement aux niveaux des contours qu'elle estime que la mesure est fiable.

L'analyse de ces mesures sur des images simples ayant subi des transformation connues (bruit, déplacement ou changement de résolution) montre les spécificités de leur réponses. elles pré-

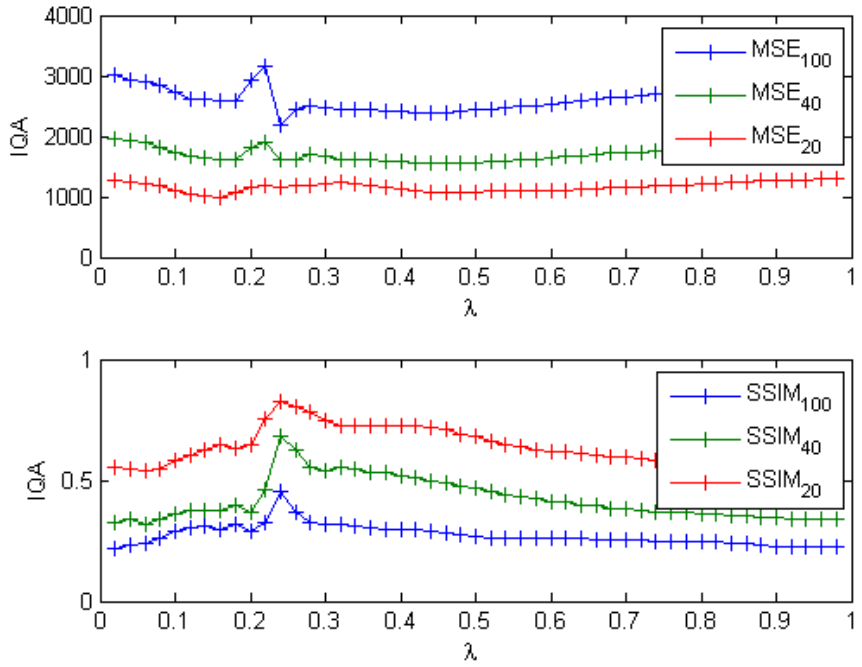


FIGURE 3.7 – Comportement des mesures face aux changements d'échelle/de résolution. : Nous avons étudié le comportement de MSE et SSIM en fonction de λ, cf. figure 3.2, paramètre qui fait varier la direction de la droite qui sépare un triangle non-plan en deux « sous-triangles » plans. Nous avons fait varier les résolutions des images en prenant trois résolutions : 100% (courbes bleues) , 40% (courbes vertes) et 20% (courbes rouges) de la pleine résolution.

sentent des réponses variables mais cohérentes à leur construction. Dans le cas particulier du changement de résolution, comme attendu, plus la résolution est faible, moins les mesures sont discriminantes.

Cette analyse préliminaire des mesures, nous permet de mieux comprendre leur comportement afin de pouvoir les introduire dans le protocole d'évaluation des mesures de cohérence photométrique pour la classification de zones planes et non-planes que nous présentons dans la partie suivante.

3.4 Proposition d'un protocole de classification des zones planes

L'objectif de l'utilisation de ce protocole de classification des zones planes est de mettre en évidence les mesures de similarité/dissimilarité qui sont les plus adaptées pour distinguer les zones planes des zones non-planes. Ainsi, étant donnée une zone triangulaire d'une scène représentée dans deux images, comme celles de la figure 3.8, nous souhaitons déterminer si cette zone correspond à une zone 3D plane ou non-plane. En estimant à partir des trois sommets d'un triangle $\{q_i, i = 1, 2, 3\}$, le plan de support π, nous connaissons *a priori* la géométrie de la scène. Nous cherchons donc à démontrer que l'évaluation d'un critère de cohérence photométrique

(similarité/dissimilarité), comme ceux présentés précédemment dans la section 3.2, entre une image de référence I et une image homologue $\tilde{I}$, correspondant à une image adjacente I' déformée par une transformation homographique connue H, permet de valider ou d'infirmer cet *a priori* géométrique.

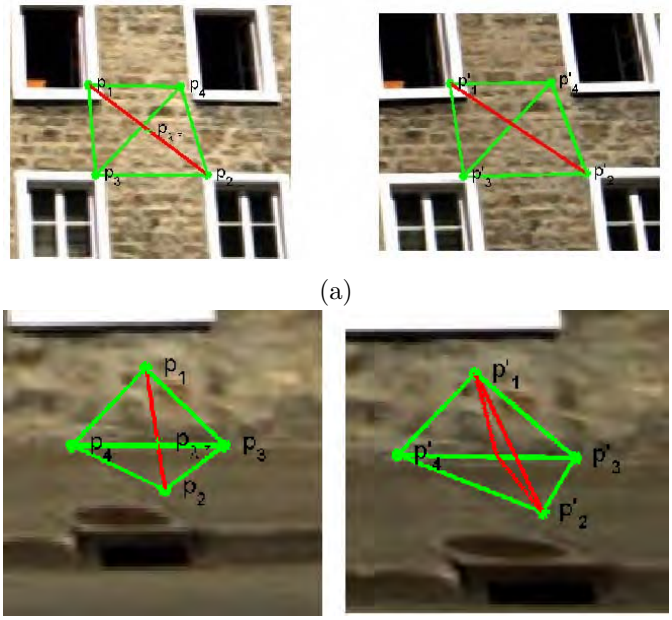


FIGURE 3.8 – Exemples d'image de référence I et d'image adjacente I' : (a) Cas plan avec une forte similarité inter-image. (b) Cas non-plan avec une faible similarité inter-image. La droite brisée représentée dans l'image adjacente correspond à la projection du segment $[p_1p_2]$ par les homographies induites par les deux plans de support $(p_1p_3p_4)$ et $(p_2p_3p_4)$.

3.4.1 Description générale du protocole

L'algorithme 1 présente une vue d'ensemble du protocole mis en place. Ce protocole permet de quantifier la similarité entre deux images telles que celles illustrées dans la figure 3.9. Nous comparons les différentes mesures de cohérence photométrique en analysant la réponse de la classification en zone planes et non-planes par rapport à une vérité terrain donnée. Nous connaissons les paramètres de calibrage et nous travaillons à partir de quatre points mis en correspondance sur les deux images. Pour une simplification des estimations des homographies, les points utilisés ont été, soit sélectionnés manuellement, soit choisis parmi un ensemble de points d'intérêt mis en correspondance de manière automatique, à l'aide d'outil SfM, tel que VisualSfM [Wu 11a]. Dans le cas de zones non-planes, deux des points se situent sur l'arrête à détecter et les deux autres sur chacune des surfaces adjacentes et visibles de l'arrête.

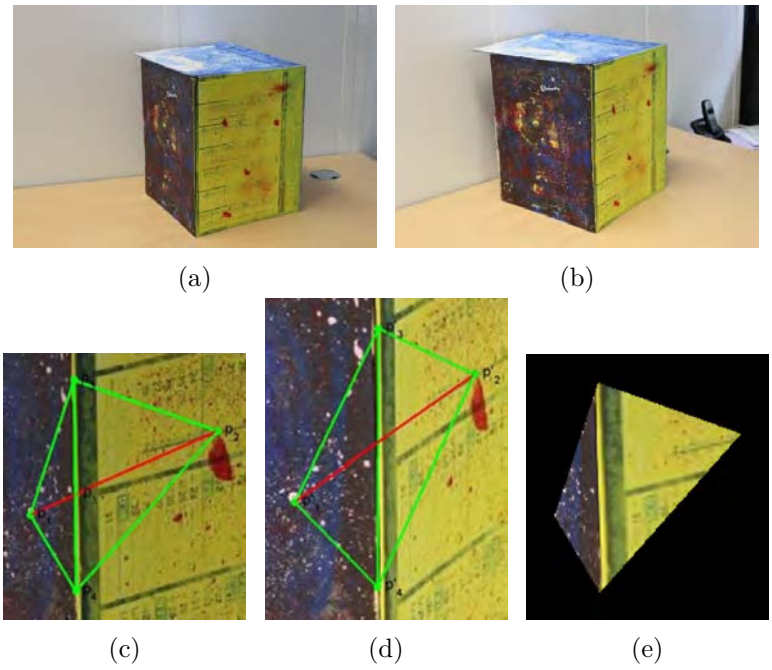


FIGURE 3.9 – Recalage par morceaux d'un triangle non-plan via des homographies : (a) Image de référence I, (b) image adjacente I', (c) zone de l'image de référence z, (d) zone correspondante z' de l'image adjacente, (e) zone adjacente recalée $\tilde{z}$ sur l'image de référence par les homographies induites par les plans de support, c'est-à-dire induites par les deux triangles séparés par le segment en rouge.

3.4.2 Estimation d'homographies

Nous calculons les homographies à partir des paramètres de calibrage de la caméra, d'une ligne et d'un point mis en correspondance sur les deux images [Hartley 04, p.331], car nous avons remarqué, de manière empirique, que cette technique est la plus fiable. En effet, la mise en correspondance de la droite est une contrainte forte lors de l'estimation de l'homographie. De plus, Nous estimons les deux homographies H_1 et H_2 induites par le plan de support π_1 et respectivement π_2 où $\{q_3, q_4, q_1\} \in \pi_1$ et $\{q_3, q_4, q_2\} \in \pi_2$. Nous définissons $\lambda^\star$ tel que $q_{\lambda^\star} = (q_1q_2) \cap (q_3q_4)$ et nous notons q_λ un point défini par $q_\lambda = \lambda q_1 + (1-\lambda)q_2$ dans l'image I avec $\lambda \in [0,1]$. Le point q'_λ est obtenu par projection de son correspondant q_λ dans l'image adjacente par l'homographie H_1 si λ et plus petit que $\lambda^\star$ ou par H_2 sinon. La troisième homographie H utilisée est estimée avec cette nouvelle mise en correspondance du point $p_\lambda \leftrightarrow q'_\lambda$. En fonction de la valeur de λ, nous pouvons construire deux images partielles et recalées sur l'image de référence $\tilde{z}_1$ et $\tilde{z}_2$. La zone d'intérêt de l'image adjacente recalée est alors $\tilde{z} = \tilde{z}_1 \cup \tilde{z}_2$.

3.4.3 Calcul de la similarité inter-images

Les auteurs de [Seitz 06] proposent de classer les mesures de photo-consistance suivant le domaine d'intégration : dans l'espace de la scène 3D ou dans l'espace image. Bien que cela semble être lié au modèle de reconstruction utilisé, ces deux approches sont cependant très similaires.

Algorithme 1 : Protocole d'évaluation des mesures de cohérence photométrique pour la classification en zones planes (P)/non-planes (NP) : Toutes les étapes sont développées dans la section 3.4. Le terme $\lambda^\star$ correspond à la valeur du paramètre λ tel que q_λ permet de séparer un triangle non-plan en deux triangles plans.

Données : 4 points d'intérêt mis en correspondance $q_1 \leftrightarrow q'_1$, $q_2 \leftrightarrow q'_2$, $q_3 \leftrightarrow q'_3$, $q_4 \leftrightarrow q'_4$ entre les deux images I et I', les paramètres de calibrage
Résultat : Classification en zones planes (P)/non-planes (NP)

// Estimation de la valeur $\lambda^\star$
$q_{\lambda^\star} \quad (q_1q_2) \cap (q_3q_4)$;
// Estimation des homographies (3.4.2)
$H_1 \quad calculHomographie(q_3, q_4, q_1)$;
$H_2 \quad calculHomographie(q_3, q_4, q_2)$;
// Calcul de la valeur IQA pour chaque λ
pour $\lambda = 0 : d\lambda : 1$ **faire**
  // Calcul du point $\mathbf{q}_\lambda \in [\mathbf{q}_1\mathbf{q}_2]$
  $q_\lambda \quad \lambda q_1 + (1-\lambda)q_2$;
  // Estimation de l'image recalée
  si $\lambda < \lambda^\star$ **alors**
    $q_{\lambda'} \quad H_1(q_\lambda)$;
    $H \quad calculHomographie(q_2, q_3, q_\lambda)$;
    $\tilde{\mathbf{z}}_1 \quad H_1(\mathbf{z}')$;
    $\tilde{\mathbf{z}}_2 \quad H(\mathbf{z}')$;
  sinon
    $q_{\lambda'} \quad H_2(q_\lambda)$;
    $H \quad calculHomographie(q_1, q_3, q_\lambda)$;
    $\tilde{\mathbf{z}}_1 \quad H(\mathbf{z}')$;
    $\tilde{\mathbf{z}}_2 \quad H_2(\mathbf{z}')$;
  $\tilde{\mathbf{z}} \quad \tilde{\mathbf{z}}_1 \cup \tilde{\mathbf{z}}_2$;
  // Calcul de la similarité inter-images avec IQA (3.4.3)
  $IQA(\lambda, \mathbf{z}, \tilde{\mathbf{z}}) \quad calculIQA(\mathbf{z}, \tilde{\mathbf{z}})$;
  // Classification en zone P/NP (3.4.4)
  si max $IQA_s(\boldsymbol{z}, \tilde{\boldsymbol{z}}) > \epsilon$ **alors**
    $\mathcal{C}(\mathbf{z}, \tilde{\mathbf{z}}) \quad$ P;
  sinon
    $\mathcal{C}(\mathbf{z}, \tilde{\mathbf{z}}) \quad$ NP;

Dans le premier cas, dans l'espace 3D de la scène, l'erreur est intégrée sur la surface. Elle est plus significative sur des petits éléments de surface. Dans le second cas, l'erreur est intégrée sur l'ensemble des images disponibles. Elle est plus pertinente sur les zones où l'information est redondante,ou bien sur les zones de grandes tailles. Nous choisissons ici, de considérer la mesure de similarité dans l'espace image, en analysant $IQA(z, \tilde{z})$

Nous cherchons à quantifier la variabilité de la similarité en fonction de la position du point q_λ, situé sur l'arrête potentielle. Pour cela, q_λ parcourt la droite $[q_1q_2]$. Dans l'exemple de région non-plane, cf. figure 3.10, nous pouvons remarquer que la similarité maximale est obtenue lorsque q_λ se trouve sur l'arrête d'intersection des deux faces au point $q^\star_\lambda$.

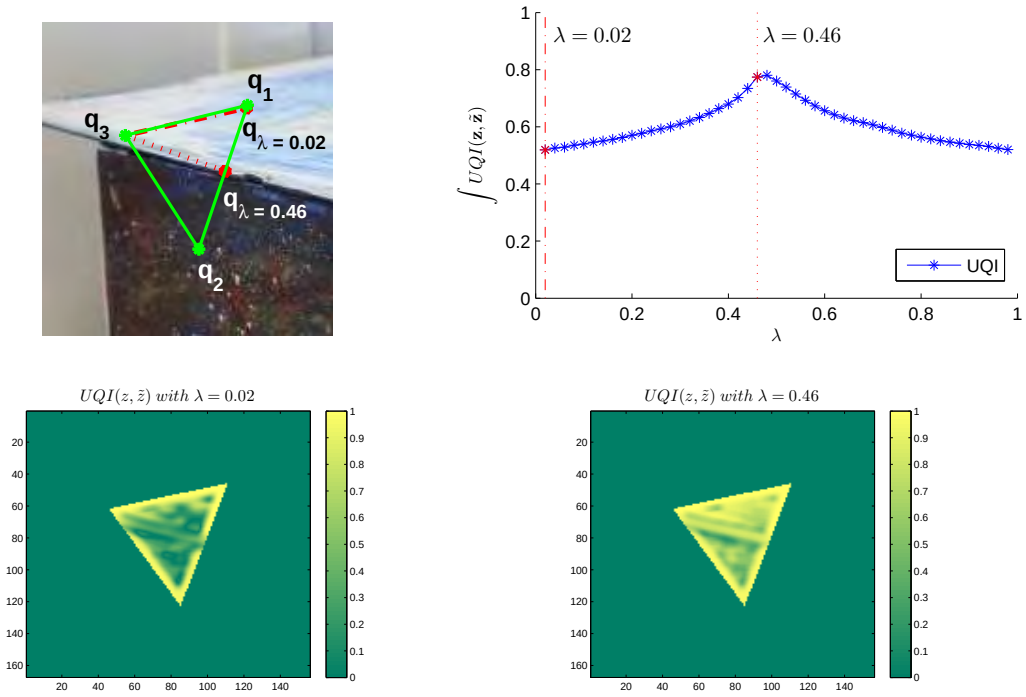


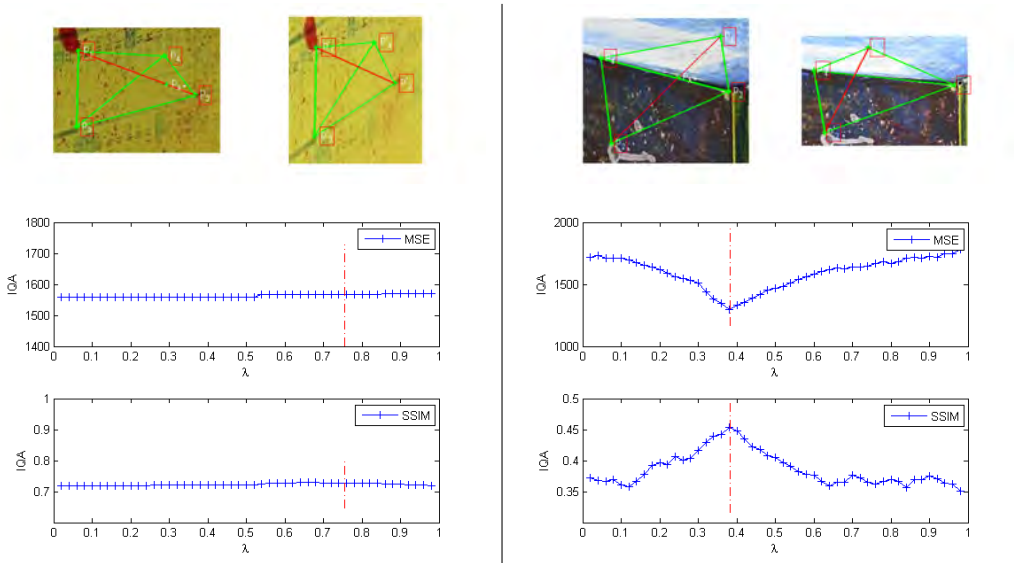
FIGURE 3.10 – Exemple de variation de la valeur d'UQI en fonction de λ dans une zone non-plane : Sur la première ligne, il s'agit de la région d'intérêt de l'image de référence I et de la courbe de la moyenne des valeurs de la mesure en fonction de λ. Sur la deuxième ligne, nous affichons la carte des valeurs de UQI$(\mathbf{z}, \tilde{\mathbf{z}})$ obtenues sur la zone non-plane pour $\lambda = 0.02$ et pour la vérité terrain avec $\lambda^\star = 0.46$ où les orientations des surfaces sont correctement estimées. Globalement, dans le cas idéal, nous obtenons une valeur d'UQI plus importante, comme attendu.

3.4.4 Classification des zones planes

Nous cherchons à identifier les zones planes des zones non-planes afin d'obtenir une meilleure segmentation de la scène observée à partir de plusieurs images. Pour cela, sur chaque zone triangulaire déterminée par au moins trois points 3D, nous estimons l'*a priori* géométrique, fourni par la plan de support. La cohérence photométrique est calculée afin de pouvoir identifier et classifier les zones planes des zones non-planes.

La variations des valeurs des mesures de cohérence photométrique est représentative du cas traité. Nous détaillons comment classer les triangles en étudiant la variation des mesures de similarité en fonction du paramètre λ qui détermine une intersection probable. Dans le cas plan, le critère de similarité/dissimilarité doit être fort/faible et constant quelle que soit la valeur de λ. Dans le cas non-plan, le critère de similarité doit atteindre un maximum global à l'intersection des deux plans. Ces comportements sont illustrés avec deux cas concrets dans la figure 3.11, avec les mesures MSE et SSIM. Toutefois, selon le images traitées, la répartition des valeurs obtenues peut être variable. Comme le montre les boîtes à moustache de la figure 3.12, il est parfois difficile de déterminer clairement la valeur d'ϵ utilisé pour la classification. En effet, nous pouvons remarquer que les deux ensembles de valeurs obtenus pour chacune des deux classes

ont une intersection non-nulle, ce qui rends plus complexe la séparation des données.



(a) Cas plan (b) Cas non-plan

FIGURE 3.11 – Variation de la cohérence photométrique dans un cas plan et un cas non-plan. Représentation des images z, z' et des courbes des valeurs obtenues avec MSE et SSIM en fonction de λ (a) dans le cas d'une zone plane et (b) dans le cas d'une zone non-plane. La droite en pointillés rouges correspond au $\lambda^\star$, vérité terrain de l'intersection des deux plans $\pi_1 \cap \pi_2$. Les deux mesures ont le comportement attendu : valeurs constantes dans le cas plan, et valeurs extremums dans le cas non-plan, pour $\lambda^\star$.

Le but de ces travaux n'est pas de réaliser une étude ou une comparaison des méthodes de classification. Nous avons donc choisi d'appliquer l'algorithme le plus simple pour distinguer les zones planes des zones non-planes : un seuillage par une valeur ϵ.

$$\mathcal{C}(\mathbf{z}, \tilde{\mathbf{z}}) = \begin{array}{ll} \text{NP} & \text{si } \ min(\text{IQA}_s(\mathbf{z}, \tilde{\mathbf{z}})) < \epsilon \\ \text{P} & \text{sinon.} \end{array} \tag{3.13}$$

3.4.5 Résultats et analyses de la classification

Données utilisées. Nous avons mis en place deux bases de données de zones triangulaires mises en correspondance. BD1 correspond à des zones d'images d'une boîte texturée différemment sur chaque face, cf. figure 3.9. Ces images sont acquises en intérieur avec des conditions de lumière maîtrisées. BD2 correspond à des images de scènes d'extérieur issues des bases suivantes : base de données publique d'Oxford [1], images acquises avec le système d'acquisition mobile l'imajbox R [2]. Nous avons retenu 87 zones pour réaliser l'évaluation (29 de BD1, 58 de BD2).

1. www.robots.ox.ac.uk/~vgg
2. www.imajing.eu

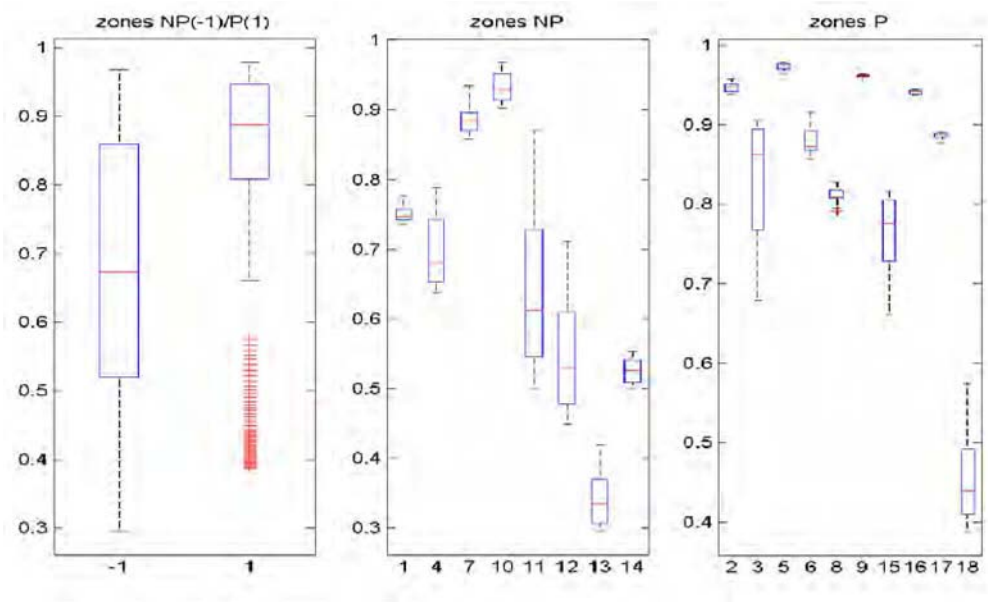


FIGURE 3.12 – Variations des valeurs de cohérence photométrique obtenue avec UQI : pour 18 zones des 87 de la base de données. Nous affichons les boîtes à moustache. De gauche à droite : moyenne et premier quartile en fonction de la vérité terrain dans les cas non-plan et dans les cas plan. Ensuite, nous affichons les détails pour les 18 éléments non-plan et plan.

Résultats. Dans cette étude, nous travaillons avec les images avec la plus haute résolution disponible car les mesures de cohérence photométrique sont plus précises et plus discriminantes, voir la section 3.3. La figure 3.13 présente la classification des zones en fonction de la valeur de similarité inter-images obtenue (faible pour le cas non-plan et élevée pour le cas plan). Les mesures analysées sont évaluées grâce aux courbes PR et ROC présentées dans la section 2.8 et illustrée dans la figure 3.14.

La classification des triangles avec le critère de photo-consistance dépend explicitement du choix de la valeur d'ϵ. Comme on pouvait s'y attendre, dans le cas d'une mesure de dissimilarité, si ϵ augmente alors les quantités de vrais positif (TP) et de faux positifs (FP) augmentent et les quantités de faux négatifs (FN) et de vrais négatifs (TN) diminuent. Cependant, nous cherchons à la fois à maximiser le nombre de TP et de TN et à minimiser le nombre de FP et de FN.

Comme nous l'avons vu, plus le voisinage pris en compte est grand, moins la mesure IQA est significative et discriminante. Que ce soit pour le paramètre r correspondant à la taille du voisinage pris en compte dans les mesures MSE_r, RC_r ou RUQI, ou pour n la taille de la fenêtre de recherche pour trouver la meilleure mise en correspondance dans SSIM, UQI, RUQI : plus la valeur est grande, plus l'erreur introduite est importante dans la quantification de la similarité. Cela implique que même dans les cas non-plans, il est possible d'obtenir une forte similarité.

Analyses. De manière générale, les résultats, présentés dans la figure 3.14, montrent que les mesures s'appuyant sur le produit scalaire (courbes rouges) et utilisant des statistiques de

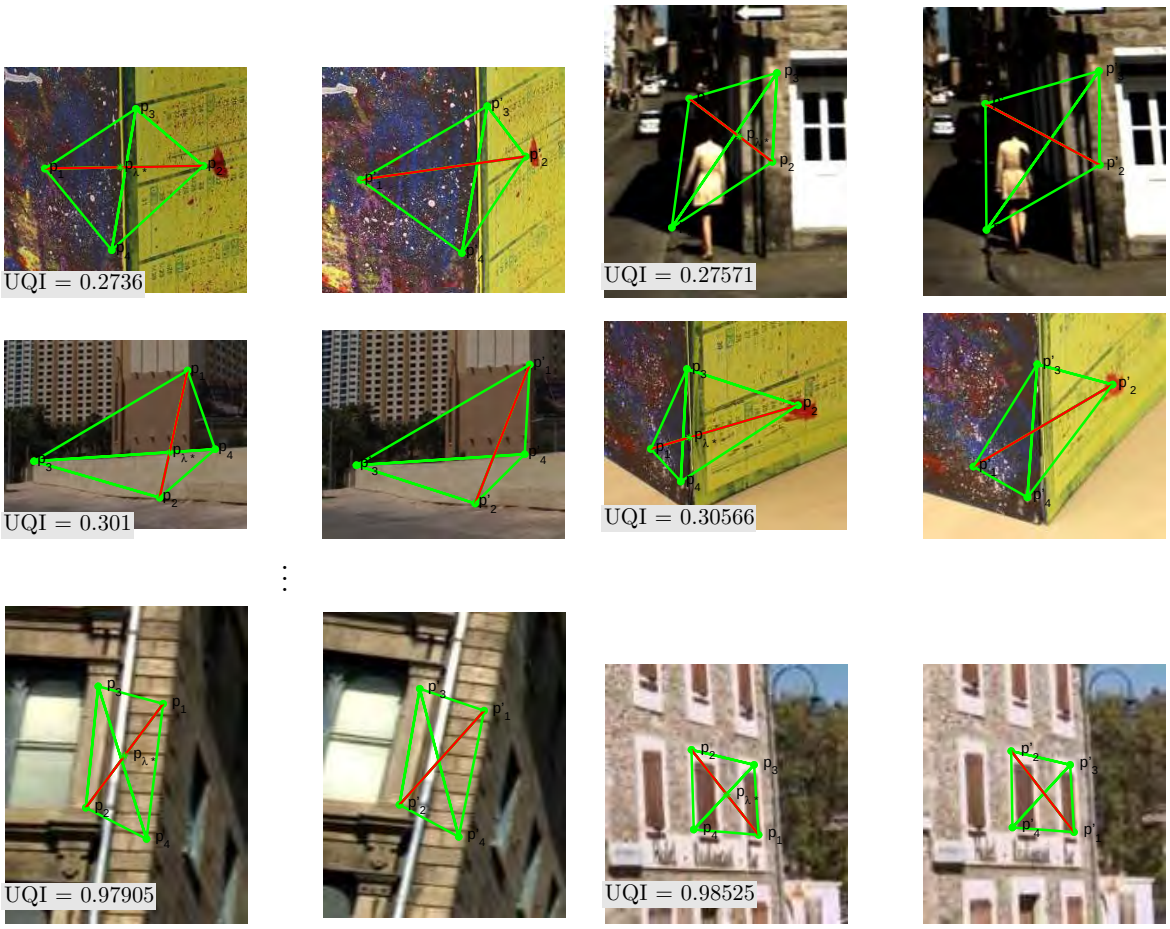


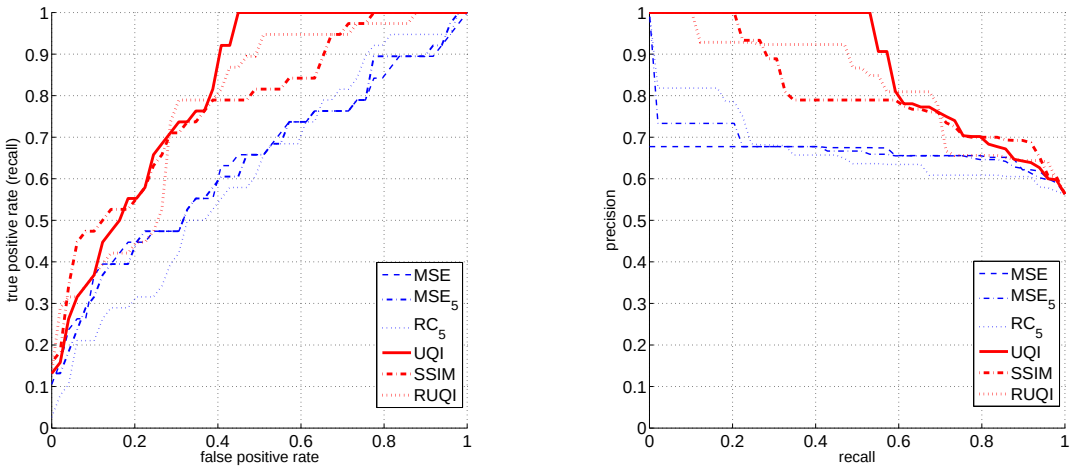
FIGURE 3.13 – Zones classées par valeurs croissantes d'UQI : i.e. de plus en plus similaire. Des faibles valeurs d'UQI sont obtenues pour les cas non-plans représentés dans la première et la deuxième ligne. Les cas plans sont classés lorsque une forte similarité est obtenue.

premier ordre sur le voisinage de chaque pixel, obtiennent de meilleure performance que les mesures s'appuyant sur la distance euclidienne (courbes bleues). Plus particulièrement, les meilleurs résultats sont obtenus avec la mesure UQI dont la courbe surpasse celles des autres.

3.5 Conclusion

Dans ce chapitre, nous avons présenté et analysé les mesures de photo-cohérence symétrique, appelées IQA, en distinguant les mesures s'appuyant sur la distance euclidienne de celle utilisant le produit scalaire, pour la classification de régions en zones planes et non-planes.

Le protocole d'évaluation présenté dans la section 3.4, permet de classifier les zones planes et les zones non-planes. Nous avons mis en avant, qu'un *a priori* de planéité permet, à l'aide d'une mesure de photo-cohérence inter-image comme UQI, de valider ou d'infirmer cette hypothèse. En effet, la mesure UQI semble se distinguer des autres mesures par ses performances dans la classification dues à son caractère intrinsèque de détecteur de contours. Cependant, ses performances sont altérées lors de grand déplacement. Les mesures étant symétriques et traitant



(a) (b)

FIGURE 3.14 – Évaluation de la classification en zone planes/non-planes de 87 triangles avec 6 mesures : (a) la courbe ROC et (b) la courbe PR en fonction du seuil ϵ utilisé lors de la séparation des deux classes.

indifféremment les deux images, cela ne permet pas de déterminer le sens de déplacement des objets d'une image à l'autre.

Chapitre 4

Applications à la segmentation en super-pixels

Sommaire

4.1 Introduction

Nous avons montré dans les chapitres précédents que les approches de super-pixels ne permettent pas toujours de distinguer des surfaces ayant les mêmes caractéristiques photométriques mais avec des orientations différentes. Ainsi, l'objectif de ce chapitre est d'obtenir une sur-segmentation d'images en super-pixels cohérents avec la géométrie de la scène, c'est-à-dire les changements d'orientations des surfaces contenues dans des scènes urbaines. Ceci permet d'être le plus précis possible au niveau de la sur-segmentation afin de minimiser les risques pour l'étape d'étiquetage sémantique.

Nous commençons par une présentation des propriétés attendues d'un super-pixel et des approches de constructeurs existants. Ensuite, nous présentons un critère pertinent pour obtenir une sur-segmentation géométrique des zones faiblement texturées ou avec des textures ambiguës. Nous proposons de prendre en compte explicitement l'information de planéité obtenue lors d'une classification préalable, en distinguant les pixels appartenant à une zone plane de ceux appartenant à une zone non-plane ou sans information géométrique *a priori*. Pour cela, la connaissance de l'orientation des surfaces est renforcée par l'estimation d'un critère de similarité photométrique entre les deux images représentant une même surface de la scène 3D. Ces informations géométriques et photométriques sont intégrées dans l'étape itérative d'agrégation des pixels d'un constructeur de super-pixels connu : *Simple Linear Iterative Clustering* (SLIC). Enfin, cette nouvelle approche de construction de super-pixels est utilisée sur des données réelles telles que des images de scènes urbaines acquises depuis un système d'acquisition en mouvement.

4.2 Sur-segmentation en super-pixels

La sur-segmentation est utilisée dans de nombreuses applications, comme la détection ou la reconnaissance d'objets [Hoiem 05b, Fulkerson 09, Toshev 10], le suivi d'objets d'intérêt [Ren 03, Liu 11]. Des comparaisons de ces approches ont déjà été réalisées [Stutz 15, Neubert 12, Hanbury 08], la plus aboutie étant celle proposée par [Achanta 10] et illustrée dans la figure 4.1. Le but de cette première partie est de partir des algorithmes répertoriés dans ces travaux afin d'identifier plus clairement les propriétés qui ont été exploitées et, en particulier, de mettre en avant le fait qu'il y a peu d'information géométrique, ou plus précisément de notion de planéité, utilisée.

4.2.1 Définition et propriétés

Nous rappelons la définition d'un super-pixel, déjà introduite dans le chapitre 1 : il s'agit d'une région fermée qui correspond à une structure locale et cohérente de niveau intermédiaire permettant de représenter un objet ou une partie d'un objet. Les propriétés attendues d'un super-pixel sont ainsi fortement dépendantes de la traduction mathématique de cette définition.

Ainsi, il est délicat de déterminer ce qu'on appelle « un bon super-pixel » [Bagon 08, Schick 11]. L'article de [Bagon 08], *What is a good image segment ?* propose une définition d'une « bonne

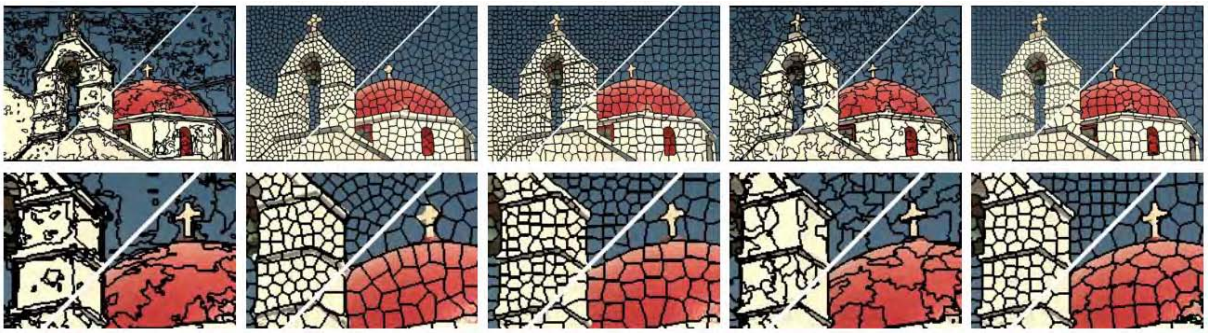


FIGURE 4.1 – Comparaison visuelle de super-pixels : de gauche à droite les super-pixels de [Felzenszwalb 04, Shi 00, Levinshtein 09, Fulkerson 09, Achanta 10]. Cette illustration est extraite de [Achanta 10]. Deux niveaux de résolution sont affichés, le plus fin étant dans le bord supérieur gauche, pour chaque résultat. Les super-pixels sont de formes assez variées : régulier/non régulier, compact/non compact et cela illustre parfaitement la variabilité des résultats obtenus suivant les propriétés attendues et les contraintes fixées lors de la construction.

région » comme le regroupement d'un faible nombre d'entités élémentaires et non triviales. En effet, le nombre d'entités est un problème clé de la segmentation, car découper l'image avec trop peu d'entités correspond au cas d'une sous-segmentation ou avec trop d'entités au cas d'une sur-segmentation.

Dans la suite, nous distinguons trois types de propriétés. Les propriétés d'apparence d'un super-pixel s'appuient sur les aspects photométriques des pixels qu'il contient. Les propriétés spatiales dépendent de la position des pixels et de leurs relations au sein du super-pixel. Enfin, nous présentons quelques propriétés spatio-temporelles qui permettent d'étendre les approches dans le cas de séquence temporelle.

Propriétés d'apparence. Une propriété attendue assez simple est que la ressemblance au sein d'un super-pixel doit être maximale. Le raisonnement complémentaire est de rechercher à respecter une propriété de ressemblance photométrique minimale entre deux super-pixels voisins, en d'autre terme à maximiser la dissimilarité des voisins. Ainsi, les critères photométriques les plus utilisés sont les statistiques de premier ordre (moyenne/écart-type) sur les valeurs d'intensité ou de couleur des pixels afin de privilégier le groupement de pixels similaires [Achanta 10]. Un autre critère consiste à utiliser les zones de forte dissimilarité, comme les **contours** [Moore 08, Moore 09, Levinshtein 09, Veksler 10]. Enfin, les méthodes [Shi 00, Felzenszwalb 04] proposent à la fois de minimiser la similarité entre deux régions tout en maximisant la similarité au sein d'une même région. Ce critère est appelé la différence intra/inter région. L'approche de [Arbeláez 11] combine également les deux problèmes de vision à savoir : la détection des contours et la segmentation en s'appuyant sur une classification des zones en fonction de leur signature dans le domaine fréquentielle. De manière plus anecdotique, nous avons également trouvé dans la littérature des approches qui recherche une propriété de similarité de la **texture** à l'intérieur d'un même super-pixel [Shi 00].

Propriétés spatiales. La **position** des pixels dans l'image est la propriété spatiale la plus utilisée [Shi 00, Felzenszwalb 04, Achanta 12]. La prise en compte de cette propriété pour la construction des super-pixels, permet de contraindre la **connexité** et de renforcer la **compacité** de chaque entité.

La propriété de **compacité** est depuis longtemps évoquée dans la littérature [Moore 09, Levinshtein 09, Liu 11, Achanta 12], mais une définition formelle n'a été proposée qu'en 2012 par [Schick 12]. Le critère de compacité est mesurée à l'aide du rapport isopérimétrique, correspondant au rapport de l'aire du super-pixel sur l'aire du cercle ayant le même périmètre. Les recherches antérieures ont montré que les algorithmes ne considérant pas de critère de compacité spatiale produisent généralement une sous-segmentation, en particulier dans les zones à faible contraste ou dans l'ombre [Levinshtein 09, Li 15].

La propriété de **connexité** est également couramment utilisée [Moore 09, Veksler 10, Duan 15]. Un super-pixel $\mathcal{S}$ est dit connexe si et seulement si pour toutes paires (p_i, p_j) de pixels dans $\mathcal{S}$, il existe un chemin passant uniquement par des pixels de $\mathcal{S}$. Un exemple de prise en compte de cette propriété consiste à effectuer un post-traitement dans les cas de super-pixels non-connexes en rattachant les « petits » (notion de taille limitée par un seuil) super-pixels à son plus « grand voisin » [Achanta 12].

Une autre propriété attendue consiste à conserver une certaine **régularité topologique**, c'est-à-dire un nombre de voisins constants comme dans les approches [Moore 08, Achanta 10]. Ceci peut permettre d'utiliser un modèle markovien [Zhang 10]. De plus, il arrive que la notion de connectivité soit utilisée comme *a priori* sur des approches par coupure de graphe [Vicente 08].

Propriété d'invariance temporelle. Des approches récentes proposent d'intégrer la notion d'invariance temporelle ou de cohérence temporelle des super-pixels le long d'un flux vidéo ou d'un flux d'images [Grundmann 10, Rubio 12]. Dans le cas d'un flux vidéo, la contrainte la plus utilisée s'appuie sur le fait que la position d'un objet entre deux images successives varie très peu et se retrouve dans un petit voisinage de l'image précédente. Par exemple, les auteurs comme [Vazquez-Reina 10] s'appuient sur la propagation non-supervisée d'étiquettes via un modèle *Conditional Random Field* (CRF) sur une grille fine de super-pixels construite de manière indépendante sur chaque image.

En conclusion, toutes les propriétés présentées dans cette section peuvent être prises en compte dans la construction des super-pixels et expliquent en partie les nombreuses approches proposées dans la littérature.

4.2.2 Familles de constructeurs

De la même manière que [Hanbury 08], nous pouvons distinguer trois types d'approches : celles s'appuyant sur la **théorie des graphes** comme [Felzenszwalb 04, Mori 05], celles utilisant des critères **statistiques** [Comaniciu 02, Achanta 12] et enfin, les approches par **croissance de**

germes [Levinshtein 10].

Approches s'appuyant sur la théorie des graphes. Généralement, une certaine régularité est exigée dans la structure afin de faire facilement le lien avec la théorie des graphes. Ainsi, la représentation en graphe peut s'appliquer aussi bien au niveau du pixel de l'image pour construire les super-pixels, que pour représenter les super-pixels et leurs relations avec leurs voisins. Les méthodes d'optimisation les plus utilisées sont la coupure de graphe [Shi 00], et l'approche d'arbre couvrant de poids minimal (*minimmum spanning tree*) [Felzenszwalb 04]. L'article de [Felzenszwalb 04] se distingue des autres car l'algorithme proposé permet de prendre en compte à la fois la similarité au sein d'un super-pixel et la dissimilarité entre deux super-pixels voisins. La régularité imposée par le fait d'utiliser un graphe n'implique pas la propriété de connexité, c'est pourquoi, dans certaines approches, elle peut être renforcée par la notion de voisinage telle qu'utilisait dans MRF ou CRF.

Approches statistiques. Elles s'inspirent des méthodes connues des k-moyennes, *k-means*, et du *mean shift* [Comaniciu 02]. Initialement utilisées dans le contexte de la classification, elles ont étaient introduites dans des approches de construction de super-pixels telles que *quick shift* [Vedaldi 08] ou SLIC [Achanta 10]. Cette dernière approche présente de nombreuses variantes mais nous la développerons dans la section 2.

Approches par croissance de germes. Les turbopixels de [Levinshtein 09] sont les plus connus parmi les approches de cette famille. De la même façon que les approches présentées auparavant, il s'agit d'une approche itérative où les super-pixels sont définis en fonction de leur état à l'itération précédente. Le critère utilisé pour mettre à jour les super-pixels prend en considération un critère d'expansion ou de rétractation dépendant de contraintes extérieures, telles que la courbure des super-pixels ou les positions relatives des super-pixels les uns par rapport aux autres.

Une autre approche de cette famille [Zeng 11] consiste à construire des super-pixels de taille variable en fonction de la structure définie par la densité de contours présents dans l'image (éparse ou dense). En effet, la méthode fournit un nombre de super-pixels proportionnel au nombre de points contours.

Enfin, le constructeur de super-pixels SEEDS, *Super-pixels Extracted via Energy-Driven Sampling* [Van den Bergh 12], propose une approche permettant de diminuer significativement le temps d'exécution. À l'aide d'un processus itératif et d'une initialisation sur une grille régulière, seuls les pixels frontaliers à deux super-pixels voisins peuvent changer de centre de rattachement en optimisant un terme prenant en compte le caractère homogène du super-pixel ainsi que la régularité de sa forme. Plus précisément, l'homogénéité est relative à la fonction de densité de l'histogramme couleur au sein de la région et la régularité de la forme dépend du nombre de super-pixels voisins localement (en chaque pixel frontalier).

4.2.3 Synthèse

Nous avons pu remarquer la grande variabilité existante dans les approches de constructeurs de super-pixels. Certains super-pixels fournissent une sur-segmentation proche de la perception [Felzenszwalb 04] alors que d'autres produisent des régions régulières en taille et en surface [Achanta 12, Mori 05]. Bien que de nombreux algorithmes soient performants, ils peuvent manquer de contrôle sur le nombre de super-pixels ou la compacité [Stutz 15]. Avec le tableau 4.1, nous proposons une synthèse de ces constructeurs.

Fam.	Références	Complexité	Apparence			Propriétés spatiales					Tps.
			Int.	*Rég.*	*Tex.*	*Pos.*	*Com.*	*Con.*	*Top.*	*C.*	
Graphes	[Shi 00]	$O(N^{2/3})$	–		–	–					
	[Felzenszwalb 04]	$O(N\log(N))$	–	–		–					
	[Grundmann 10]	$O(N\log(N))$	–	–		–					–
	[Moore 08]	$O(N^{3/2}log(N))$	–					–	–	–	
	[Veksler 10]		–	–			–	–			
	[Liu 11]	$O(N\log(N))$	–				–	–	–		
Statistiques	[Comaniciu 02]	$O(kN^2)$	–								
	[Vedaldi 08]	$O(N^2)$	–			–					
	[Achanta 12]	$O(N)$	–		–	–	–	–		–	–
	[Wang 12]	$O(N_{it}(k.N)^{1/2})$	–				–	–	–		
	[Arbeláez 11]		–		–	–	–	–	–		
	[Li 15]	$O(N)$	–								
Germes	[Levinshtein 09]	$O(N)$	–			–	–	–	–	–	
	[Zeng 11]	$O(N.N_{it})$	–			–		–			
	[Van den Bergh 12]	$O(N)$	–			–	–	–		–	
	[Duan 15]	$O(N\log(N))$	–			–		–	–	–	

Table 4.1 – Familles (Fam.) d'approches permettant la construction de super-pixels. La complexité dépend des termes N, le nombre de pixels, N_{it}, le nombre d'itérations de l'approche et k, le nombre de super-pixels. Les propriétés listées sont celles présentées dans la section 4.2.1 : statistiques sur l'intensité ou la colorimétrie (*Int.*) la différence intra/inter régions (*Rég.*), utilisation de la texture (*Tex.*), utilisation de la position (*Pos.*). Mais également, nous indiquons si les notions de compacité (*Com.*), de connexité (*Con.*), de régularité topologique (*Top.*) sont prises en compte. Enfin nous précisons si les contours (*C.*) sont utilisés. La colonne Tps. nous permet d'indiquer lorsque l'approche propose une variante adaptée à une séquence d'images et ainsi à utiliser une information dense pour la propagation entre image.

Les approches de sur-segmentation en super-pixels permettent de réduire la complexité des calculs, mais toutes ne fournissent pas le même niveau de détail et surtout le même type de résultat. Quelle que soit l'approche, il arrive que certains super-pixels ne soient pas cohérents avec la géométrie de la scène et une région peut alors couvrir deux zones appartenant à deux objets différents car ils sont très similaires en couleur. Nous souhaitons prendre en compte cet aspect en modifiant une approche de la littérature, à savoir SLIC qui fournit un cadre assez simple pour effectuer des modifications et qui, de plus, est très utilisée dans la littérature. Avant de proposer la modification, nous présentons en détail cette approche SLIC.

4.3 Détails sur l'algorithme *Simple Linear Iterative Clustering* (SLIC)

La méthode proposée par [Achanta 12], *Simple Linear Iterative Clustering* (SLIC), permet la construction de super-pixels en regroupant les pixels d'après leur similarité et leur proximité spatiale. Ceci est réalisé dans un espace à 5 dimensions dont 3 dimensions correspondent à la couleur du pixel dans l'espace couleurs CIELAB, système perceptuellement uniforme, c'est-à-dire qu'il permet d'être le plus proche de la perception humaine et en particulier, de la perception des différences entre deux couleurs. Les 2 dernières dimensions sont relative à la position du pixel dans l'image.

4.3.1 Algorithme

Cette méthode récente et simple, est composée de trois étapes, cf. figure 4.2, décrites dans l'algorithme 2. Elle est de faible complexité algorithmique $O(N)$, où N est le nombre de pixels dans l'image.

Algorithme 2 : Algorithme général de sur-segmentation en super-pixels (SLIC) : l'ensemble des pixels associés au centre $\{C_m\}$ est noté $\{R_m\}$. Le critère d'arrêt de l'étape itérative dépend uniquement du nombre d'itérations maximal est $N_{it} = 10$. Les trois fonctions InitCentres, CalculRegions et RenforceConnexité sont détaillées dans la section 4.3.1.

Données : Image I
Résultat : Image segmentée $\{R_m\}$

$\{C_m\}^1$ InitCentres(I)
pour *i=1 :*N_{it} **faire**
  $\{R_m\}^i$ CalculRegions($I, \{C_m\}^i, D_{SLIC}$)
  $\{C_m\}^{i+1}$ MiseAJourCentres($\{R_m\}^i$)
$\{R_m\}$ RenforceConnexité ($\{R_m\}^k$)

Les centres des super-pixels sont déterminés par InitCentres(I) suivant une grille régulière dont la dimension $S \times S$ est fixée par le nombre de pixels de l'image N et le nombre de super-pixels souhaités K avec $S = \overline{N\ K}$. L'approche proposée par [Wang 12] est très similaire à SLIC mais diffère à l'initialisation en utilisant des régions hexagonales et non rectangulaires.

Les super-pixels sont le résultat d'un processus itératif en deux étapes. Tout d'abord, la fonction CalculRegions permet l'agrégation des pixels à un centre dans un voisinage $2S \times 2S$ par minimisation d'une distance D_{SLIC}. Nous développons cette distance dans la section 4.3.2. Cette première étape génère donc une nouvelle version des super-pixels. Puis la fonction MiseAJourCentres met à jour la position des centres en fonction de la moyenne des pixels associés à chaque super-pixel construit à l'étape précédente. Le cas d'arrêt est soit fixé par le nombre d'itérations, N_{it}, (les auteurs le fixe de manière empirique), soit jusqu'à l'obtention d'une faible variation des positions des centres entre deux itérations successives. En pratique, les auteurs ont proposé un

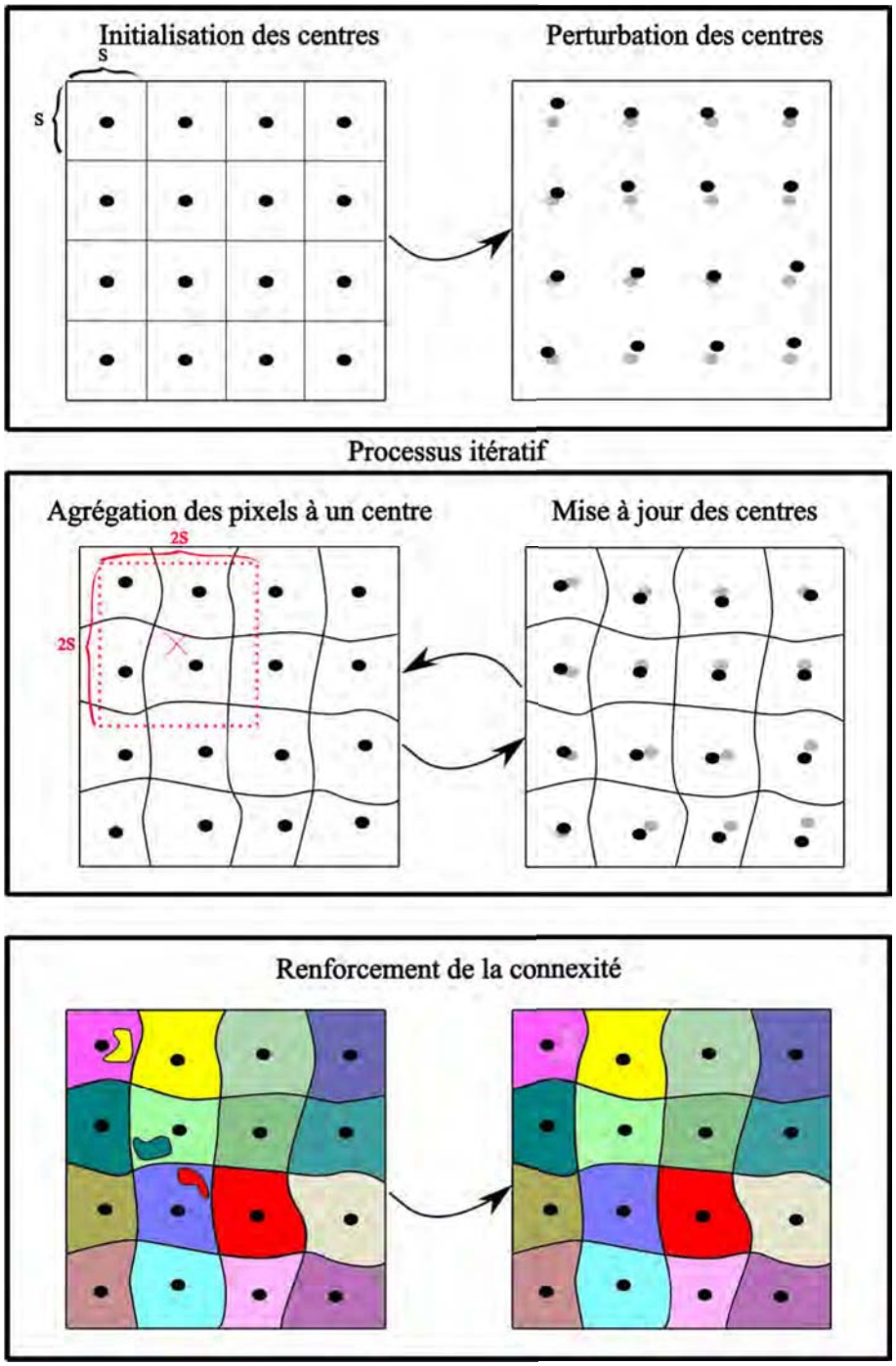


FIGURE 4.2 – Illustration de l'algorithme sur le constructeur de super-pixels SLIC : les centres sont initialisés suivant une grille régulière de taille $S \times S$. Ils sont ensuite éventuellement faiblement déplaces afin de ne pas être localisés sur un contour. La construction des super-pixels est un processus itératif en deux étapes : agrégation des pixels à un centre en minimisant une distance et mise à jour des centres. La dernière étape consiste à renforcer la connexité en fusionnant des super-pixels de taille non significative (seuil à déterminer) à des super-pixels voisins et de taille significative.

code qui fixe $N_{it} = 10$.

Lors de la dernière étape, le renforcement de la connexité (fonction RenforceConnexité), consiste à rattacher les entités les plus petites, i.e. celles dont la surface est inférieure à $\frac{N}{4K}$ pixels, au super-pixel voisin le plus grand.

4.3.2 Distance d'agrégation

La distance d'agrégation D_{SLIC} est définie par la somme pondérée de deux distances euclidiennes relatives respectivement à la couleur et à la position du pixel étudié, p_j, par rapport à un centre donné, p_i :

$$D_{SLIC}(p_j, p_i) = d_c(p_j, p_i) + \frac{m}{S} d_s(p_j, p_i), \tag{4.1}$$

où

- $d_c(p_j, p_i) = (l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2$ distance couleur entre le centre p_j, de couleur (l_j, a_j, b_j) et le pixel courant p_i, de couleur (l_i, a_i, b_i),
- $d_s(p_j, p_i) = (x_j - x_i)^2 + (y_j - y_i)^2$ distance de position entre le centre p_j, de position (x_j, y_j) et le pixel courant p_i, de position (x_i, y_i),
- m, paramètre de compacité,
- $S = \quad \overline{N\ K}$, constante spatiale.

Lorsque l'espace couleur CIELAB est utilisé, le paramètre de compacité m varie dans l'intervalle [0,40]. Plus la valeur de m est grande, plus les super-pixels sont compacts. La difficulté réside dans le fait de trouver un bon équilibre entre ces deux termes (couleur, position). Les auteurs préconisent d'utiliser $m = 10$ comme valeur par défaut.

4.3.3 Variante sans paramètre

Les auteurs de cet algorithme proposent une extension sans paramètre de compacité, appelée SLICO (pour SLIC zéro). Alors que SLIC utilise le même paramètre de compacité pour toute l'image, avec SLICO l'utilisateur n'a plus besoin d'initialiser cette valeur car elle est choisie de manière dynamique et adaptative pour chaque itération. En effet, pour chaque centre, à chaque itération n, les distances $d_{c,n}$ et $d_{s,n}$ sont normalisées par les valeurs maximales obtenues à l'itération précédente, c'est-à-dire : $m_{s,n-1} = \max(d_{s,n-1})$ et $m_{c,n-1} = \max(d_{c,n-1})$.

Dans ce cas, la distance d'agrégation devient :

$$D_{SLICO,n} = d_{c,n} + d_{s,n} \tag{4.2}$$

avec

$$\begin{aligned} d_{s,n}(p_j, p_i) &= \frac{d_{s,n}}{m_{s,n-1}} \text{ avec } d_{s,0} = d_{s_0} \\ d_{c,n}(p_j, p_i) &= \frac{d_{c,n}}{m_{c,n-1}} \text{ avec } d_{c,0} = d_{c_0}. \end{aligned} \tag{4.3}$$

où d_{s_0} et d_{c_0} sont les valeurs par défaut utilisées dans SLIC. Dans la suite de ce travail, nous avons

choisi d'utiliser cette variante afin de ne pas avoir à choisir manuellement ou par apprentissage le paramètre de compacité et nous allons introduire une contrainte géométrique s'appuyant sur la notion de planéité dans la distance d'agrégation utilisée.

4.4 Introduction de l'approche de super-pixels géométriques

L'objectif du travail présenté est d'obtenir une décomposition de l'image I en un nombre inconnu K d'entités géométriquement et sémantiquement cohérentes à partir de plusieurs vues calibrées d'une scène qui peut être considérée plane par morceaux. Nous disposons des paramètres de la caméra et d'une reconstruction éparse de points 3D correspondant à la structure de la scène. Ces points peuvent être obtenus par une approche SfM comme dans [Wu 11a], cf. chapitre 2.

Dans la suite, nous décrivons l'information géométrique que nous utilisons et comment nous l'intégrons dans l'étape d'agrégation des pixels en super-pixels. Enfin, nous proposons une évaluation de cette nouvelle approche de construction de super-pixels, en analysant l'influence des choix à faire pour les différents paramètres.

4.4.1 Extraction et intégration de l'information géométrique

Nous décrivons ici l'information géométrique que nous proposons d'extraire afin de l'intégrer au constructeur de super-pixels [Achanta 12]. Plus précisément, nous proposons de construire deux cartes : une carte de planéité et une carte de similarité, comme illustrées dans la figure 4.3.

Carte de planéité. Tout d'abord, nous supposons que nous sommes capables d'estimer les plans dominants de la scène avec une approche classique de type RANSAC, cf .annexe 4.7. Ensuite, il s'agit d'en déduire la normale $\vec{n}$ à la surface 3D plane à laquelle appartient chaque pixel p. Pour des raisons de simplification, la carte de planéité représente l'étiquette d'appartenance à un plan, i.e. le numéro du plan pour lequel la valeur de similarité photométrique est maximale.

Cartes de similarité. Pour chaque région plane, connaissant les paramètres du plan et la géométrie épipolaire, nous pouvons estimer l'homographie induite par le plan de support, comme nous l'avons proposé dans [Bauda 13]. Cette homographie permet de calculer la région recalée $\tilde{z}$, alignée sur la région de référence. Les deux zones z et $\tilde{z}$ sont alors comparées avec une mesure IQA. Dans le chapitre 3, nous avons montré que les mesures s'appuyant sur le produit scalaire sont les plus performantes pour réaliser la classification en zone plane et en zone non-plane. Plus précisément, la mesure UQI de [Wang 02], un cas particulier de SSIM, donne les meilleures résultats dans nos expérimentations [Bauda 15b]. C'est donc cette mesure qui nous fournit la carte de similarité finale que nous utilisons.

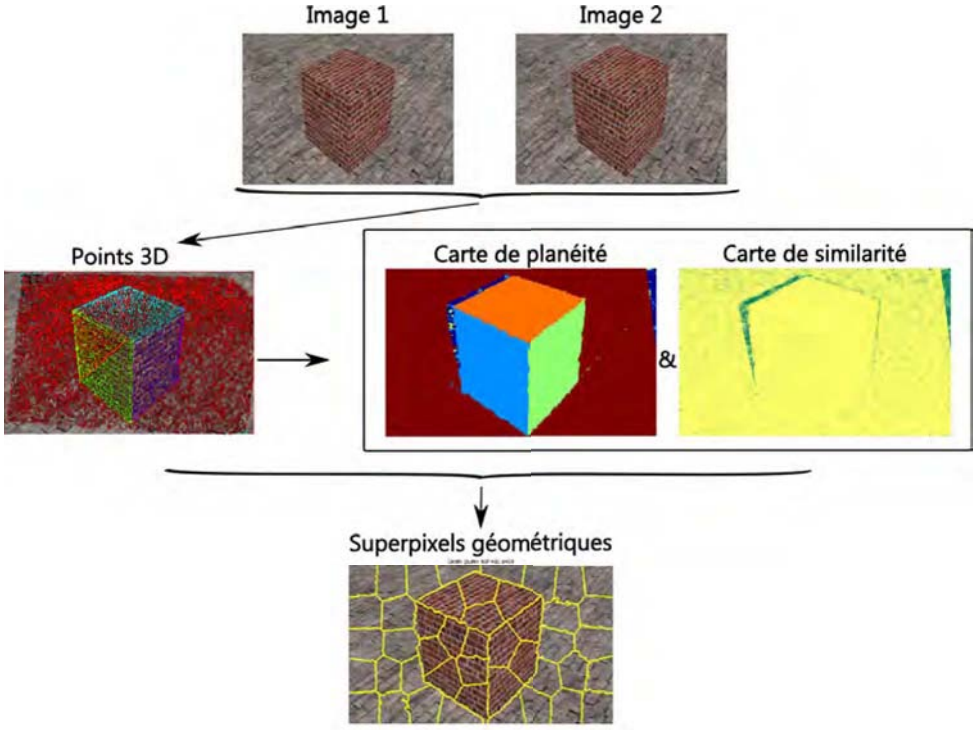


FIGURE 4.3 – Intégration de l'information géométrique dans la construction de super-pixels : La première ligne présente les deux images couleurs calibrées d'entrée (ici, il s'agit de simples images de synthèse. La seconde ligne présente la projection du nuage de points 3D dans l'image 2D, puis, la carte de planéité, qui indique l'appartenance de chaque pixel à un plan et, pour terminer, la carte de similarité inter-image qui correspond à la valeur d'UQI. Le calcul de ces deux dernières cartes est présenté dans la section 4.4.1. La dernière ligne expose enfin le résultat de la méthode de sur-segmentation géométrique cohérente avec les surfaces de la scène.

Détection des pixels non-fiables dans les deux cartes. Quelle que soit la carte calculée, il nous est apparu important d'éliminer tous les cas où le résultat obtenu pour la carte de planéité ou la carte de similarité ne semble pas fiables. C'est-à-dire, dans les deux cas, les pixels ayant une faible valeur de similarité (inférieure à un seuil $< \epsilon$ à fixer) ne sont pas considérés nous leur affectons la valeur $\varnothing$.

4.4.2 Distance d'agrégation proposée

Les deux cartes présentées précédemment sont intégrées dans la construction de super-pixels géométriques, et plus précisément, la distance d'agrégation que nous proposons. Cette nouvelle énergie à minimiser intègre l'information géométrique grâce à l'utilisation du terme d_g. Le paramètre $\lambda \in [0, 1]$ pondère les deux termes (D_{SLICO} et d_g) de la somme qui composent la distance d'agrégation proposée :

$$\begin{aligned} D_{GEOM} &= \sqrt{\lambda.D_{SLICO} + (1-\lambda).d_g} \\ &= \sqrt{\lambda.(d_{c,.} + d_{s,.}) + (1-\lambda).d_g}. \end{aligned} \tag{4.4}$$

Ce terme d_g prend en compte la géométrie de la scène en fusionnant l'information donnée par les cartes de planéité et de similarité. Il est défini par :

$$d_g(p_j, p_i) = 1 - d_n(p_j, p_i).d_{UQI}(p_j). \tag{4.5}$$

où

$$d_{UQI}(p_j, p_i) = UQI(p_j).\mathbb{1}_{UQI>}$$

La distance d_{UQI} correspond à la valeur de similarité inter-images fournie par la carte de similarité dans laquelle la valeur au niveau du germe considéré est maximale. La distance entre normales est notée d_n et indique l'appartenance d'un germe au même plan qu'un pixel voisin. Il est possible de formuler cette distance d_n dans deux situations différentes : le cas particulier exact et le cas général simplifié.

Cas particulier exact. Si nous supposons que l'orientation des surfaces est connue en chaque pixel de l'image, comme avec les données de synthèse présentées dans [Bauda 15a], alors nous savons que la variation d'orientation est donnée par le cosinus de l'angle que forme les normales en ces deux points :

$$d_n(p_j, p_i) \quad = \quad \frac{1 + cos(\vec{n}_j, \vec{n}_i)}{2} \tag{4.6}$$

Cas général simplifié. Dans le cas simplifié, les orientations des surfaces reconstruites sont obtenues par approximations. Nous ne considérons que deux états : soit deux pixels appartiennent au même plan (1), soit ils n'appartiennent pas au même plan (0). Nous ne considérons pas les variations d'orientations mais seulement l'appartenance des pixels à un même plan :

$$d_n(p_j, p_i) = \begin{array}{ll} 1 & \text{si } \vec{n}_i = \vec{n}_j \\ 0 & \text{sinon} \end{array}$$

Le comportement des trois termes $d_{s,.}$, $d_{c,.}$ et d_g de la distance proposée D_{GEOM} est illustré figure 4.4. Nous observons bien que, comme attendu, c'est le terme d_g qui met en évidence les deux plans présents dans la zone étudiée.

4.4.3 Analyse du comportement des super-pixels géométriques proposés

Nous commençons par une analyse de l'influence des choix des paramètres. Puis, nous évaluons notre constructeur de super-pixels géométriques de deux façons différentes. Tout d'abord, globalement sur toute l'image où nous avons été capable de calculer les cartes de similarité et de planéité en étudiant la boîte englobante de l'enveloppe convexe formée par les points d'intérêt utilisés dans le calcul de l'homographie. Ensuite, de manière locale, en ciblant les zones d'intérêt pour lesquelles, notre approche fournit un résultat supérieur à celui fourni par SLICO. À partir

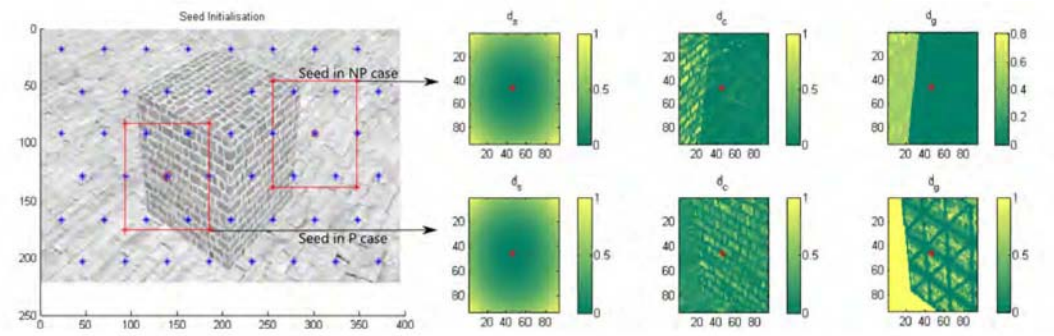


FIGURE 4.4 – Analyse du comportement des termes utilisés dans la distance d'agrégation proposée : Nous rappelons que plus la couleur du pixel est proche du vert, plus la distance est petite. Nous visualisons les valeurs obtenues pour d_{s_0}, d_{c_0} et d_g dans deux cas où l'information spatiale et colorimétrique ne sont pas suffisantes pour discriminer les deux surfaces adjacentes et la distance géométrique permet de distinguer les deux surfaces. Sur la première ligne : le germe se trouve sur le plan support (sol) alors que sur la deuxième, le germe appartient à une face de la boîte.

de maintenant, nous noterons l'approche que nous avons proposée GEOM.

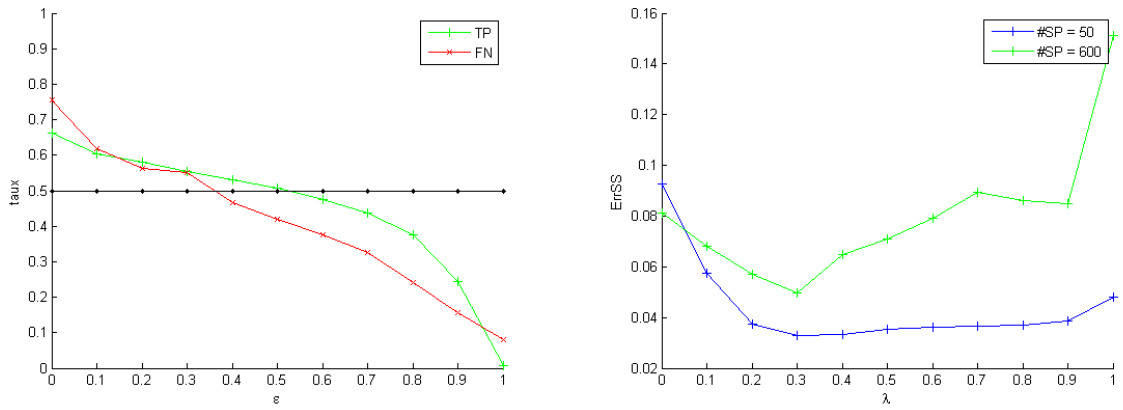
Analyse des paramètres. Nous constructeur nécessite le choix de trois paramètres, à savoir, le nombre de super-pixels attendus, le paramètre de pondération λ et la valeur du seuil de similarité ϵ. Nous avons choisi de faire varier le nombre de super-pixels entre 50 et 600 car cela nous permet de couvrir un ensemble de valeurs suffisant pour étudier le comportement en fonction du choix pour ce paramètre. Lorsque le nombre de super-pixels diminue, les contours fournis par l'information géométrique sont conservés, la densité de super-pixels diminue et leur taille augmente.

La valeur du paramètre λ influence la pondération entre les deux termes D_{SLICO} et d_g. Lorsque $\lambda = 1$ notre énergie correspond à l'approche SLICO, et dans le cas où $\lambda = 0$ seule l'information géométrique est prise en compte. Ce dernier cas suppose que l'information géométrique extraite ne comporte pas d'erreur. Concrètement, il est important de conserver un certain équilibre entre les deux aspects.

La valeur de ϵ détermine le seuil à partir duquel les pixels ayant une valeur de similarité inférieure à ce seuil sont considérés comme peu fiables et ne doivent pas être pris en compte dans le calcul de la distance. Ils sont alors affectés à la classe de rejet, où l'information n'est pas suffisante pour prendre une décision sur leur appartenance ou non à un plan donné de la scène. Si $\epsilon = 0$ cela revient à négliger l'information géométrique. Dans le cas où ϵ est faible, nous considérons que l'information géométrique est peu fiable. Si $\epsilon = 1$, nous considérons que l'information géométrique est exacte et sans erreur, cela n'est valable que dans très peu de cas particulier, car il faudrait une mise en correspondance 2D $\leftrightarrow$ 3D exacte et une similarité inter-image sans bruit et à très haute-résolution.

Nous avons fait varier tous les paramètres listés pour un jeu d'images représentatif. Les

évaluations quantitative sont représentées dans la figure 4.5. Le choix de la valeur ϵ se fait sur le taux de pixels bien classés (TP et FN) lors de l'estimation de la carte de planéité (dans les première étape de notre algorithme). La figure 4.5a représente ces deux taux en fonction d'ϵ. Nous choisissons pour la suite de l'évaluation $\epsilon = 0.4$, valeur pour laquelle nous obtenons le meilleur compromis en terme de résultats obtenus, c'est-à-dire un peu plus de 50% de pixels bien détectés et un peu moins de 50% d'erreur pour le taux de pixels étiquetés à tord sur les autres classes. Le choix de la valeur de λ s'effectue sur l'erreur de sous-segmentation lors de la construction des super-pixels. La figure 4.5b représente l'erreur de sous-segmentation en fonction du paramètre λ pour 50 et 600 super-pixels calculés avec $\epsilon = 0.4$. Pour 50 super-pixels, il y a une faible variation de valeur autour du minimum atteint, alors que pour 600 super-pixels un minimum global est atteint pour $\lambda = 0.3$. Dans les expérimentations qui suivent, nous avons donc choisi un compromis avec $\lambda = 0.4$.



(a) (b)

FIGURE 4.5 – Analyse du comportement de notre approche en fonction des paramètres ϵ et λ : En (a) les taux des TP (*true positive*) et des FN (*false negative*) en fonction d'ϵ. En (b) erreur de sous-segmentation en fonction de λ lorsque $\epsilon = 0.4$ pour 50 et 600 super-pixels.

Analyse globale. Une comparaison visuelle des résultats entre SLICO et GEOM est donnée dans la figure 4.6. Nous pouvons noter que les super-pixels obtenus par GEOM, comme attendu, adhérent plus précisément aux contours correspondant au changement de surfaces.

Une comparaison quantitative a également été réalisée afin d'évaluer la sur-segmentation à l'aide de l'erreur de sous-segmentation et de la valeur de rappel. Représentées sur les graphiques de la figure 4.7, ces mesures qui ont été détaillées dans la partie décrivant l'évaluation 2.8 montrent que le rappel est plus élevé avec GEOM et que l'erreur de sous-segmentation est plus faible avec GEOM également. Ainsi, nous obtenons de meilleures performances qu'avec les super-pixels de l'état de l'art SLICO. Nous pouvons noter que pour un nombre de super-pixels donné, par exemple entre 100 et 200, l'erreur relative vaut entre $\frac{0,5-0,43}{0,43} \quad 0,16$ et $\frac{0,57-0,52}{0,52} \quad 0,10$, c'est-à-dire que notre approche améliore de 10 à 16% le rappel de la sur-segmentation.

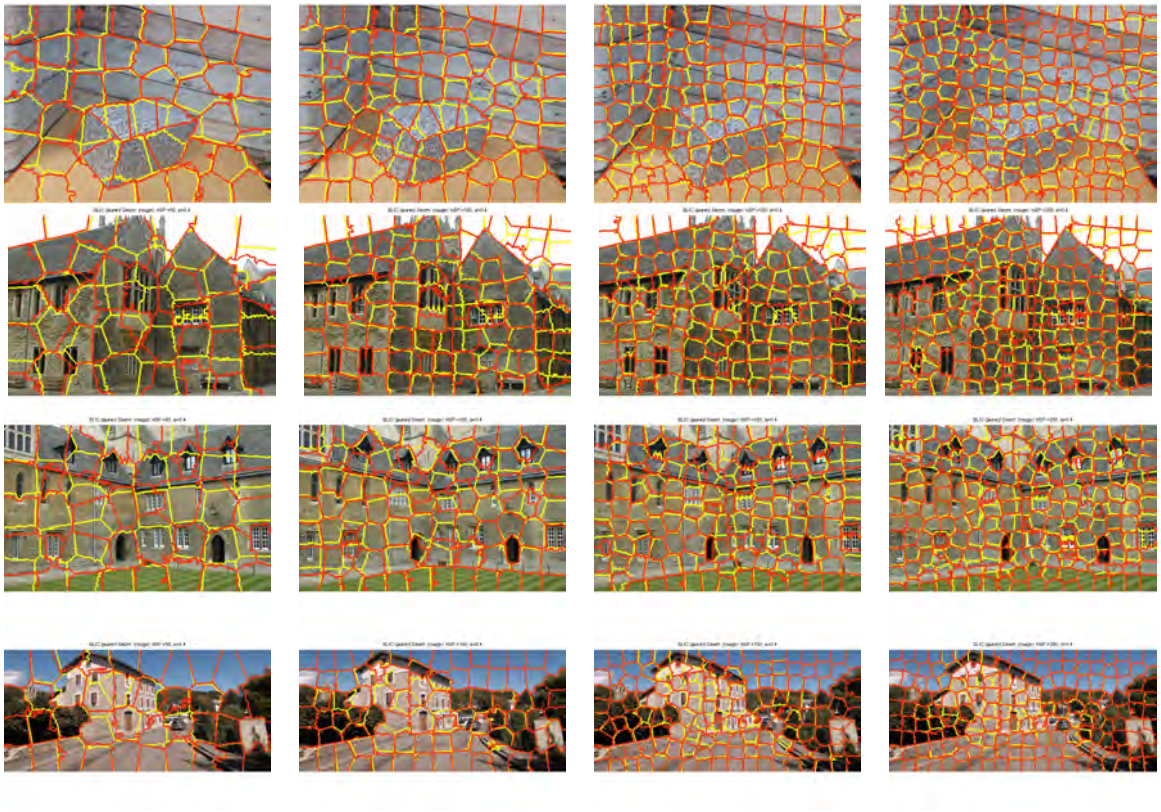


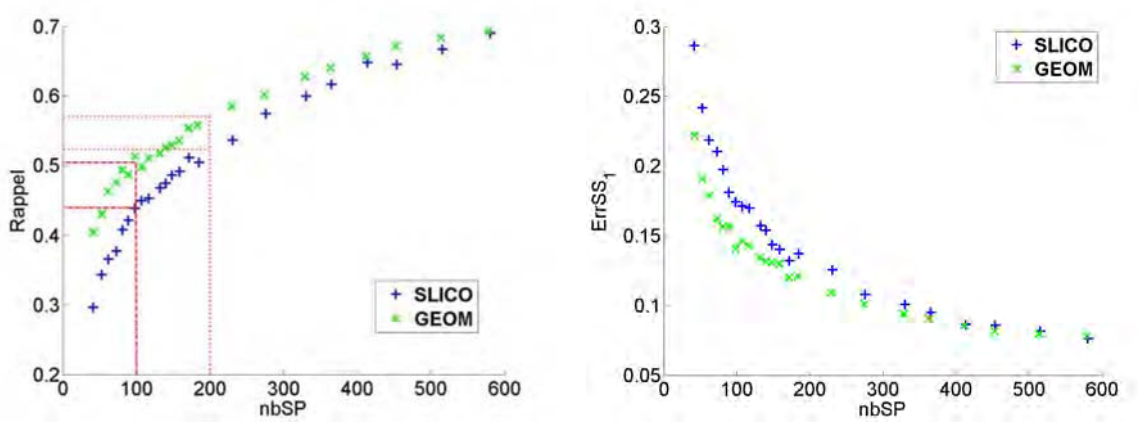
FIGURE 4.6 – Comparaison visuelle de l'approche SLICO et de notre approche GEOM : Sur 4 images de notre corpus, nous visualisons les contours des super-pixels obtenus avec l'approche SLICO (en jaune) et notre approche GEOM (en rouge), pour 50, 100, 150 et 200 super-pixels.

L'amélioration est plus significative pour les sur-segmentations composées d'un faible nombre de super-pixels. Or, nous savons que la difficulté et l'enjeu des approches de sur-segmentation réside dans l'amélioration des performances pour un nombre de super-pixels raisonnable (une ou deux centaines maximum dans notre corpus, mais, cela dépend du type des scènes étudiées) afin de réduire la complexité algorithmique du processus global.

Analyse locale. Pour permettre une analyse plus approfondie du comportement de GEOM, nous détaillons les résultats obtenus localement sur des zones contenant des contours entre deux surfaces ayant des orientations différentes et dont la texture est proche, voire similaire. La figure 4.8 présente les résultats obtenus. Pour cela nous utilisons deux mesures : le rappel, défini de la même manière que pour l'analyse globale, et la moyenne des distances des points de contours de la vérité terrain aux points de contours obtenus par la sur-segmentation proposée. Cette moyenne, notée $m_{\text{dContours}}$ est définie par :

$$m_{\text{dContours}} = \frac{1}{N_{\text{VT}}} \sum_{p_i \text{ VT}} d_i(p_i) \tag{4.7}$$

où $d_i(p_i)$ est la distance du pixel p_i appartenant à la vérité terrain, notée VT, au contour le plus proche proposé par la sur-segmentation calculée. Le terme N_{VT} est le nombre de pixels sur le



(a) (b)

FIGURE 4.7 – Rappel et erreur de sous-segmentation obtenus pour SLICO et notre approche GEOM : Les courbes représentent (a) le rappel et (b) l'erreur de sous-segmentation. Plus de détail est donné dans le texte.

contour de la vérité terrain dans la zone analysée.

Ainsi, l'analyse locale montre que les super-pixels géométriques améliorent la qualité de la sur-segmentation obtenue, en particulier aux frontières entre deux surfaces de texture similaire mais d'orientations différentes. Dans la partie suivante, nous proposons d'utiliser cette sur-segmentation en super-pixels, GEOM, à la segmentation sémantique de scènes urbaines.

4.5 Méthode de segmentation de scènes urbaines utilisant les super-pixels géométriques

4.5.1 Vue générale de l'approche proposée

Notre algorithme 3 est illustré par la figure 4.9. Ainsi, les données 2D et 3D sont utilisées pour extraire deux types d'informations : les plans et la similarité inter-images appelée également cohérence photométrique, cf. l'approche GEOM exposée dans la section 4.4. Il reprend tous les éléments déjà abordés dans cette précédente section, mais, cette fois, nous explicitons tous les aspects spécifiques à l'analyse de scènes urbaines réelles : l'estimation des plans avec RANSAC et la dernière étape de segmentation sémantique.

4.5.2 Classification des points 3D pour l'estimation des plans

Nous utilisons ici une approche s'appuyant sur la méthode de RANSAC, dont l'algorithme générale est détaillé dans l'annexe 4.7. Dans notre contexte de segmentation sémantique en plans, cette approche permet d'estimer les paramètres du plan dominant à partir d'un nuage de points 3D de la scène à analyser. L'algorithme RANSAC a pour paramètre le seuil de tolérance utilisé pour valider l'appartenance d'un point au plan hypothétique, qui correspond à l'erreur

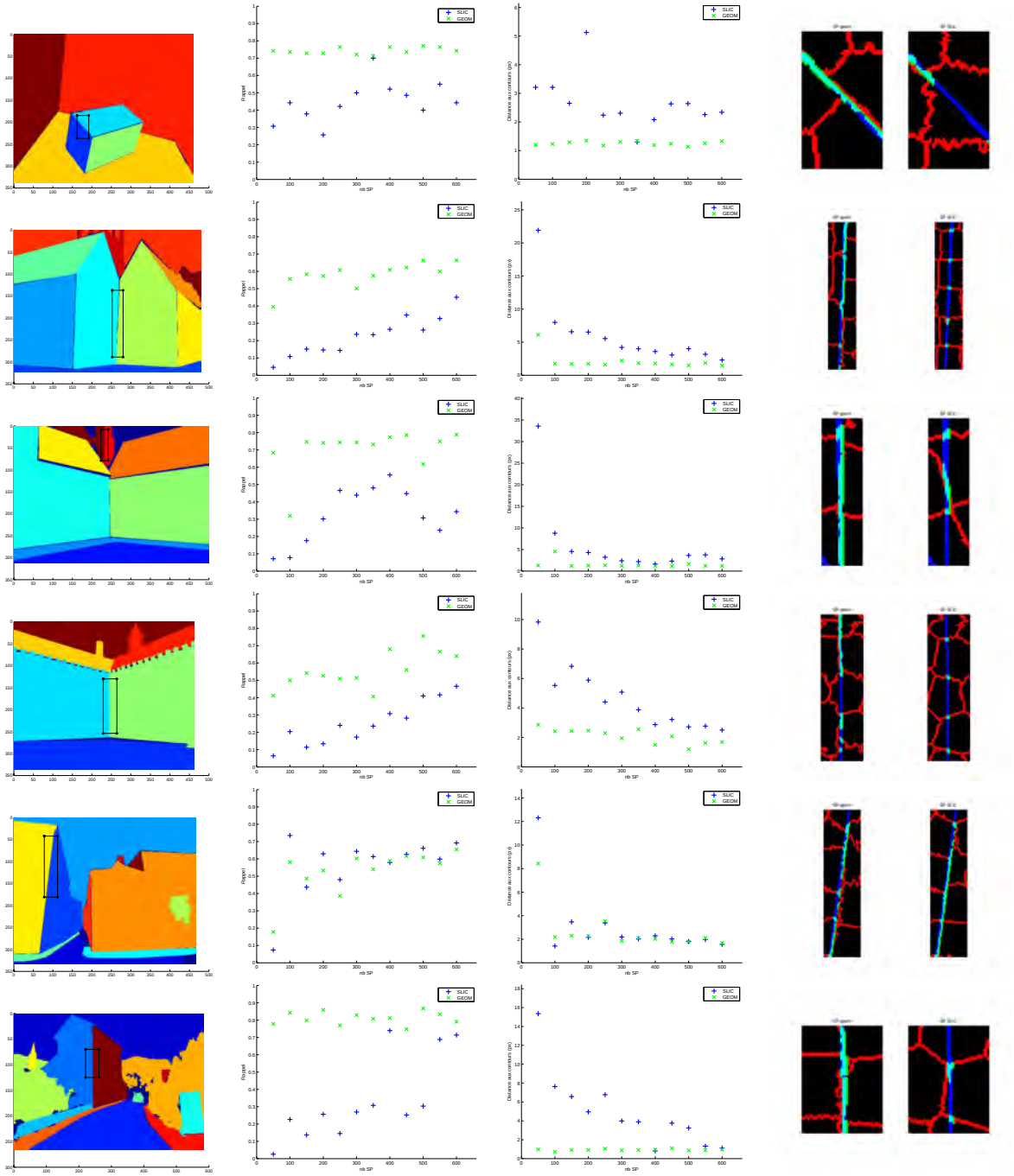


FIGURE 4.8 – Comparaison des sur-segmentations obtenues à la frontière entre deux surfaces de même texture et d'orientations différentes : GEOM (× vert) et SLICO (+ bleu). De gauche à doite : la vérité terrain avec la zone analysée encadrée en noir, la valeur du rappel en fonction du nombre de super-pixels, la distance aux contours $m_{\text{dContours}}$ telle que définie dans l'équation 4.7, et enfin, visualisation du résultat de sur-segmentation obtenue pour 200 super-pixels avec GEOM et SLICO. De plus, la vérité terrain est en bleu (faux négatifs), les pixels bien détectés en vert (vrais positifs) et les autres pixels en rouge (faux positifs).

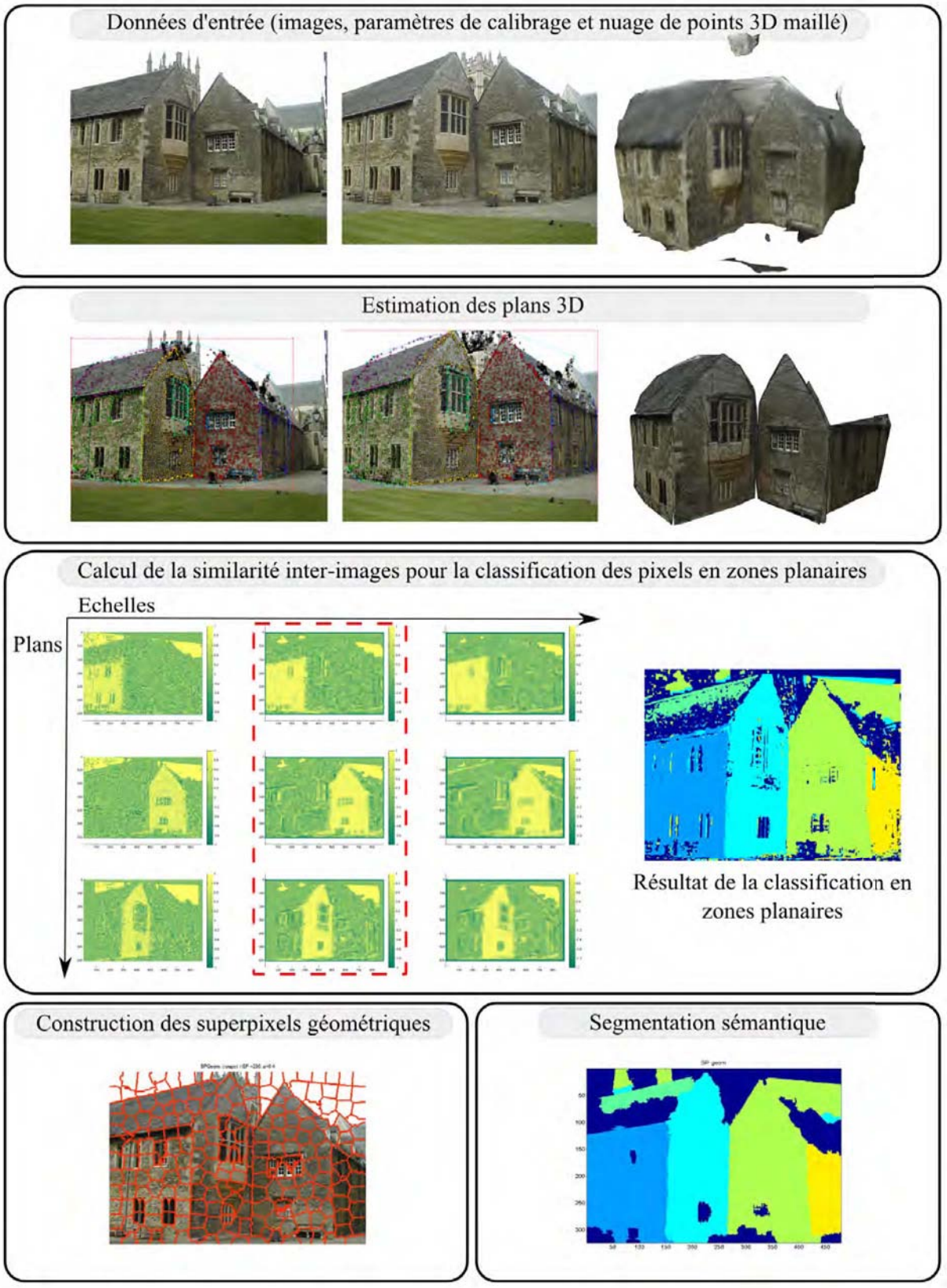


FIGURE 4.9 – Vue d'ensemble de l'application de la segmentation aux scènes urbaines : les détails sont donnés dans l'algorithme 3 et la section 4.5.1.

Algorithme 3 : Vue générale de l'algorithme proposé de segmentation en plans d'une scène $\mathcal{S}$, utilisant des super-pixels géométriques.

Données : Images $\{I, I'\}$ de la scène $\mathcal{S}$, matrices de projection $P = K[Id|0]$ et $P' = K[R|t]$ de chaque image, nuage de points 3D $\{xyz\}$, points d'intérêt 2D appariés $\{xy\} \in I$ et $\{x'y'\} \in I'$.

Résultat : Segmentation $\mathcal{C}$ de l'image I représentant la scène $\mathcal{S}$.

// 3D : Estimation de np plans dans la scène

$\Pi_{1:np}$ ClassificationPtsPlan(xyz)

// 2D : Calcul de la similarité inter-images pour chaque plan

pour *i=1 :np* **faire**

- $(xy_i, x'y'_i)$ ExtractionPtsPlan($xy, x'y', \Pi_i$)
- H_i CalculHomographie($xy_i, x'y'_i, P'$)
- $\tilde{I}_i$ $H_i(I')$
- IQA_i CalculIQA($I, \tilde{I}_i$)

// Classification des pixels aux zones planes pour la segmentation

(L, V) $\max(IQA_{1:np})$

$\mathcal{C}$ $L(V > \epsilon)$

// Construction des super-pixels géométriques

SP SuperpixelsGeom($I, \mathcal{C}, IQA$)

// Segmentation sémantique

Lf SegSem($I, \mathcal{C}, IQA_{1:np}$)

autorisée dans l'estimation du plan et qui est notée ϵ. Nous avons étendu cette approche par ces trois ajouts :

- possibilité qu'un point puisse appartenir à plusieurs plans, en utilisant le critère de recouvrement des plans proposé dans [Bartoli 07] ;
- maximisation du nombre de points dans le plan, mais également minimisation de la distance moyenne des points au plan hypothétique ;
- détection de plans multiples, en itérant simplement N_p fois RANSAC où N_p est le nombre de plans maximum à détecter.

Les deux premières contributions permettent de renforcer la robustesse de la détection du modèle.

Les choix pour ϵ et N_p sont assez délicats et de emprique, nous avons choisi la combinaison qui permet d'obtenir les meilleurs résultats.

4.5.3 Segmentation sémantique en plans

La dernière étape de notre approche consiste à affecter une étiquette à chaque super-pixel. Pour cela nous proposons pour chaque super-pixel de transférer l'étiquette la plus fréquente obtenue au niveau pixel lors de la classification en plans.

4.6 Expérimentations

4.6.1 Données testées

Nous disposons de 9 séquences d'images 2D d'une même scène 3D, cf. tableau récapitulatif 4.2 :

- 1 séquence de synthèse représentant une boîte posée sur un support ;
- 1 séquence réelle représentant le même cas ;
- 7 séquences réelles de scènes urbaines de deux bases de données différentes.

	Scène		Nb images	Nb Pts		Err	VT
				éparse	dense		
1		boxSyntetic	13	4593	28804	1.23	x
2		boxTextured	27	1846	20525	1.20	x
3	oxford	WadhamCollege	5	1392	15348	1.82	x
4		ValbonneChurch	15	537	5728	2.50	x
5		MertonCollegeIII	3	1808	4188	0.24	x
6		MertonCollegeI	3	2070	9262	0.51	x
7	imajing	M0603004	14	388	24865	3.36	x
8		F0914014	15	1566	18502	0.86	x
9		C0525004	8	1570	11355	4.12	x

TABLE 4.2 – Ensemble des données de scènes urbaines utilisées : pour chaque scène, il est précisé le nombre d'images utilisées lors de la reconstruction, le nombre de points obtenus pour la reconstruction éparse (obtenu parVisualSfM) et 10% du résultat dense (obtenu par CMPMVS). Le terme Err correspond à l'erreur de reprojection dans le cas MEE, où nous disposons à la fois de la position des points d'intérêt utilisés pour l'estimation du nuage de points 3D et de la position de leurs images reprojetées, enfin la disponibilité de la vérité terrain (VT) est notée (x).

Les données 3D peuvent être éparses ou denses, cf. figure 4.9, le premier bloc :

- Les données dites **éparses** sont obtenues à partir de l'approche VisualSfM [Wu 11a]. Les correspondances de points 2D sont les points d'intérêt appariés et les points 3D correspondent au résultat de l'optimisation de la triangulation des données 2D et de l'estimation des matrices de projection. Ce qui signifie que nous sommes en mesure d'évaluer quantitativement l'erreur de reprojection (distance entre le point 2D et la reprojection de son point 3D associé) et ainsi d'évaluer la qualité des données fournies.
- Les données dites **denses** sont obtenues par densification du modèle tridimensionnel épars avec l'approche CMPMVS [Jancosek 11]. CMPMVS propose deux nuages denses, le premier complet et un second ne contenant que 10% des points. Celui à 10% contient déjà beaucoup de points (par rapport au nuage éparse) représentatifs de la scène à analyser. De plus, l'approche RANSAC sur le nuage de points complet ne donne pas de meilleurs résultats et le temps de calcul est important. En conséquence, par la suite nous travaillons sur le nuage contenant 10% des points. Enfin, les données fournies en sortie de ce programme ne permettent pas d'estimer l'erreur de reprojection puisqu'il n'y a aucune information sur

les points 2D appariés et nous ne sommes pas en mesure d'évaluer la qualité des données fournies.

Comme nous disposons, de deux types de nuages de points : un épars (E) et un dense (D) (pour lequel, finalement, seuls 10% des points sont considérés), nous pouvons envisager plusieurs configurations possibles pour l'estimation des plans puis le calcul des homographies, et ainsi, nous avons donc testé trois approches notées respectivement : MEE, MDD, MDE et résumées dans le tableau 4.3.

	MEE	MDD	MDE
Estimation des plans (points 3D)	épars	dense	dense
Calcul des homographies (points 2D)	épars	dense	épars

TABLE 4.3 – Présentation des trois méthodes testées en fonction du type de points utilisés dans les étapes d'estimation des plans et de calcul des homographies.

Nous proposons d'analyser trois aspects :

- l'influence du type de données d'entrée (éparse ou dense) sur les résultats intermédiaires ;
- l'influence de l'échelle dans le calcul de la similarité inter-images sur les résultats intermédiaires ;
- la comparaison et l'analyse de la segmentation sémantique en zones planes finale.

Enfin, nous présenterons quelques cas représentatifs.

4.6.2 In uence des données d'entrée

Comme espéré, l'estimation des plans est plus robuste dans le cas d'un nuage dense car le nombre de points décrivant les plans est plus important. En revanche, de manière générale, nous avons pu remarquer que l'approche **MDD** fournit de moins bons résultats que **MEE** ou encore **MDE**. Cela est probablement dû au fait que le processus de densification de CMPMVS introduit une erreur trop importante pour être traitée par l'approche proposée.

En utilisant, l'approche MDE nous maximisons la robustesse de l'estimation des plans en utilisant le nuage dense, et nous optimisons le recalage en prenant en compte les appariements fiables fournis par l'approche éparse.

4.6.3 In uence de l'échelle sur la similarité inter-images

Pour une optimisation de la détection des vrais positifs lors de la classification, il est nécessaire d'intégrer l'information sur une échelle appropriée. Pour avoir une meilleure appréciation de l'influence du paramètre d'échelle, des résultats visuels sont présentés sur l'image de synthèse, dans la figure 4.10. En particulier, nous détaillons les résultats pour deux groupes de trois points. Le premier groupe de points (cyan, bleu, magenta) se situe entre deux plans estimés (les $2^{\text{ème}}$ et $3^{\text{ème}}$ lignes correspondent à la carte de similarité obtenue en considérant respectivement chacun de ces plans), alors que le deuxième groupe (rouge, jaune, vert) est sur le contour représenté

par le plan du sol occulté par un plan estimé du cube (la 1$^{\text{ère}}$ ligne correspond à la carte de similarité obtenue en considérant le plan du sol).

Ainsi, sur la 1$^{\text{ère}}$ ligne, le premier groupe de points obtient une faible valeur de similarité quelle que soit l'échelle de la fenêtre de corrélation utilisée. En revanche, le deuxième groupe de points obtient une similarité très différente à faible échelle alors que tous les points du groupe atteignent une même valeur moyenne à grande échelle. Nous pouvons noter que seul le comportement du point en rouge permet de significativement affirmer qu'il appartient à ce plan, puisque c'est lui qui présente la variation la plus significative, avec une grande similarité à petites échelles qui diminue à grande échelle. On peut mettre en parallèle le comportement des valeurs obtenues sur la ligne 4, où le point rouge présente une similarité faible et variable à petites échelles et de plus en plus élevée à grande échelle.

Les lignes 2 et 3 présentent une forte similarité pour les points du premier groupe quelle que soit l'échelle alors que les valeurs de similarité pour le deuxième groupe sont variables mais faibles à petite échelle puis convergent vers une faible valeur de similarité à grande échelle.

Enfin, la 4$^{\text{ème}}$ ligne représente pour chaque échelle la valeur maximale retenue après le calcul de la similarité en considérant chacun des plans (ici 4 : le sol et les trois faces visibles). Nous pouvons remarquer la faible variation obtenue pour chaque courbe. Cependant, lorsque l'échelle augmente, nous pouvons noter que la valeur des trois points positionnés sur le contour d'occultation (deuxième groupe) diminue.

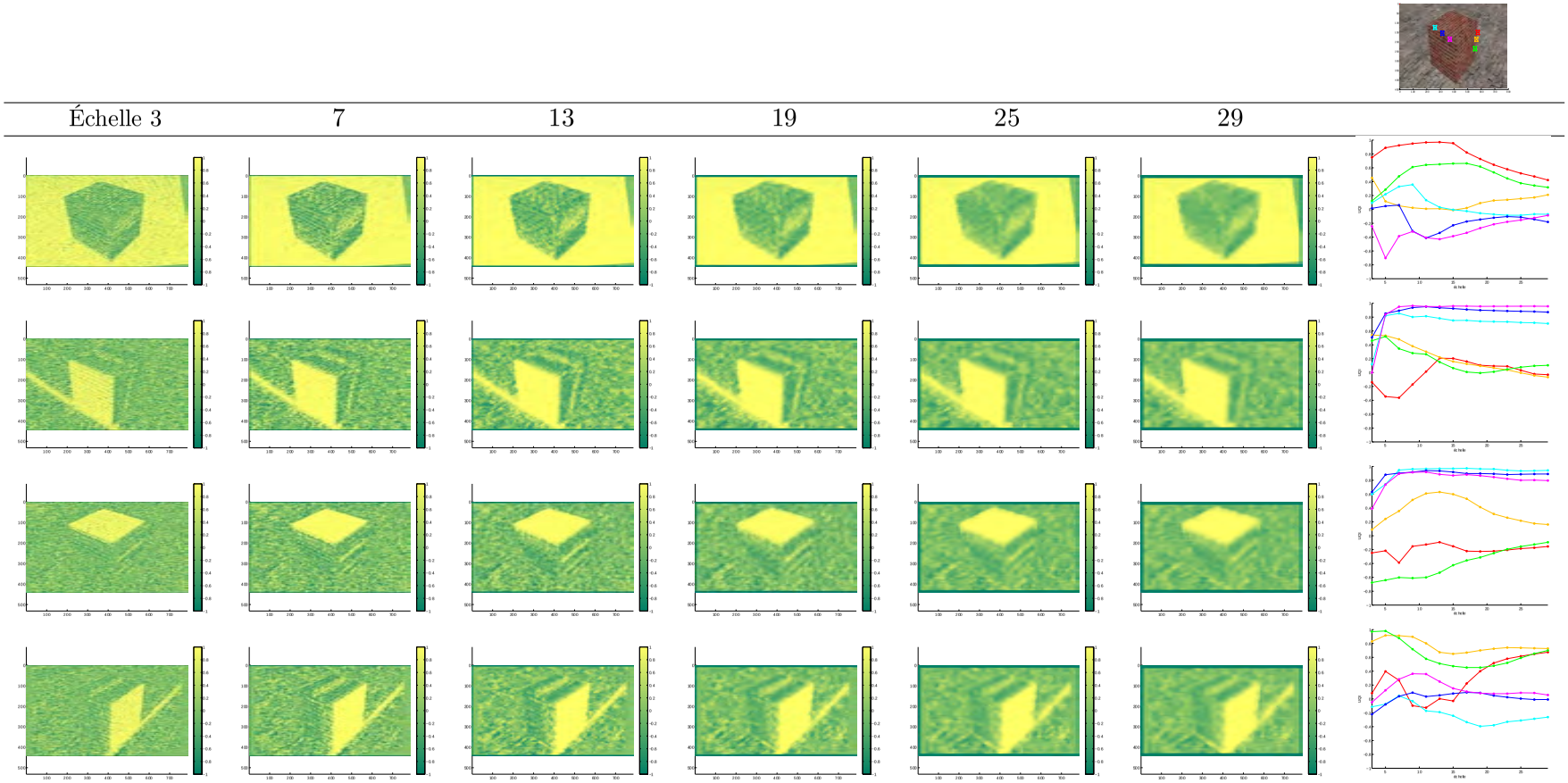


FIGURE 4.10 – Analyse de l'influence de la résolution sur la similarité inter-images : sur l'image de synthèse (boîte synthétique), pour chaque taille de fenêtre de corrélation $[3, 29]$, les trois premières lignes représentent les cartes des valeurs de similarité d'UQI obtenues et la dernière ligne le maximum obtenu sur chaque plan. Les courbes, à droite, représentent les valeurs de similarité obtenues (axe des ordonnées) en chaque point pour différentes échelles (axe des abscisses).

4.6.4 Évaluation de la segmentation sémantique

Nous proposons ici une visualisation de l'évaluation comparative de la segmentation sémantique en zones planes. Nous avons analysé nos résultats sur deux niveaux de détails. Le premier niveau de détails considère l'ensemble des bâtiments. Le second niveau, qui est plus fin, correspond à l'analyse de chaque façade séparément. Ne travaillant que sur les zones planes dominantes, telles que les façades, nous remarquons une amélioration significative lorsque l'on s'intéresse à l'analyse quantitative considérant ce niveau de segmentation. Nous détaillons quelques cas particuliers dans les figures 4.12 et 4.13.

Pour chaque image, pour chaque niveau de détails, nous comparons trois approches d'étiquetage :

1. Niveau pixel qui correspond à la carte de planéité obtenue comme résultat intermédiaire dans GEOM, que nous notons PIXEL ;
2. Niveau du super-pixel avec SLICO ;
3. Niveau du super-pixel avec GEOM.

C'est pourquoi toutes les illustrations de cette section représentent : la vérité terrain, la classification au niveau pixel, la classification utilisant GEOM et la classification SLICO. Chaque matrice de confusion permet de comparer pour un niveau de détail donné, la vérité terrain avec le résultat obtenu pour une des trois approches de classification.

Comme nous avons pu le voir lors de l'évaluation avec l'erreur de sous-segmentation et avec le rappel, notre approche améliore les résultats en particulier dans le cas d'un faible nombre de super-pixels. C'est pourquoi dans la suite notre analyse porte en particulier sur l'approche avec 50 ou 100 super-pixels.

Images réelles et images de synthèse d'une boîte posée sur un support. L'exemple de la boîte de synthèse, correspondant à l'illustration 4.11, présente des résultats très similaires au cas de la boîte réelle. Ces deux scènes correspondent à un cas très simplifié d'un bâtiment où l'ensemble des plans visibles sont largement représentés par le nuage de points 3D. Les surfaces sont texturées et nous disposons de bonnes mises en correspondance faiblement perturbées. Le calcul de l'erreur de reprojection est présentée dans le tableau 4.2 avec les données utilisées. L'approche PIXEL donne des résultats très proche de l'approche GEOM qui présente des performances bien supérieures, en termes de matrices de confusion, à l'approche SLICO et cela quel que soit le niveau de détail utilisé. De plus, nous pouvons remarquer que l'approche PIXEL ne permet pas classer les zones occultées alors que GEOM va associer ces zones à des plans photométriquement similaires.

Images réelles de scènes urbaines. Avec l'exemple de oxford Merton College III, figure 4.12, nous utilisons 50 super-pixels, $\epsilon = 0.4$ et $\lambda = 0.4$. L'évaluation au niveau de l'ensemble des bâtiments permet de constater l'apport de l'utilisation d'entités compactes comme

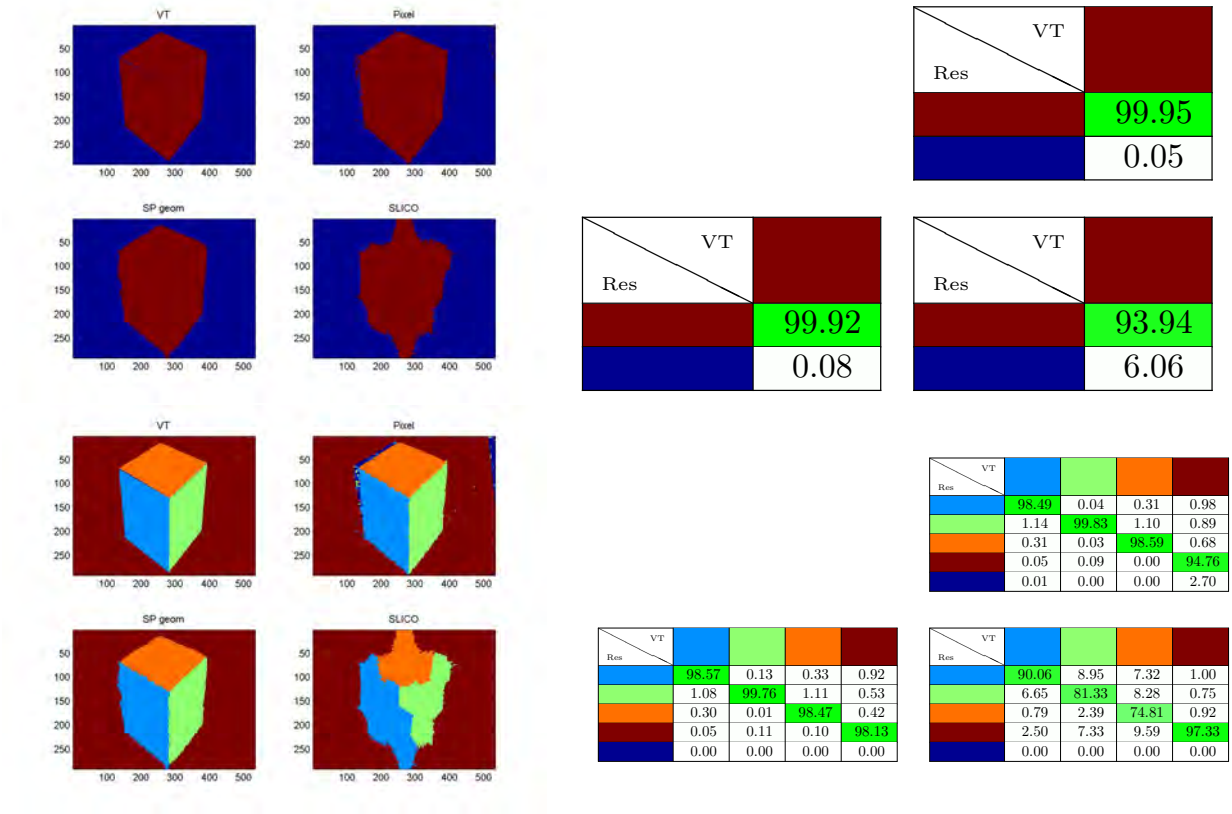


FIGURE 4.11 – Comparaison de la segmentation sémantique en plans sur des données de synthèse avec 3 approches : PIXEL, GEOM, SLICO. Visualisation des cartes de résultats ainsi que des matrices de confusion associées aux classes considérées (par bâtiments ou par surfaces ayant la même orientations), cf. section 4.6. Ici, nous utilisons 50 super-pixels, $\epsilon = 0.4$ et $\lambda = 0.4$.

les super-pixels. Les analyses à un niveau de détail plus fins, montre que les valeurs obtenues par l'approche utilisant des super-pixels géométriques sont plus élevées pour le taux de vrais positifs et présentes des erreurs plus faibles. Nous pouvons noter la difficulté a obtenir une annotation correct sur les plans correspond au toit car les points 3D associés sont généralement mal et peu représentés. par exemple pour le bâtiment représenté en bleu sur la vérité terrain et composé de deux façades verticales et d'un toit, notre approche améliore de 1 à 15% pour chacune des zones par rapport aux deux autres approches.

De manière similaire à l'exemple de oxford Merton College III, et avec les même paramètres que précédemment, les résultats obtenus dans la figure 4.13 sont encourageants. Cependant nous pouvons remarquer que notre vérité terrain n'est parfois pas assez fine. Lors de la classification en plans, les fenêtres étant un peu en retrait de la façade ont parfois été détectées. Ces entités ont alors été classifiés de la même étiquette. Nous pouvons noter sur cette exemple la difficulté à classé des plans ayant la même orientation mais sur des supports physiques disjoints.

Ces exemples ne sont qu'un échantillon de l'ensemble des scènes possibles, mais sont représentatifs de leurs variabilités. L'analyse de l'influence de la densité, de l'exactitude, de la résolution où des échelles de traitement ou du niveau de détails, nous ont permis de démontrer l'uti-

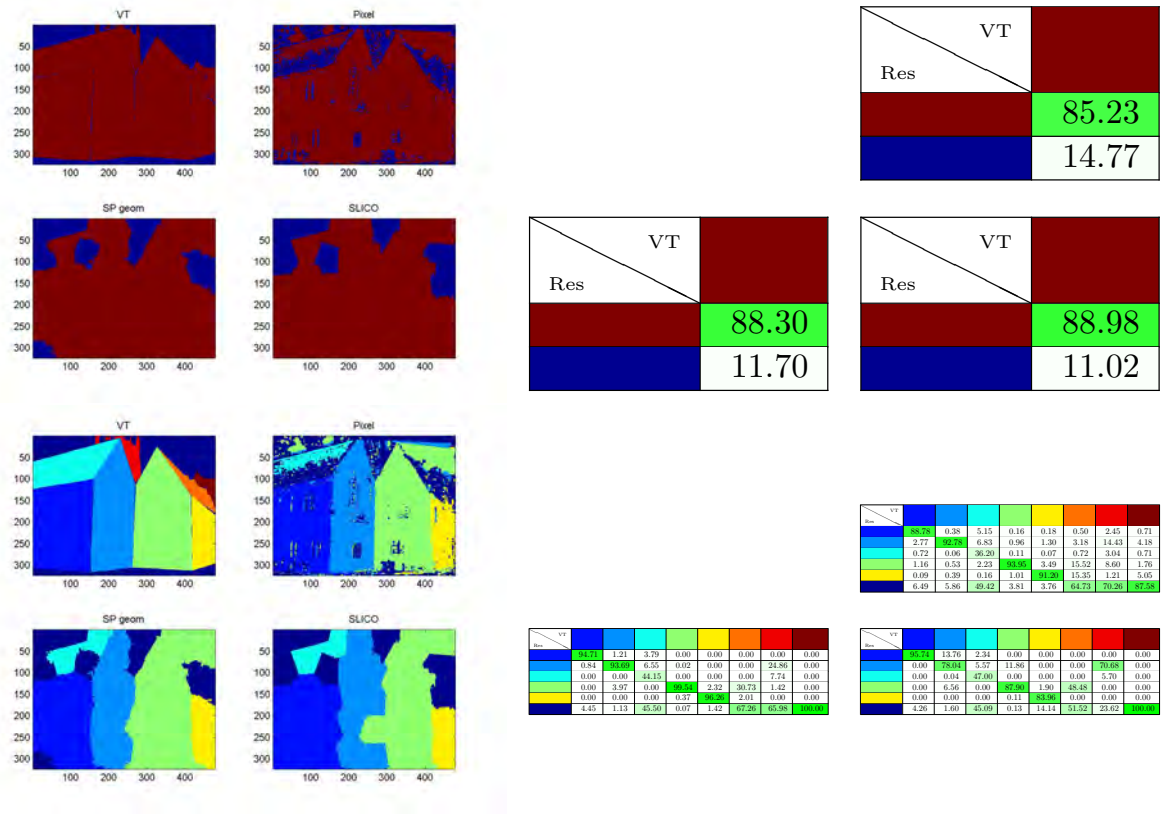


FIGURE 4.12 – Évaluation comparative des trois approches (PIXEL, GEOM, SLICO) lors de l'étiquetage pour la segmentation sémantique en plans sur les données d'Oxford Merton College III. Visualisation du résultats et matrice de confusion associée aux classes considérées (bâtiments ou planes zones des bâtiments). L'analyse de ces résultats est faite dans le texte.

lité t'intégrer de l'information géométrique dans la construction de primitive intermédiaire : les super-pixels géométrique afin d'obtenir une meilleur segmentation sémantique en plans. En effet, l'approche de super-pixels intégrant de l'information géométrique, permet de gérer des entités cohérentes (compacité, connexité) et de densifier l'étiquetage, tout en respectant la géométrie de la scène, en particulier lors du changement d'orientation des surfaces texturées. Les surfaces trop petites dans l'image, non ou mal représentées dans le nuage de point ne peuvent pas être traitées.

Afin d'avoir une vision globale de l'apport de notre approche GEOM sur la segmentation sémantique finale et de la comparer avec l'approche PIXEL et l'approche SLICO, nous proposons de quantifier le taux moyen de vrais positifs et le taux moyen de faux obtenus. Cette évaluation est représentée dans le tableau 4.4 où les taux moyens sont calculés sur les 9 images utilisées (dont 7 de scènes urbaines) avec une vérité terrain associée. Nous pouvons noter que GEOM améliore (de 6% par rapport à PIXEL et de 3% par rapport à SLIC) la détection de vrais positifs tout en diminuant la quantité de faux positifs.

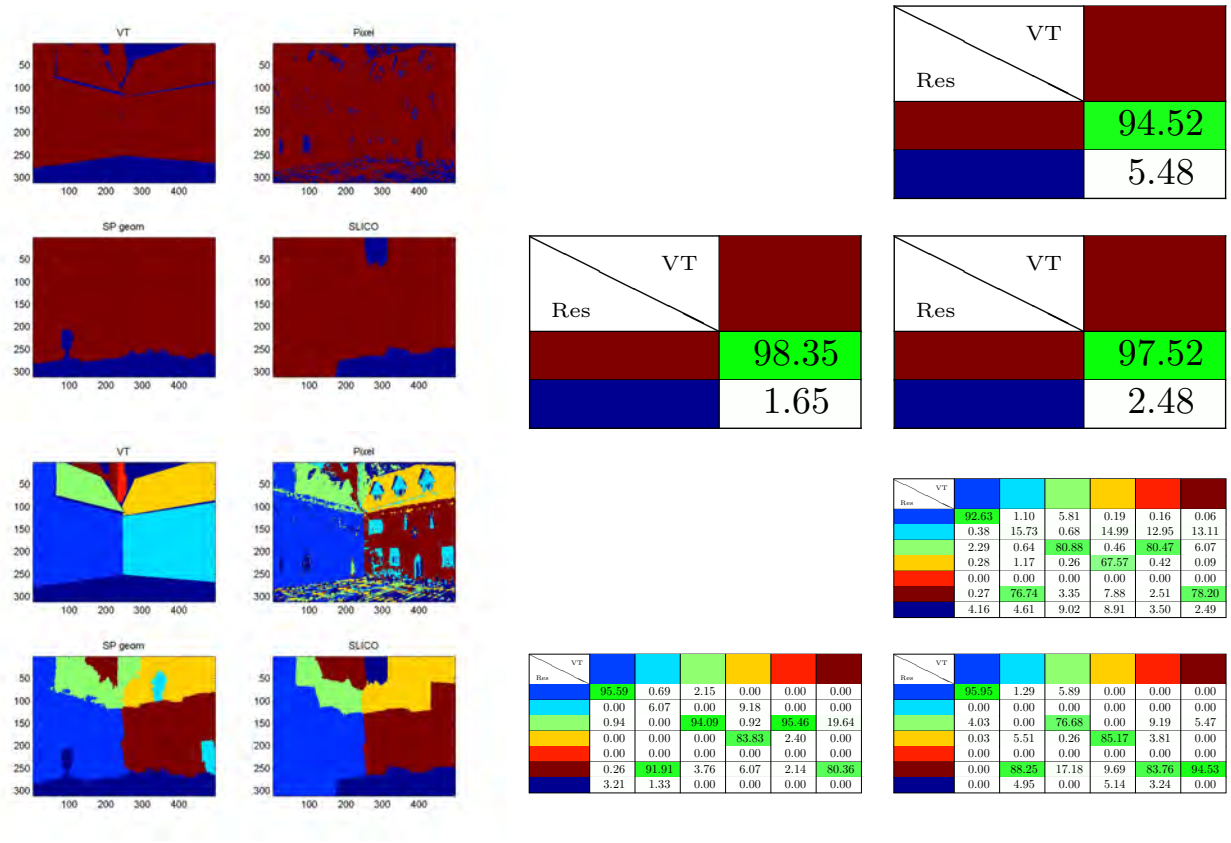


FIGURE 4.13 – Visualisation de l'étiquetage sur les données d'Oxford Merton College I pour la segmentation en plans avec les trois approches (PIXEL, GEOM, SLICO) suivant deux niveaux d'évaluation, chacune accompagnée de la matrice de confusion associée. L'analyse de ces résultats est faite dans le texte.

4.6.5 Synthèse sur l'algorithme de segmentation sémantique en plans

Dans cette partie, nous avons proposé une approche de segmentation sémantique en plans. Cette approche utilise les super-pixels géométriques présentés précédemment, qui intègre de l'information géométrique provenant d'un système de calcul de structure et de déplacement, et de l'information photométrique provenant des images. Les expérimentations ont montré l'intérêt d'intégrer de l'information géométrique dès la construction des entités de niveau intermédiaire afin de rendre plus robuste l'approche de sur-segmentation au changement surface texturé avec faible variation de couleur comme l'arrête entre deux façades.

4.7 Conclusion

Dans ce chapitre nous avons proposé une nouvelle approche de segmentation en plans utilisant notre méthode de constructeur de super-pixels qui intègre de l'information géométrique fournit par le multi-vues. Nous avons montré, dans un premier temps l'intérêt, d'intégrer de l'information géométrique telle que la notion de surface plane et de similarité inter-images dans une extension de l'approche de super-pixels SLIC. Ces travaux ont donnés lieu à une publication [Bauda 15a].

Approche	taux de TP	taux de F
PIXEL	81.21 [0.574 0.979]	7.40 [0.005 0.162]
SLICO	83.70 [0.573 0.969]	5.79 [0.022 0.110]
GEOM	87.28 [0.598 0.981]	4.73 [0.004 0.097]

TABLE 4.4 – Évaluation globale de la segmentation sémantique en plans : sur les 9 images pour les trois approches (PIXEL, SLICO, GEOM). Les colonnes correspondent au type d'approche utilisée, à la moyenne des vrais positifs et à la valeur moyenne de faux, chacune accompagnée des valeurs minimale et maximale ($[\min, \max]$), pour l'ensemble des images.

Nous avons évalué de manière quantitative l'apport de cette contribution à la fois au niveau global et au niveau local, en particulier à l'intersection des surfaces texturées dont l'orientation tridimensionnelle est connue. Ceci nous a permis de constater une amélioration de 10 à 16% du rappel lors de la quantification globale et d'une diminution de la distance aux contours. Dans un second temps nous avons proposé un algorithme permettant d'obtenir une segmentation sémantique des zones planes, qui a été présenté sous une forme initiale dans [Bauda 13]. Cette approche de segmentation utilise les super-pixels géométriques proposés précédemment et est totalement automatique et non-supervisée.

Conclusions et perspectives

Au cours de cette thèse nous avons cherché à obtenir une segmentation d'images de scènes urbaines acquises en séquence en intégrant différentes sources d'information (2D/3D) en vue de l'obtention d'une description sémantique. Les contributions ont été guidées par les concepts suivants : l'utilisation de l'information provenant d'une reconstruction tridimensionnelle disponible afin de renforcer la sur-segmentation en super-pixels. En effet, l'obtention d'une classification correcte nécessite une segmentation cohérente à la scène.

Dans un premier temps, nous avons proposé un protocole d'évaluation des mesures de cohérence photométrique pour la classification des pixels appartenant à une zone plane ou non-plane. Ce critère inter-images quantifie la similarité (ou la non-similarité) entre une image de référence et l'image adjacente recalée via l'homographie induite par le plan de support. Les expérimentations ont mis en avant la mesure UQI [Wang 02], qui présente les meilleures performances pour ce type de classification (P/NP). Cette approche a été publiée dans un article [Bauda 15b].

Ensuite, nous avons proposé une approche de sur-segmentation d'images cohérente avec la géométrie de la scène. Elle combine l'approche des super-pixels SLICO [Achanta 12] et l'intégration d'informations géométriques (orientation des plans dominants de la scène) renforcée par un critère de cohérence photométrique inter-images. Cette méthodologie a donnée lieu a deux articles préliminaires à l'approche finale [Bauda 13, Bauda 15a].

Enfin, l'ensemble de ces contributions a été mis au profit d'un algorithme générique qui propose une segmentation finale proche d'une segmentation sémantique en plans. Les résultats de la segmentation finale ont été obtenus avec une approche dite « descriptive » où nous avons cherché à extraire l'information de la géométrie de la scène afin de renforcer la sur-segmentation d'image et d'obtenir des entités (super-pixels ou groupe de super-pixels) cohérentes avec la photométrie (couleur) et la géométrie (orientation et planéité).

Il faut noter, que le rendu du modèle tridimensionnel simplifié est plus réaliste lors de l'estimation des plans dominants qu'avec l'interpolation sur chaque facette (sortie de CMPMVS).

À l'issu de cette thèse, nous pensons cependant que la solution proposée peut être améliorée. La reconnaissance de la route et l'interprétation des marquages pour la reconnaissance de route ou de carrefour sont également des domaines très actifs de la recherche qui peuvent enrichir le système de compréhension de scènes urbaines. L'extraction et l'appariement de droites dans la scène aident à la compréhension des orientation des surfaces (route et objets verticaux). Nous pouvons également envisager une extension de notre approche prenant en compte ces outils.

Il est possible de combiner, par exemple, les deux étapes d'estimation de la planéité et de calcul de similarité avec une approche telle que *sift ow* [Liu 08], qui définit un descripteur pour chaque pixel et le met en correspondance avec un pixel de la seconde image en utilisant la notion

de planéité comme contrainte. Ce processus fournit une mise en correspondance dense des pixels de l'image qui peut être interprétée comme une information de profondeur.

Tout un ensemble d'approches existantes utilisent des méthodes par apprentissage dit profond (ou *deep learning* en anglais) pour l'obtention d'une segmentation sémantique. La prochaine étape de ce travail pourrait consister à intégrer les super-pixels cohérents avec la géométrie dans des approches d'apprentissage. En ce qui concerne les objets particuliers comme des piétons, des voitures ou du textes, de nombreux travaux sont également disponibles.

Ransac : *RANdom SAmple Consensus*

Ransac corresponpond à l'abbréviation *RANdom SAmple Consensus* [Fischler 81]. Il s'agit d'un processus itératif permettant d'estimer les paramètres d'un modèle mathématique à partir d'un ensemble de données observé qui contient des valeurs aberrantes, appelées *outliers*. Il est utilisé par exemple pour estimer une droite dans un ensemble de points 2D. Cette algorithme est non-déterministe dans le sens où il produit un résultat correct avec une certaine probabilité qui augmente à mesure que le nombre d'itérations est grand. Cette approche a été revisité sur un grand ensemble de données variées (2D, 3D, ND), multi-hypothèses avec le j-linkage [Toldo 08]. Nous y faisons référence dans l'approche proposée lors de l'estimation des plans dans un nuage de points 3D.

Algorithme 4 : Algorithme général de Ransac.

Données :

- données - un ensemble d'observation,
- modèle - un modèle à ajuster aux données,
- n - le nombre minimum de données nécessaires pour ajuster le modèle,
- k - le nombre maximal d'itération (ou condition d'arrêt de convergence) de l'algorithme,
- t - une valeur seuil pour déterminer la représentativité d'une donnée au modèle,
- d - le nombre de données proches des valeurs nécessaire pour faire valoir une correspondance du modèle aux données.

Résultat :

- meilleurModèle - les paramètres du modèle qui correspondent le mieux aux données (ou zéro si aucun bon modèle a été trouvé),
- meilleurEnsemblePoints - données à partir desquelles ce modèle a été estimé,
- meilleureErreur - l'erreur de ce modèle par rapport aux données.

// Initialisation des paramètres
itérateur ← 0
meilleurModèle ← *aucun*
meilleurEnsemblePoints ← *aucun*
meilleureErreur ← $infini$
// Recherche du meilleur cas
tant que *itérateur* $< k$ **faire**
  incrémention de l'itérateur
  pointsAléatoires ← valeurs choisies au hasard à partir des données
  modèlePossible ← paramètres du modèle correspondant aux pointsAléatoires
  ensemblePoints ← pointsAléatoires
  pour chaque *point des données pas dans pointsAléatoires* **faire**
    si *le point s ajuste au modèlePossible avec une erreur inférieure à t* **alors**
      Ajouter un point à ensemblePoints
    si *le nombre d éléments dans* ensemblePoints *est* $> d$ *(ce qui implique que nous avons peut-être trouvé un bon modèle, on teste maintenant dans quelle mesure il est correct)* **alors**
      modèlePossible ← paramètres du modèle réajusté à tous les points de ensemblePoints
      erreur ← une mesure de la manière dont ces points correspondent au modèlePossible
    si *erreur* $<$ meilleureErreur **alors**
      (nous avons trouvé un modèle qui est mieux que tous les précédents, le garder jusqu'à ce qu'un meilleur soit trouvé)
      meilleurModèle ← modèlePossible
      meilleurEnsemblePoints ← ensemblePoints
      meilleureErreur ← erreur
// Résultat final
retourne meilleurModèle, meilleurEnsemblePoints, meilleureErreur

Bibliographie

[Achanta 10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua & S. Susstrunk. SLIC superpixels. Rapport de recherche, École Polytechnique Fédérale de Lausanne (EPFL), 2010.

[Achanta 12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua & S. Susstrunk. *SLIC Superpixels Compared to State-of-the-art Superpixel Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2012.

[Alexe 10] B. Alexe, T. Deselaers & V. Ferrari. *What is an object ?* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2010.

[Alvarez 12] J. Alvarez, T. Gevers, Y. LeCun & A. López. *Road Scene Segmentation from a Single Image.* Dans European Conference on Computer Vision (ECCV), 2012.

[Arbeláez 09] P. Arbeláez, M. Maire, C.C. Fowlkes & J. Malik. *From Contours to Regions : An Empirical Evaluation.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

[Arbeláez 11] P. Arbeláez, M. Maire, C. Fowlkes & J. Malik. *Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2011.

[Aschwanden 92] P. Aschwanden & W. Guggenbül. *Experimental results from a comparative study on correlation type registration algorithms.* Dans W. Förstner & S. Ruwiedel, rédacteurs, Robust computer vision : Quality of Vision Algorithms. 1992.

[Badrinarayanan 15] V. Badrinarayanan, A. Handa & R. Cipolla. *SegNet : A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling. arXiv preprint*, 2015.

[Bagon 08] S. Bagon, O. Boiman & M. Irani. *What Is a Good Image Segment ? A Uni ed Approach to Segment Extraction.* Dans European Conference on Computer Vision (ECCV), 2008.

[Bao 14] S. Y. Bao, A. Furlan, L. Fei-Fei & S. Savarese. *Understanding the 3D layout of a cluttered room from multiple images.* Dans Winter Conference on Applications of Computer Vision (WACV), 2014.

[Bartoli 07] A. Bartoli. *A random sampling strategy for piecewise planar scene segmentation.* Dans Computer Vision and Image Understanding (CVIU), 2007.

[Bauda 13] M.-A. Bauda, S. Chambon, M. Spangenberg & V. Charvillat. *Segmentation de scènes urbaines par combinaison d information.* Dans ORASIS, journées francophones des jeunes chercheurs en vision par ordinateur, 2013.

[Bauda 15a] M.-A. Bauda, S. Chambon, P. Gurdgos & V. Charvillat. *Geometry-Based Superpixel Segmentation Introduction of Planar Hypothesis for Superpixel Construction.* Dans International Conference on Computer Vision Theory and Applications (VISAPP), 2015.

[Bauda 15b] M.-A. Bauda, S. Chambon, P. Gurdgos & V. Charvillat. *Image Quality Assessment for Photo-consistency Evaluation on Planar Classi cation in Urban Scenes.* Dans International Conference on Pattern Recognition Applications and Methods (ICPRAM), 2015.

[Bazin 12] J.-C. Bazin & M. Pollefeys. *3-line RANSAC for orthogonal vanishing point detection.* Dans IROS, 2012.

[Birchfield 98] S. Birchfield & C. Tomasi. *A Pixel Dissimilarity Measure That Is Insensitive to Image Sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1998.

[Bleyer 11] M. Bleyer, C. Rother, P. Kohli, D. Scharstein & S. Sinha. *Object Stereo - Joint Stereo Matching and Object Segmentation.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2011.

[Bódis-Szomorú 14] A. Bódis-Szomorú, H. Riemenschneider & L. Van Gool. *Fast, Approximate Piecewise-Planar Modeling Based on Sparse Structure-from-Motion and Superpixels.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2014.

[Bódis-Szomorú 15] A. Bódis-Szomorú, H. Riemenschneider & L. Van Gool. *Superpixel Meshes for Fast Edge-Preserving Surface Reconstruction.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 7-12 June 2015.

[Boucher 05] A. Boucher & T. Le. *Comment extraire la semantique d une image ?* Dans Conference Internationale Sciences Electroniques, Technologies de l'Information et des Telecommunications (SETIT), 2005.

[Brostow 08] G. Brostow, J. Shotton, J. Fauqueur & R. Cipolla. *Segmentation and Recognition Using Structure from Motion Point Clouds.* Dans European Conference on Computer Vision (ECCV), 2008.

[Brun 96] L. Brun. *Segmentation d images à base topologique.* Thèse de doctorat, Université de Bordeaux I, 1996.

[Canny 86] J. Canny. *A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1986.

[Chan 01] T. F. Chan & L. A. Vese. *Active Contours Without Edges. Transactions on Image Processing*, 2001.

[Chatfield 09] K. Chatfield, J. Philbin & A. Zisserman. *E cient retrieval of deformable shape classes using local self-similarities.* Dans IEEE International Conference on Computer Vision Workshop (ICCV), 2009.

[Chauve 10] A.-L. Chauve, P. Labatut & J.-P. Pons. *Robust piecewise-planar 3D reconstruction and completion from large-scale unstructured point data.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2010.

[Collins 96] R.T. Collins. *A space-sweep approach to true multi-image matching.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 1996.

[Comaniciu 02] D. Comaniciu & P. Meer. *Mean shift : a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2002.

[Coughlan 03] J.M. Coughlan & A.L. Yuille. *Manhattan world : Orientation and outlier detection by bayesian inference. Neural Computation*, 2003.

[Dai 13] J. Dai, Y.N. Wu, J. Zhou & S. Zhu. *Cosegmentation and Cosketch by Unsupervised Learning.* Dans IEEE International Conference on Computer Vision (ICCV), 2013.

[Davis 06] J. Davis & M. Goadrich. *The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning - ICML 06*, 2006.

[Delage 07] E. Delage, H. Lee & A. Ng. *Automatic Single-Image 3D Reconstructions of Indoor Manhattan World Scenes. Robotics Research*, 2007.

[Desolneux 02] A. Desolneux, L. Moisan & J.-M. Morel. *Computational Gestalts and Perception Thresholds. Journal of Physiology*, 2002.

[Dey 10] V. Dey, Y. Zhang & M. Zhong. *A review on image segmentation techniques with remote sessing perspective.* Dans International Society for Photogrammetry and Remote Sensing (ISPRS), 2010.

[Duan 15] L. Duan & F. Lafarge. *Image partitioning into convex polygons.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2015.

[Erbs 11] F. Erbs, A. Barth & U. Franke. *Moving vehicle detection by optimal segmentation of the Dynamic Stixel World. 2011 IEEE Intelligent Vehicles Symposium (IV)*, 2011.

[Fawcett 06] T. Fawcett. *An Introduction to ROC Analysis. Pattern Recognition Letters*, 2006.

[Felzenszwalb 04] P. F. Felzenszwalb & D. P. Huttenlocher. *E cient Graph-Based Image Segmentation. International Journal of Computer Vision (IJCV)*, 2004.

[Fischler 81] M. A. Fischler & R. C. Bolles. *Random Sample Consensus : A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Commun. ACM*, 1981.

[Floros 12] G. Floros & B. Leibe. *Joint 2D-3D temporally consistent semantic segmentation of street scenes.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2012.

[Fouhey 10] D. Fouhey, D. Scharstein & A. Briggs. *Multiple Plane Detection in Image Pairs Using J-Linkage.* Dans International Conference on Pattern Recognition (ICPR), 2010.

[Fouhey 13] D. Fouhey, A. Gupta & M. Hebert. *Data-Driven 3D Primitives for Single Image Understanding.* Dans IEEE International Conference on Computer Vision (ICCV), 2013.

[Fulkerson 09] B. Fulkerson, A. Vedaldi & S. Soatto. *Class segmentation and object localization with superpixel neighborhoods.* Dans IEEE International Conference on Computer Vision (ICCV), 2009.

[Furukawa 09] Y. Furukawa, B. Curless, S. M. Seitz & R. Szeliski. *Manhattan-world stereo.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

[Furukawa 10] Y. Furukawa & J. Ponce. *Accurate, Dense, and Robust Multi-View Stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2010.

[Gallup 10] D. Gallup, J.-M. Frahm & M. Pollefeys. *Piecewise planar and non-planar stereo for urban scene reconstruction.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2010.

[Geiger 12] A. Geiger, P. Lenz & R. Urtasun. *Are we ready for Autonomous Driving ? The KITTI Vision Benchmark Suite.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2012.

[Getreuer 11] P. Getreuer. *Linear Methods for Image Interpolation. Image Processing On Line*, 2011.

[Gould 08] S. Gould, J. Rodgers, D. Cohen, G. Elidan & D. Koller. *Multi-Class Segmentation with Relative Location Prior. International Journal of Computer Vision (IJCV)*, 2008.

[Grompone von Gioi 12] R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel & G. Randall. *LSD : a Line Segment Detector. Image Processing On Line*, 2012.

[Grundmann 10] M. Grundmann, V. Kwatra, M. Han & I. Essa. *E cient hierarchical graph-based video segmentation.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2010.

[Habbecke 06] M. Habbecke & L. Kobbelt. *Iterative Multi-View Plane Fitting.* Dans Vision, Modeling, and Visualization (VMV), 2006.

[Hanbury 08] A. Hanbury. *How Do Superpixels A ect Image Segmentation ?* Dans Iberoamerican Congress on Pattern Recognition (CIARP), Lecture Notes in Computer Science, 2008.

[Häne 13] C. Häne, C. Zach, A. Cohen, R. Angst & M. Pollefeys. *Joint 3D Scene Reconstruction and Class Segmentation.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2013.

[Harris 88] C. Harris & M. Stephens. *A combined corner and edge detector.* Dans In Proc. of Fourth Alvey Vision Conference, 1988.

[Hartley 04] R. I. Hartley & A. Zisserman. Multiple view geometry in computer vision. Second édition, 2004.

[Hernández 10] C. Hernández & G. Vogiatzis. *Shape from Photographs : A Multi-view Stereo Pipeline.* Dans Computer Vision : Detection, Recognition and Reconstruction. 2010.

[Hochbaum 09] D.S. Hochbaum & V. Singh. *An e cient algorithm for Co-segmentation.* Dans IEEE International Conference on Computer Vision (ICCV), 2009.

[Hoiem 05a] D. Hoiem, A.A. Efros & M. Hebert. *Automatic photo pop-up. ACM Trans. Graph.*, 2005.

[Hoiem 05b] D. Hoiem, A.A. Efros & M. Herbert. *Geometric context from a single image.* Dans IEEE International Conference on Computer Vision (ICCV), 2005.

[Hoiem 07] D. Hoiem. *Seeing the world behind the image Spatial Layout for 3D Scene Understanding.* Thèse de doctorat, 2007.

[Hoiem 08] D. Hoiem, A.A. Efros & M. Herbert. *Closing the loop on scene interpretation.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2008.

[Hoiem 11] D. Hoiem, A.A. Efros & M. Hebert. *Recovering occlusion boundaries from an image. International Journal of Computer Vision (IJCV)*, 2011.

[Hu 12] R. Hu, L. Fan & L. Liu. *Co-Segmentation of 3D Shapes via Subspace Clustering. Computer Graphics Forum (Proceedings of SGP)*, 2012.

[Jancosek 11] M. Jancosek & T. Pajdla. *Multi-view Reconstruction Preserving Weakly-supported Surfaces.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2011.

[Joulin 12] A. Joulin, F. Bach & J. Ponce. *Multi-Class Cosegmentation.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2012.

[Kass 88] M. Kass, A. Witkin & D. Terzopoulos. *Snakes : Active contour models.* Dans International Journal of Computer Vision (IJCV), 1988.

[Kim 12] G. Kim & E. Xing. *On multiple foreground cosegmentation.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2012.

[Kutulakos 00] K. Kutulakos. *Approximate N-View Stereo.* Dans European Conference on Computer Vision (ECCV), 2000.

[Labatut 09] P. Labatut, J.-P. Pons & R. Keriven. *Robust and E cient Surface Reconstruction From Range Data. Computer Graphics Forum*, 2009.

[Lafarge 12] F. Lafarge & C. Mallet. *Creating large-scale city models from 3D-point clouds : a robust approach with hybrid representation. International Journal of Computer Vision (IJCV)*, 2012.

[Levinshtein 09] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson & K. Siddiqi. *TurboPixels : Fast Superpixels Using Geometric Flows. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2009.

[Levinshtein 10] A. Levinshtein, C. Sminchisescu & S. Dickinson. *Optimal contour closure by superpixel grouping. Lecture Notes in Computer Science (including subseries Lecture Notes in Arti cial Intelligence and Lecture Notes in Bioinformatics)*, 2010.

[Levy-Schoen 68] A. Levy-Schoen. *Eye movements and vision. Neuropsychologia*, 1968.

[Lezama 14] J. Lezama, R. Grompone von Gioi, Gregory Randall & J.-M. Morel. *Finding Vanishing Points via Point Alignments in Image Primal and Dual Domains.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), June 2014.

[Li 15] Z. Li & J. Chen. *Superpixel Segmentation Using Linear Spectral Clustering.* 2015.

[Lin 15] G. Lin, C. Shen, I. Reid & A. van den Hengel. *E cient piecewise training of deep structured models for semantic segmentation. CoRR*, abs/1504.01013, 2015.

[Liu 08] C. Liu, J. Yuen, A. Torralba, J. Sivic & W.T. Freeman. *SIFT Flow : Dense Correspondence Across Di erent Scenes.* Dans European Conference on Computer Vision (ECCV), 2008.

[Liu 10] B. Liu, S. Gould & D. Koller. *Single image depth estimation from predicted semantic labels.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2010.

[Liu 11] Y. Liu, W. Zhou, H. Yin & N. Yu. *Tracking Based on SURF and Superpixel.* Dans International Conference on Image and Graphics (ICIG), 2011.

[Lowe 04] D. G. Lowe. *Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV)*, 2004.

[Martin 04] D. Martin, C. Fowlkes & J. Malik. *Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2004.

[Matsuo 13] K. Matsuo & Y. Aoki. *Depth Interpolation via Smooth Surface Segmentation Using Tangent Planes Based on the Superpixels of a Color Image.* Dans IEEE International Conference on Computer Vision Workshop (ICCV), 2013.

[Mikolajczyk 04] K. Mikolajczyk & C. Schmid. *Scale & A ne Invariant Interest Point Detectors. International Journal of Computer Vision*, 60(1) :63–86, 2004.

[Miksik 15] O. Miksik, Y. Amar, V. Vineet, P. Pérez & P.H.S. Torr. *Incremental dense multi-modal 3D scene reconstruction.* Dans International Conference on Intelligent Robots and Systems (IROS), 2015.

[Mičušík 09] B. Mi u ík & J. Ko ecká. *Semantic segmentation of street scenes by superpixel co-occurrence and 3D geometry.* Dans IEEE International Conference on Computer Vision (ICCV), 2009.

[Mičušík 10] B. Mi u ík & J. Ko ecká. *Multi-view Superpixel Stereo in Urban Environments. International Journal of Computer Vision (IJCV)*, 2010.

[Moore 08] A. P. Moore, S. Prince, J. Warrell, U. Mohammed & G. Jones. *Superpixel lattices.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2008.

[Moore 09] A.P. Moore, S. J. D. Prince, J. Warrell, U. Mohammed & G. Jones. *Scene shape priors for superpixel segmentation.* Dans IEEE International Conference on Computer Vision (ICCV), 2009.

[Mori 04] G. Mori, X. Ren, A.A. Efros & J. Malik. *Recovering human body con gurations : combining segmentation and recognition. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2004.

[Mori 05] G. Mori. *Guiding model search using segmentation.* Dans IEEE International Conference on Computer Vision (ICCV), 2005.

[Musialski 12] P. Musialski, P. Wonka, D.G. Aliaga, M. Wimmer, L. Van Gool & W. Purgathofer. *A Survey of Urban Reconstruction.* Dans EUROGRAPHICS 2012 State of the Art Reports, 2012.

[Nawaf 14] M.M. Nawaf, M.A. Hasnat, D. Sidibé & A. Trémeau. *Color and ow based superpixels for 3D geometry respecting meshing.* Dans IEEE Winter Conference on Applications of Computer Vision, 2014.

[Neubert 12] P. Neubert & P. Protzel. Superpixel benchmark and comparison. 2012.

[Palubinskas 14] G. Palubinskas. *Mystery behind similarity measures mse and SSIM.* Dans IEEE International Conference on Image Processing (ICIP), 2014.

[Rand 71] W.M. Rand. *Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association*, 1971.

[Raza 13] S. H. Raza, M. Grundmann & I. Essa. *Geoemetric Context from Video.* Dans IEEE Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2013.

[Ren 03] X. Ren & J. Malik. *Learning a classi cation model for segmentation.* Dans IEEE International Conference on Computer Vision (ICCV), 2003.

[Ren 07] X. Ren & J. Malik. *Tracking as Repeated Figure/Ground Segmentation.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2007.

[Rosenberger 06] C. Rosenberger. *Contribution à l évaluation d algorithmes de traitement d images.* Habilitation à diriger des recherches, 2006.

[Rother 06] C. Rother, T. Minka, A. Blake & V. Kolmogorov. *Cosegmentation of Image Pairs by Histogram Matching - Incorporating a Global Constraint into MRFs.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2006.

[Rubio 12] J. Rubio, J. Serrat, A. López & N. Paragios. *Unsupervised co-segmentation through region matching.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2012.

[Saxena 08] A. Saxena, M. Sun & A.Y. Ng. *Make3D : Depth Perception from a Single Still Image.* Dans Association for the Advancement of Artificial Intelligence (AAAI), 2008.

[Saxena 09] A. Saxena, M. Sun & A.Y. Ng. *Make3D : Learning 3D Scene Structure from a Single Still Image. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2009.

[Scharwächter 13] T. Scharwächter, M. Enzweiler, U. Franke & S. Roth. *E cient Multi-cue Scene Segmentation.* Dans 35th German Conference Pattern Recognition(GCPR), 2013.

[Schick 11] A. Schick & R. Stiefelhagen. *Evaluating image segments by applying the description length to sets of superpixels.* Dans IEEE International Conference on Computer Vision Workshop (ICCV), 2011.

[Schick 12] A. Schick, M. Fischer & R. Stiefelhagen. *Measuring and evaluating the compactness of superpixels.* Dans International Conference on Pattern Recognition (ICPR), 2012.

[Seitz 06] S. Seitz, B. Curless, J. Diebel, D. Scharstein & R. Szeliski. *A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2006.

[Shao 03] H. Shao, T. Svoboda & L. Van Gool. ZuBuD Zürich buildings database for image based recognition. Rapport Technique, 2003.

[Shechtman 07] E. Shechtman & M. Irani. *Matching Local Self-Similarities across Images and Videos.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2007.

[Shi 00] J. Shi & J. Malik. *Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2000.

[Shokron 10] S. Shokron & C. Marendaz. Comment voyons-nous ? Le Pommier, 2010.

[Sinha 08] S.N. Sinha, D. Steedly & R. Szeliski. *Interactive 3D architectural modeling from unordered photo collections. ACM Transactions on Graphic*, 2008.

[Sinha 09] S.N. Sinha, D. Steedly & R. Szeliski. *Piecewise planar stereo for image-based rendering.* Dans IEEE International Conference on Computer Vision (ICCV), 2009.

[Sinha 14] S.N. Sinha, D. Scharstein & R. Szeliski. *E cient High-Resolution Stereo Matching using Local Plane Sweeps.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2014.

[Sturgess 07] P. Sturgess, K. Alahari, C. Russell & P. Torr. *What , Where & How Many ? Combining Object Detectors and CRFs.* 2007.

[Stutz 15] D. Stutz. *Superpixel Segmentation : An Evaluation.* Dans International Conference on Pattern Recognition (ICPR). 2015.

[Szeliski 10] R. Szeliski. Computer vision : Algorithms and applications. Springer, 2010.

[Tighe 13] J. Tighe & S. Lazebnik. *Superparsing - Scalable Nonparametric Image Parsing with Superpixels. International Journal of Computer Vision (IJCV)*, 2013.

[Toldo 08] R. Toldo & A. Fusiello. *Robust multiple structures estimation with j-linkage.* Dans European Conference on Computer Vision (ECCV), 2008.

[Toldo 10] R. Toldo & A. Fusiello. *Photo-consistent planar patches from unstructured cloud of points.* Dans European Conference on Computer Vision (ECCV). 2010.

[Tombari 07] Federico Tombari, Stefano Mattoccia & Luigi Di Stefano. *Segmentation-based Adaptive Support for Accurate Stereo Correspondence.* Dans Proceedings of the 2Nd Pacific Rim Conference on Advances in Image and Video Technology, 2007.

[Torralba 02] A. Torralba & A. Oliva. *Depth estimation from image structure. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2002.

[tos 07] *Image Matching via Saliency Region Correspondences*, 2007.

[Toshev 10] A. Toshev, B. Taskar & K. Daniilidis. *Object detection via boundary structure segmentation. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2010.

[Tretyak 12] E. Tretyak, O. Barinova, P. Kohli & V. Lempitsky. *Geometric image parsing in man-made environments. International Journal of Computer Vision*, 2012.

[Unnikrishnan 05] R. Unnikrishnan, C. Pantofaru & M. Hebert. *A Measure for Objective Evaluation of Image Segmentation Algorithms.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2005.

[Van den Bergh 12] M. Van den Bergh, X. Boix, G. Roig, B.de Capitani & L. J. Van Gool. *SEEDS : Superpixels Extracted via Energy-Driven Sampling.* Dans European Conference on Computer Vision (ECCV), 2012.

[Van Gool 13] L. Van Gool, A. Martinovic & M. Mathias. *Towards Semantic City Models.* 2013.

[Vantaram 12] S.R. Vantaram & E. Saber. *Survey of contemporary trends in color image segmentation. Journal of Electronic Imaging*, 2012.

[Vazquez-Reina 10] A. Vazquez-Reina, S. Avidan, H. Pfister & E. Miller. *Multiple Hypothesis Video Segmentation from Superpixel Flows.* Dans European Conference on Computer Vision (ECCV), 2010.

[Vedaldi 08] A. Vedaldi & S. Soatto. *Quick Shift and Kernel Methods for Mode Seeking.* Dans European Conference on Computer Vision (ECCV), 2008.

[Veksler 10] O. Veksler, Y. Boykov & P. Mehrani. *Superpixels and supervoxels in an energy optimization framework.* Dans European Conference on Computer Vision (ECCV), 2010.

[Vicente 08] S. Vicente, V. Kolmogorov & C. Rother. *Graph cut based image segmentation with connectivity priors. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2008.

[Vincent 91] L. Vincent & P. Soille. *Watersheds in Digital Spaces : An E cient Algorithm Based on Immersion Simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1991.

[Vogel 13] C. Vogel, K. Schindler & S. Roth. *Piecewise Rigid Scene Flow.* Dans IEEE International Conference on Computer Vision (ICCV), 2013.

[Wang 02] Z Wang & AC Bovik. *A universal image quality index. IEEE Signal Processing Letters*, 2002.

[Wang 04] Z. Wang, A.C. Bovik, H.R. Sheikh & E.P. Simoncelli. *Image Quality Assessment : From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing*, 2004.

[Wang 09] Z. Wang & A.C. Bovik. Error : Love it or leave it ? 2009.

[Wang 12] J. Wang & X. Wang. *VCells : Simple and E cient Superpixels Using Edge-Weighted Centroidal Voronoi Tessellations. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2012.

[Wang 14] R. Wang, J. Choi & G. Medioni. *3D Modeling from Wide Baseline Range Scans using Contour Coherence.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2014.

[Witt 13] J. Witt & U. Weltin. *Robust Stereo Visual Odometry Using Iterative Closest Multiple Lines.* Dans Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE/RSJ, 2013.

[Wu 11a] C. Wu. Visualsfm : A visual structure from motion system. 2011.

[Wu 11b] C. Wu, S. Agarwal, B. Curless & S.M. Seitz. *Multicore bundle adjustment.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2011.

[Xiong 15] B. Xiong, M. Jancosek, S. Oude Elberink & G. Vosselman. *Flexible building primitives for 3D building modeling. International Society for Photogrammetry and Remote Sensing (ISPRS)*, 2015.

[Xu 14] J. Xu, A.G. Schwing & R. Urtasun. *Tell Me What You See and I Will Show You Where It Is.* Dans IEEE Computer Vision and Pattern Recognition (CVPR), 2014.

[Zagrouba 94] E. Zagrouba. *Construction de facettes tridimensionnelles par mise en correspondance de régions en stéréovision.* Thèse de doctorat, Toulouse, INPT, 1994.

[Zeng 11] G. Zeng, P. Wang, J. Wang, R. Gan & H. Zha. *Structure-sensitive superpixels via geodesic distance.* Dans IEEE International Conference on Computer Vision (ICCV), 2011.

[Zhang 10] H. Zhang, J. Xiao & L. Quan. *Supervised label transfer for semantic segmentation of street scenes.* Dans European Conference on Computer Vision (ECCV), 2010.

[Zhang 13] X. Zhang, Y.H. Yang, Z. Han, H. Wang & C. Gao. *Object Class Detection : A Survey. ACM Computer Survey*, 2013.

[Zhao 11] L.G. Zhao & C.K. Wu. *A Planes Detection Algorithm Based on Feature Distribution.* Dans IEEE International Conference on Intelligent Networking and Collaborative Systems (INCoS), 2011.

Glossaire

CMVS *Clustering Views for Multi-view Stereo*. 27

CRF *Conditional Random Field*. 84, 85

GAT *Graphic Annotation Tool*. 48

GNSS *Global Navigation Satellite System*. 31

GPS *Global Position System*. 31

IMU *Inertial Measurement Unit*. 31

IQA *Image Quality Assessment*. 46, 47, 62, 72, 77, 78, 90

LiDAR *Light Detecting And Ranging scans*. 8

LSD *Line Segment Detector*. 9, 41

MRF *Markov Random Field*. 15, 17, 19, 85

MSE *Mean Square Error*. 62, 63, 66, 70, 75

MSE$_r$ *Mean Square Error* sur un voisinage r. 63, 66, 77

MVS *Multi View Stereo*. 12, 14, 21

OLT *Object Labeling Tool*. 48, 49

PMVS *Patch-based Multi-view Stereo Software*. 27

PR Précision-Rappel. 52, 55, 76

RC$_r$ r-coherence. 63, 65, 66, 70, 77

ROC *Receiver Operator Characteristic*. 52, 55, 56, 76

RUQI *Universal Quality Image* sur un voisinage r. 65, 66, 67, 68, 70, 77

SfM *Structure from Motion*. 12, 14, 15, 16, 51, 72, 90

SLIC *Simple Linear Iterative Clustering*. 86

SSIM *Structure SImilarity Measure*. 62, 64, 65, 66, 68, 70, 75, 77, 90

SVM *Support Vector Machine*. 16

UQI *Universal Quality Image*. xvi, 65, 66, 68, 70, 74, 76, 77, 78, 90, 103

YACVID *Yet Another Computer Vision Index To Datasets*. 49

ZNCC *Zero mean Normalised Cross-Correlation*. 65