UNIVERSITAT POLITÈCNICA
DE CATALUNYA

Ph.D. Thesis Dissertation

Monocular Depth Estimation for Image Segmentation and Filtering

Author: Mariella Dimiccoli

Advisor: Prof. Philippe Salembier

Image Processing Group
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya

Barcelona, July 2009

Ai miei genitori

Abstract

This Ph.D. dissertation addresses the problem of estimating depth ordering information from single images, a key issue in image understanding that in recent years has focused the interest of the community. Motivation behind this tendency is provided by several important applications that could beneficiate of advances in the field such as automatic object removal, image indexing, 3D scene reconstruction and synthesis.

In contrast to state-of-the-art works, this Ph.D. dissertation investigates a general low-level approach to the problem of monocular depth estimation, in which the depth ordering is directly inferred from a set of monocular depth cues without relying on any previously learned contextual information nor on any assumption on the image structure. New methods for monocular depth cue detection are proposed, the problem of depth cue integration is analysed, and depth ordering information is exploited to improve classical color segmentation, and create new depth-oriented filtering applications. The investigation for depth cue integration leads to the development of two distinct frameworks based on different strategies: a diffusion based strategy and a region merging based strategy.

The former is based on the use of a nonlinear filter which iteratively extends initial depth values arisen from monocular depth cues to the entire image domain until stability is attained. The result is a flexible framework that allows the integration of several monocular depth cues and that gives a correct interpretation of a plurality of vision phenomena.

The latter strategy is based on the construction of a hierarchical region-based representation of images, that incorporates depth ordering information provided by depth cues, as well as on a graph formalization, which encodes depth relationships between regions and allows to infer a global, consistent depth ordering.

Resumen

Esta tesis doctoral aborda el problema de la estimación de profundidad en imágenes monoculares, un tema clave en análisis de imágenes, que en los últimos años ha focalizado la atención de la comunidad científica. La razón de este interés se debe principalmente al considerable número de aplicaciones que podrían beneficiar de avances en el campo, como la recuperación automática de objectos, la indexación de imágenes, la reconstrucción 3D y la síntesis de imágenes.

En contraste con el estado del arte, esta tesis doctoral investiga un enfoque general y de bajo nivel al problema de la estimación de profundidad en imágenes monoculares, en el cual el orden de profundidad se infiere directamente de un conjunto de indicios de profundidad, sin apoyarse sobre información contextual previamente aprendida, ni sobre suposiciones sobre la estructura de la imagen. Se proponen nuevos métodos para la detección de indicios de profundidad, se analiza el problema de la integración de indicios de profundidad, y se explota la información de profundidad para mejorar la segmentación clásica basada en color y crear nuevas aplicaciones de filtrado orientadas a la profundidad. La investigación de la integración de indicios de profundidad lleva al desarollo de dos marcos de trabajo basados en dos estrategias diferentes: una estrategia basada en difusión y una estrategia basada en fusión de regiones.

El primer marco de trabajo se basa en el uso de un filtro no lineal que iterativamente extende los valores iniciales de profundidad a todo el dominio de la imagen hasta alcanzar la estabilidad. El resultado es un marco de trabajo flexible, que permite la integración de varios indicios de profundidad y que da una correcta interpretación de una pluralidad de fenómenos visuales.

El segundo marco de trabajo se basa en la construcción de una representación jerárquica de la imagen, que incorpora información de profundidad proporcionada por los indicios de profundidad, así como sobre una formalización gráfica, que codifica las relaciones de profundidad entre regiones permitiendo inferir un orden de profundidad global y consistente.

Résumé

Cette thèse adresse le problème de l'estimation de profondeur à partir d'une seule image, un problème clé pour la compréhension automatique des images, qui dans les dernières années a focalisé l'intérêt de la communauté. La motivation derrière cette tendance vient de l'importance des applications qui pourraient bénéficier d'avances dans le domaine comme l'élimination automatique des objets dans les images, l'indexation des images, la reconstruction et la synthèse 3D.

Contrairement à l'état de l'art, cette thèse étudie une approche générale et de bas niveau au problème de l'estimation monoculaire de profondeur, où l'ordre de profondeur est inféré directement à partir d'un ensemble d'indices locaux de profondeur sans utiliser d'information contextuelle préalablement apprise ni des suppositions sur la structure de l'image. Des nouvelles méthodes pour la détection des indices de profondeur sont proposées. Le problème de l'intégration des indices locaux de profondeur pour obtenir une interprétation globale est analysé et l'information de profondeur est exploitée pour améliorer la segmentation classique reposant uniquement sur l'information de couleur et pour créer de nouvelles applications de filtrage. La recherche sur l'intégration des indices locaux de profondeur aboutit au développement de deux cadres de travail basés sur deux stratégies différentes: une stratégie basée sur la diffusion et une stratégie basée sur le fusionnement des régions.

La première stratégie repose sur l'utilisation d'un filtre non-linaire qui étend itérativement les valeurs initiales de profondeur provenant des indices locaux de profondeur, au domaine de l'image. Le résultat est un cadre de travail flexible qui permet l'intégration de plusieurs indices de profondeur et qui donne une interprétation correcte de plusieurs phénomènes visuels.

La dernière stratégie fait appel à la construction d'une représentation orientée région et hiérarchique de l'image. L'utilisation de l'information sur l'ordre de profondeur donné par les indices locaux de profondeur ainsi qu'une formalisation graphique, qui encode les relations de profondeur entre régions, permet d'inférer un ordre de profondeur global et consistant.

Ringraziamenti

Innanzitutto, non rigrazierò mai abbastanza Philippe per avermi dato la possibilità di scrivere questa tesi, per la sua preziosa guida durante questi anni, per gli innumerevoli consigli, per il suo essere esigente, per la continua disponibilità e simpatia, per aver dato spazio alle mie idee, per aver saputo evidenziare i miei difetti ed aiutarmi a correggerli almeno in parte, e non ultimo per avermi sempre incoraggiata e dato fiducia.

Un sincero e sentito grazie è per Jean-Michel Morel, per avermi dato la grande opportunità nonchè il piacere di lavorare a suo fianco e lasciarmi contagiare dal suo entusiasmo, per l'interesse dimostrato al mio lavoro, per i suoi preziosi consigli e per l'attenta revisione di questa tesi.

Ringrazio Stéphanie Jehan-Besson e Lionel Moisan per aver accettato di revisare questa tesi ed assieme a loro Ferran Marqués, Josep Ramon Casas e Daniel Bennequin per aver essere stati parte del mio Tribunale di Tesi.

Ringrazio i membri del "Grupo de Imagen" per avermi fornito tutto il supporto di cui ho avuto bisogno. Ringrazio Carlos per la sua irreprensibile assistenza informatica. Un caloroso grazie ai dottorandi del GPS con cui ho condiviso i caffè, i pranzi, le serate, i viaggi, i non pochi momenti di crisi ma anche tanti altri piacevoli momenti. Ringrazio anche i membri del "Centre de Mathématique et de Leurs Applications" per la calda accoglienza dimostrata durante il mio stage all'ENS-Cachan. Ringrazio tanto anche i miei amici di sempre, che ignorano il tempo e la distanza.

Soprattutto, ringrazio di cuore i miei genitori, Erminia e Giuseppe, per essere sempre presenti e dimostrarmi il loro amore. Perchè tutti i miei piccoli e grandi traguardi raggiunti sono in fondo anche un po' i loro.

Grazie....

Contents

Chapter 1

Introduction

1.1 Motivation

Most of digital images are merely a projection of a 3D scene. As a consequence of the projection, objects spatially separated in the 3D world might interfere with each other in the projected 2D plane and each of them occludes part of the ground. Decomposing 2D image data into different objects and determining how objects and surfaces interact in the scene from their 2D projection is usually an effortless task for human vision, but it still represents one of the major challenges that both neuroscience and computer vision are facing nowadays.

The importance of binocular disparity as source of information about the 3D structure of the world emerged from the work of Wheatstone [Whe38], in 1838, when the invention of the stereoscope allowed demonstrating that viewers could achieve a vivid three dimensional perception of an object depicted by a pair of appropriately hand-drafted drawings. Since then, binocular disparity has been for more than one century considered the undisputed source of 3D surface structure. Only after the foundation of the Gestalt School of Psychology, the role of monocular factors has been evidenced. In many chapters of his *Gesetze des Sehens* [Met75] dedicated to depth perception, Metger demonstrates that depth can also be perceived in absence of binocular correspondence.

Although these results were well known at the time Computer Vision emerged as a new discipline, a great deal of effort has been invested by the community in coming up with algorithms to recover depth from stereo [Mar82] and from other cues that requires multiple images, such as structure from motion [Hel25] or depth from defocus [Pen85]. This lack of interaction between Gestalt Psychology and Computer Vision is mainly due to the qualitative nature of Gestalt theory. The mathematical definition of digital image was ignored by Gestaltists and the related issues of blur and noise in image formation were even not qualitatively considered. These practical realities have somehow mitigated the evident strength of monocular factors as source of

depth information in computer-based systems. Indeed, the problems of segmentation with depth and monocular depth cue detection are intimately bound together: a good segmentation mask gives extremely clean contours, useful for monocular depth cue detection, while monocular depth cues constitute an excellent source of information to be exploited in low-level segmentation.

In recent years, motivated by the number of applications such as object removal, image understanding, 3D scene reconstruction and synthesis, that could beneficiate of advances in the field and encouraged by the enormous progresses in machine learning over the last decade, the computer vision community has focused its interest on recovering the spatial layout from single images. State-of-the-art techniques aim to learn the structure of the visual world from a set of training images, in order to attempt depth recovery in unseen test images. Such approaches allow to incorporate prior experience about the structure of the environment as for instance that blue patches are more likely to be the sky and green patches are more likely to be grass on the ground and therefore green patches should be closer to the viewpoint than blue patches.

While learning is crucial for recognition, the issue of whether it really influences basic depth segregation and grouping remains controversial. Nakayama and Shimojo [Nak92] proposed a theory that emphasizes the importance of learning but only in the high stages of the visual perception process, while most visual information is provided by *inherent* cues, which are a direct response to the retinal stimulation. In more recent years, Kellman and Shipley [Kel01] have proposed a framework for object perception that includes depth cues into the grouping process and does not incorporate any feedback from high stage processes to the low ones. They argued that object perception can undoubtedly proceed without such feedback, and likely does so in cases where there is no obvious involvement of learned information. The lack of reliable methods for computing inherent monocular depth cues still represents one of the major limitations not only for a more effective exploration of this fundamental issue but also for a potential exploitation of these cues in computer-based applications.

Standing in stark contrast to state-of-the-art approaches and having image segmentation and filtering as main objectives, this Ph.D. dissertation proposes new methods for monocular depth cue detection in single images that rely neither on previously learned information about the structure of the world nor on any assumption on the image structure. These depth cue detectors open the door to the introduction of an important intermediate layer in applications such as image segmentation and filtering. The integration of depth ordering information provided by depth cues into low-level processing allows not only a more accurate segmentation, which preserves low level structures, but also gives a new image representation, closer to the real world, in which the scene is considered as a set of independent objects with an associated relative depth. This spatial understanding of images can be used as a foundation for other visual tasks, enabling a wide variety of applications such as for example automatic object removal, image indexing, 3D reconstruction and synthesis.

1.2 Approach

This work focuses on a set of inherent depth cues which are pervasive in natural images and are a direct consequence of the process of physical generation of an image. Firstly, the detection of occlusion and all of its various manifestations such as transparency, camouflage and visual completion are investigated. Mainly due to its complementarity with occlusion, we consider a second type of cue: convexity, which in absence of occlusion can be very useful in distinguishing between foreground regions and the background.

In most cases, monocular depth cues provide only a local gradient of depth and therefore, in order to obtain a global depth interpretation, local depth information arisen from different depth cues need to be integrated and extended to the entire image domain. In addition, since different monocular depth cues may provide conflicting information, the integration process must envisage a mechanism to deal with conflicting situations. To these goals, two different approaches have been studied. The first approach is diffusion-based: it consists in iteratively propagating local depth information provided by depth cues to the entire image domain by using a nonlinear filter, which allows to recover both occlusion boundaries and the relative distance from the viewpoint of depicted objects without the need of any explicit segmentation. Conflicting depth interpretations are therefore solved by the diffusion process itself. The second approach consists in constructing a hierarchical region-based representation of images, which incorporates depth ordering information provided by depth cues. By pruning the tree representation, a partition is obtained having the property that pairs of neighboring regions belong to different levels of depth. In order to obtain a global depth ordering, depth relationships between the regions of the partition are encoded in a graph formalization, which allows to easily detect and solve possible conflicting interpretations. The assignment of a relative depth to each region of the partition on the tree representation enables the implementation of a novel depth-oriented filter, which allows one to remove regions belonging to a given level of depth and replacing them with a visually plausible background. The restitution of the removed regions is performed through the use of an image completion technique based on Priority-Belief Propagation.

1.3 Research Contributions

This Ph.D. dissertation provides fourfold contributions related to the estimation of depth in single images. The first contribution is the development of new monocular depth cue detectors, the second consists in the proposal of two different frameworks for depth cues integration, the third is the idea of creating a hierarchical region-based representation of images which exploits monocular depth information, the fourth contribution concerns new depth-oriented filtering applications. The details of the research contributions and publications in each chapter are as

follows.

The main results in chapter 3 address the design of monocular depth cue detectors. Different approaches for the detection of occlusion, transparency and convexity are investigated. Some of the results appear in three conference papers:

- [Dim08] M. Dimiccoli, J.M. Morel, and P. Salembier, "Monocular Depth by Nonlinear Diffusion," in *Proceeding of Sixth Indian Conference on Computer Vision, Graphics and Image processing (ICVGIP)*, December 2008, Bhubaneswar, India.
- [Dim09a] M. Dimiccoli and P. Salembier, "Exploiting T-junctions for Depth Segregation in Single Images," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2009, Taipei, Taiwan.
- [Dim09b] M. Dimiccoli and P. Salembier, "Hierarchical region-based representation for Segmentation and Filtering with Depth in Single Images," in *Accepted for presentation at International Conference on Image Processing (ICIP)*, November 2009, Cairo, Egypt.

Chapter 4 addresses the problem of depth cue integration. Two different lines of research are explored, leading to the proposal of a diffusion-based framework and a region-merging based framework relying on the construction of a hierarchical region-based representation of images that incorporates monocular depth information. These research contributions are presented in following papers:

- [Dim08] M. Dimiccoli, J.M. Morel, and P. Salembier, "Monocular Depth by Nonlinear Diffusion," in *Proceeding of Sixth Indian Conference on Computer Vision, Graphics and Image processing (ICVGIP)*, December 2008, Bhubaneswar, India.
- [Dim09a] M. Dimiccoli and P. Salembier, "Exploiting T-junctions for Depth Segregation in Single Images," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2009, Taipei, Taiwan.
- [Dim09b] M. Dimiccoli and P. Salembier, "Hierarchical region-based representation for Segmentation and Filtering with Depth in Single Images," in *Accepted for presentation at International Conference on Image Processing (ICIP)*, November 2009, Cairo, Egypt.

Chapter 5 develops a depth-oriented filter, which allows to remove image regions following a depth criterion and to replace them in a visually plausible way. The depth-oriented filter, as well as an early version of the main idea which contributes to its realization, have been published in following conference papers:

- [Dim07] M. Dimiccoli and P. Salembier, "Perceptual Filtering with Connected Operators and Image Inpainting," in *Proceeding of International Symposium on Mathematical Morphology (ISMM)*, October 2007, Rio de Janeiro, Brazil.
- [Dim09b] M. Dimiccoli and P. Salembier, "Hierarchical region-based representation for Segmentation and Filtering with Depth in Single Images," in *Accepted for presentation at International Conference on Image Processing (ICIP)*, November 2009, Cairo, Egypt.

1.4 Outline

This Ph.D. dissertation is divided into six chapters, the first of which is this introduction. Chapter 2 presents a summary of related work. Chapter 3 introduces new methods for monocular depth cue detection, focusing on the cues of occlusion, transparency, visual completion and convexity. Chapter 4 investigates the problem of monocular depth cue integration: a diffusion based and a region-merging based frameworks are proposed and a comparative analysis of their performances is discussed. Chapter 5 presents depth-based applications to image filtering. Finally, chapter 6 summarizes the major findings of this Ph.D. dissertation and discusses limitations and possible future lines of research.

Introduction

Motivation

La plupart des images digitales sont simplement la projection d'une scène 3D sur un plan 2D. Comme conséquence de la projection, les objets qui sont séparés dans la scène 3D peuvent interférer les uns avec les autres dans le plan de l'image et chaque objet peut occulter une partie du fond. Décomposer une image 2D en objets différents et déterminer comment les objets et les surfaces interagissent dans la scène à partir de leur projection est une tâche triviale pour la vision humaine mais elle représente encore un défit majeur que la Neuroscience et la Vision par Ordinateur essayent de résoudre actuellement.

L'importance de la disparité binoculaire comme source d'information sur la structure 3D du monde émergea par le travail de Wheatstone [Whe38], en 1838, quand l'invention du stéréoscope a permis de démontrer qu'on peut percevoir un objet en trois dimensions à partir d'une pair de dessins de l'objet convenablement dessins. A partir de ce travail, la disparité binoculaire a été considérée pendant plus d'un siècle comme la principale source d'information sur la structure 3D des surfaces. Il faudra attendre la fondation de l'école de psychologie de la Gestalt pour que le rôle des facteurs monoculaires soient mit en avant. Dans plusieurs chapitres de son *Gesetze des Sehens* [Met75], dédiés à la perception de la profondeur, Metger démontra que la profondeur peut aussi être perçue en absence de correspondance binoculaire.

Bien que ces résultats étaient bien connus au temps où la Vision par Ordinateur émergea comme une nouvelle discipline, un grand effort a été fait par la communauté pour développer des algorithmes pour estimer la profondeur en utilisant l'information stéréo [Mar82] ou d'autres indices de profondeur qui nécessitent plusieurs images comme la Structure par Mouvement [Hel25] et la Profondeur par de-focalisation [Pen85]. Cette absence d'interaction entre la Psychologie de la Gestalt et la Vision par Ordinateur est due principalement à la nature qualitative de la théorie de la Gestalt. La définition mathématique d'image digitale a été ignorée pas les Gestaltiques et les problèmes relatifs au lissage et au bruit dans le processus de formation de l'image n'ont même pas été considérés. Ces réalités pratiques ont d'une certaine façon limité l'évidente force des facteurs monoculaires comme source d'information de profondeur dans les systèmes

de vision automatiques. En fait, les problèmes de segmentation d'image avec information de la profondeur et de détection des indices monoculaires de profondeur sont intimement liés: une bonne segmentation donne des contours extrêmement clairs, qui sont utiles pour détecter les indices de profondeur, alors que les indices monoculaires de profondeur sont une excellente source d'information à être exploitée dans la segmentation de bas niveau.

Au court des dernières années, motivés pas le nombre d'applications comme l'élimination automatique d'objets, l'interprétation automatique des images, la reconstruction et la synthèse 3D, et encouragés par les énormes progrès en apprentissage pendant la dernière décade, la communauté de Vision par Ordinateur a concentré son intérêt sur la récupération de la structure 3D à partir d'une seule image. Les techniques de l'état de l'art visent à apprendre la structure visuelle du monde à partir d'un ensemble d'images d'entrainement, pour pouvoir obtenir la profondeur dans des images de test. Ces approches permettent d'incorporer l'expérience a priori sur la structure de l'environnement comme par exemple que les patches bleus font probablement partis du ciel et que les patches verts font probablement partis du sol et que les patches verts sont probablement plus proches du point de vue que les patches bleus.

Alors que l'apprentissage est crucial pour la reconnaissance, le problème de savoir s'il influence réellement les processus de ségrégation en profondeur et de groupement est encore une controverse. Nakayama et Shimojo [Nak92] ont proposé une théorie qui met en avant l'importance de l'apprentissage mais seulement dans les étapes les plus hautes du processus de perception, alors que la plupart de l'information visuelle est donnée par les indices de profondeur qui sont une réponse directe à la stimulation rétinal. Récemment, Kellman et Shipley [Kel01] ont proposé un cadre de travail pour la perception visuelle qui inclut les indices de profondeur dans le processus de groupement et n'incorpore aucun retour d'information des étapes les plus hautes aux plus basses. Ils soutiennent que la perception des objets peut certainement procéder sans ce retour d'information et quelle procède probablement comme ça dans les cas où l'information apprise dans le passé ne joue pas un rôle déterminant. Le manque de méthodes fiables pour détecter des indices inhérents de profondeur représente encore une des plus grandes limitations non seulement pour une exploration plus effective de ce problème fondamental mais aussi pour une exploitation potentielle de ces indices de profondeur dans des applications de vision par ordinateur.

Par opposition aux approches de l'état de l'art et considérant la segmentation et le filtrage des images comme objectifs principaux, cette thèse propose des nouvelles méthodes pour la détection des indices de profondeur à partir d'une seule image. Ces méthodes n'utilisent pas d'information préalablement apprise et ne fait aucune supposition sur la structure de l'image. Ces nouveaux détecteurs ouvrent la porte à l'introduction d'un niveau intermédiaire dans les applications de segmentation et de filtrage d'images. L'intégration d'information d'ordre de profondeur donnée par les indices de profondeur dans le traitement bas niveau permet non

seulement une segmentation plus précise, qui préserve les structures de bas niveau, mais donne aussi une représentation de l'image plus proche du monde réel, où la scène est considérée comme un ensemble d'objets indépendants avec une profondeur relative associée. Cette compréhension spatiale des images peut être utilisée comme fondement pour d'autres tâches visuelles, habilitant une grande variété d'applications comme par exemple l'élimination automatique des objets, l'indexation des images, la reconstruction 3D et la synthèse d'images.

Approche

Ce travail se centre sur un ensemble d'indices de profondeur qui sont largement répandus dans les images naturelles et sont une conséquence directe du processus de génération physique de l'image. Dans un premier temps, la détection d'occultation et ses différentes manifestations comme la transparence, le camouflage, l'achèvement visuel sont étudiés. Due principalement à sa complémentarité avec l'occultation, nous considérons un deuxième type d'indice de profondeur: la convexité, qui en absence d'occultation peut résulter très utile pour distinguer les régions du premier plan et les régions du fond.

Dans la plupart des cas, les indices monoculaires de profondeur donnent seulement un gradient local de profondeur. Pour obtenir une interprétation globale de profondeur, l'information de profondeur locale qui provient des différents indices de profondeur doit être intégrée et étendue au domaine entier de l'image. De plus, différents indices de profondeur peuvent donner une information conflictuelle. Le processus d'intégration doit donc envisager un mécanisme pour traiter les situations conflictuelles. Pour poursuivre ces objectifs, deux approches différentes ont été étudiées. La première approche se base sur un processus de diffusion: il consiste à propager itérativement l'information locale de profondeur donnée par les indices de profondeur au domaine de l'image en utilisant un filtre non-linéaire, qui permet de récupérer à la fois les contours d'occultation et les distances relatives du point de vue des objets dans l'image sans la nécessité d'une segmentation explicite. Les interprétations de profondeur conflictuelles sont gérées par le processus de diffusion. La deuxième approche consiste à construire une représentation hiérarchique et orientée région de l'image, qui incorpore l'information d'ordre de profondeur donnée par les indices de profondeur. En élaguant la représentation en arbre de l'image, on obtient une partition où toute paire de régions voisines appartiennent à des niveaux différents de profondeur. Pour obtenir un ordre de profondeur globale, les relations de profondeur entre les régions de la partition sont codées en un graphe, qui permet de détecter facilement les situations conflictuelles et de les résoudre. L'estimation d'une profondeur relative à chaque région de la partition habilite l'implémentation d'un nouveau filtre dont le paramètre est la profondeur. Ce filtre permet d'éliminer les régions de l'image qui appartient à un certain niveau de profondeur et de les replacer avec un fond visuellement plausible. La restitution des régions

éliminées est faite à traves une technique d'achèvement visuelle qui se base sur Priority-Belief Propagation.

Contributions de recherche

Cette thèse propose quatre contributions relatives à l'estimation de profondeur dans une seule image. La première contribution est le développement de nouveaux détecteurs d'indices locaux de profondeur; la deuxième consiste en la proposition de deux différents cadres de travail pour l'intégration des indices de profondeur; la troisième est l'idée de construire une représentation hiérarchique, orientée région de l'image, qui exploite l'information de profondeur monoculaire; la quatrième contribution concerne une nouvelle application de filtrage orientée à la profondeur. Les détailles des contributions de recherche et des publications dans chaque chapitre sont décrits à continuation.

Les principaux résultats dans le chapitre 3 concernent le développement de détecteurs des indices de profondeur monoculaire. Différentes approches pour la détection d'occultation, de transparence et de convexité sont étudiées. Les résultats correspondants ont été publiés dans trois articles de conférence:

- [Dim08] M. Dimiccoli, J.M. Morel, and P. Salembier, "Monocular Depth by Nonlinear Diffusion," in *Proceeding of Sixth Indian Conference on Computer Vision, Graphics and Image processing (ICVGIP)*, December 2008, Bhubaneswar, India.
- [Dim09a] M. Dimiccoli and P. Salembier, "Exploiting T-junctions for Depth Segregation in Single Images," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2009, Taipei, Taiwan.
- [Dim09b] M. Dimiccoli and P. Salembier, "Hierarchical region-based representation for Segmentation and Filtering with Depth in Single Images," in *Accepted for presentation at International Conference on Image Processing (ICIP)*, November 2009, Cairo, Egypt.

Le chapitre 4 traite du problème de l'intégration des indices de profondeur. Deux axes différents de recherche sont explorés: le premier axe repose sur la diffusion et le deuxième fait appel au fusionnement de région. Les contributions de recherche sont présentées dans les articles suivants:

- [Dim08] M. Dimiccoli, J.M. Morel, and P. Salembier, "Monocular Depth by Nonlinear Diffusion," in *Proceeding of Sixth Indian Conference on Computer Vision, Graphics and Image processing (ICVGIP)*, December 2008, Bhubaneswar, India.

- [Dim09a] M. Dimiccoli and P. Salembier, "Exploiting T-junctions for Depth Segregation in Single Images," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2009, Taipei, Taiwan.
- [Dim09b] M. Dimiccoli and P. Salembier, "Hierarchical region-based representation for Segmentation and Filtering with Depth in Single Images," in *Accepted for presentation at International Conference on Image Processing (ICIP)*, November 2009, Cairo, Egypt.

Le chapitre 5 développe un filtre sélectif en fonction de la profondeur. Il permet d'éliminer des régions dans l'image en suivant un critère de profondeur et de remplacer les régions éliminées par une forme visuellement plausible. Ce filtre, ainsi qu'une première version de l'idée principale qui contribue à sa réalisation, ont été publiés dans les articles de conférences suivants:

- [Dim07] M. Dimiccoli and P. Salembier, "Perceptual Filtering with Connected Operators and Image Inpainting," in *Proceeding of International Symposium on Mathematical Morphology (ISMM)*, October 2007, Rio de Janeiro, Brazil.
- [Dim09b] M. Dimiccoli and P. Salembier, "Hierarchical region-based representation for Segmentation and Filtering with Depth in Single Images," in *Accepted for presentation at International Conference on Image Processing (ICIP)*, November 2009, Cairo, Egypt.

Plan

Cette thèse est divisée en six chapitres, le premier desquels étant cette introduction. Le chapitre 2 présente un résumé des travaux de l'état de l'art liés au sujet de cette thèse. Le chapitre 3 introduit des méthodes nouvelles pour la détection des indices de profondeur monoculaire, en mettant l'accent sur les indices d'occultation, de transparence, d'achèvement visuelle et convexité. Le chapitre 4 étudie le problème de l'intégration de l'information de profondeur monoculaire: une approche reposant sur la diffusion et une approche faisant appel au fusionnement de région sont proposées et une analyse comparative de leurs performances est présentée. Le chapitre 5 présente une application de filtrage basé sur l'information de profondeur. Finalement, le chapitre 6 synthétise les contributions majeures de cette thèse, discute les limitations et propose certaines directions de recherche futures.

Chapter 2

Background

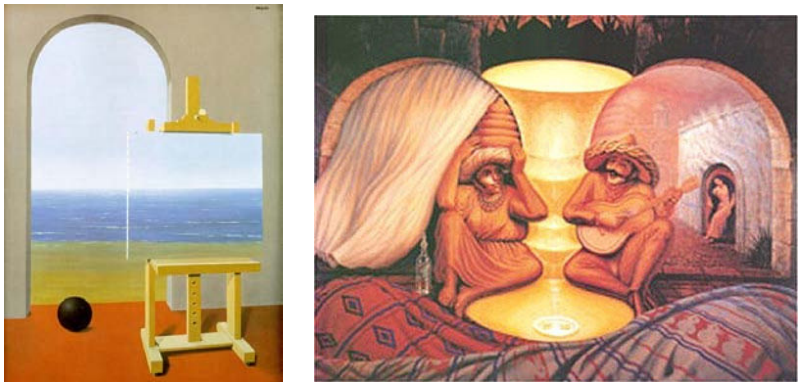
This second chapter provides a brief historical context for the estimation of depth, ranging from early theory of vision to more recent work in computer vision. Section 2.1 introduces the main issues related to depth perception in vision, focusing mainly on the factors that convey information about depth and on the theories attempting to explain their interactions. The connection between depth and object perception is established and the relevance of some monocular depth cues in segmentation and grouping is justified. Section 2.2 overviews the work related to the estimation of depth in single images in computer-based systems, providing a survey on computational models of cues as well as on algorithms that recover depth from single images. Finally, section 2.3 gives an overview of the work developed in this Ph.D. dissertation making explicit the contributions with respect to the state-of-the-art.

2.1 Depth Perception in Vision

Despite the complexity of physical factors that act to generate 2D images on the retina, the visual system is remarkably adept at decomposing 2D image data into objects and recovering their depth relationships. For centuries, scholars have pondered the mechanism underlying this spatial understanding. Though the debate continues on nearly every aspect of depth perception, some understanding is emerging and the current knowledge is broad. This chapter reviews only selected aspects of these topics which are relevant for this dissertation.

2.1.1 Theoretical Frameworks for Depth Perception

During the first half of the 20th century, three main theoretical perspectives on vision have emerged. Helmholtz, often credited as the founder of the scientific study of visual perception, developed a theory known as *unconscious inference* [Hel25], following which retinal images do not provide direct access to objects because of their intrinsic ambiguity and therefore visual



(a) (b)

Figure 2.1: To interpret ambiguous images as these surrealistic paintings, we exploit knowledge of the world around us, learned over many years. (a) *The human condition* by René Magritte. (b) *Forever Always* by Octavio Ocampo.

perception is a matter of inferring a probable interpretation for incomplete data. Inference is based on an accumulation of evidences from a variety of cues as well as on a long history of visual experiences, which provides us with the context to interpret images. For example, our ability to interpret carefully crafted images as those in Figure 2.1 relies on our learned knowledge about the objects involved in the scene.

At about the same time Helmoltz published his research results, a new paradigm of vision was developed within the *Gestalt School of Psychology*, officially initiated in 1912 by an article of Wertheimer [Wer23]. The core of it was the idea that the world is a sensible coherent whole, that reality is organized into meaningful parts, and that natural units have their own structure. The human mind can discover these structures, by understanding the internal rules and principles of the phenomenon itself. Under this perspective, Gestaltists consider human perception as the result of a construction process driven by a set of elementary grouping laws. These laws are supposed to act for every new percept before any high level cognitive process [Met75]. In the founding paper of Wertheimer [Wer23], one can distinguish two kinds of grouping laws. The first kind corresponds to elementary grouping laws that start from the atomic local level to recursively construct larger and larger groups (gestalts). According to this theory, there are six main factors that determine how we group things according to visual perception:

- Proximity: the objects closest together are more likely to form a group (see Figure 2.2 (a)).
- Symmetry: symmetrical items are more likely to group together (see Figure 2.2 (b)).
- Similarity: objects similar in size or shape are more likely to form a group (see Figure 2.2 (c)).
- Continuity: once a pattern is formed, it is more likely to continue even if the elements are

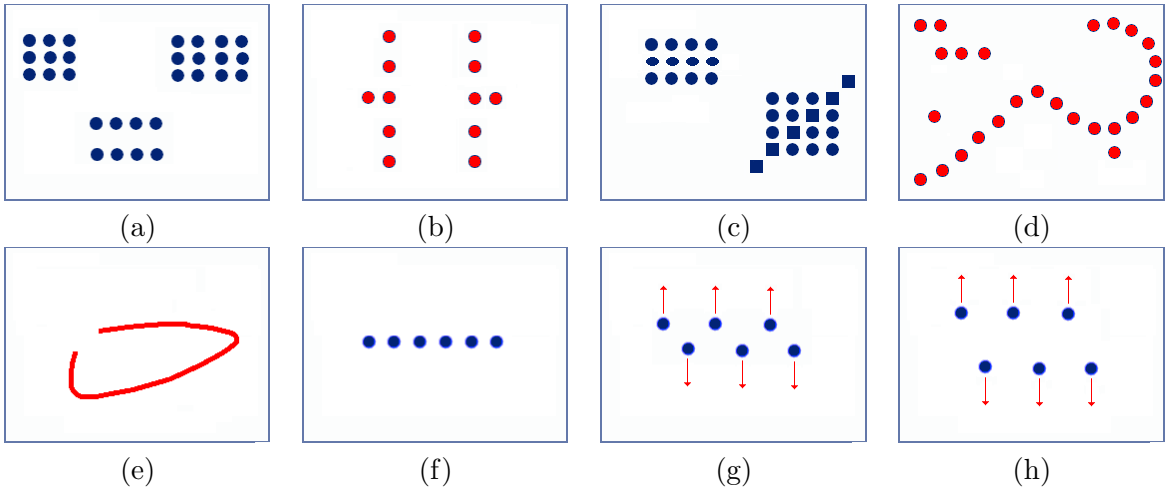


Figure 2.2: Grouping laws: (a) Proximity. (b) Symmetry. (c) Similarity. (d) Continuity. (e) Closure. (f) Common fate: in (g) and (h), odd dots go up and even dots go down so they are grouped together.

redistributed (see Figure 2.2 (d)).

- Closure: our brains add missing components to complete a larger pattern (see Figure 2.2 (e)).
- Common fate: items moving in the same direction are more likely to group together (see Figure 2.2 (f, g, h)).

The second kind covers principles governing the interaction, collaborative or conflictive, between partial gestalts obtained by elementary grouping laws. In fact, since Gestalt laws are independent and act on the same building elements, conflicts may occur between different interpretations. Three cases are possible. The first one, called *collaboration*, occurs when two grouping laws act simultaneously and give rise to two overlapping groups. The second one, called *masking*, occurs when two grouping laws compete and one of them wins inhibiting the other. The last one, called *conflict*, occurs when both grouping laws are potentially active and none of the grouping laws wins clearly.

Contrary to Helmholtz, who completely rejected the possibility of interpreting images without relying on prior experience, Gestaltists believed that retinal images provide direct access to objects but also recognized that visual data are somehow related to neural functioning in the brain. A complete rejection for every form of neural processes has been instead proposed by Gibson [Gib50], who developed an *ecological approach*, following which depth perception does not rely on ambiguous cues in images. Rather, the visual space is defined in terms of surfaces and depth is recovered by directly sensing complex relationships between optical properties that uniquely specify the 3D relationships between surfaces without the need of any neural process.

Central in his theory is that depth discontinuities, together with surfaces, is what gives rise to the perception of the overall surface layout of a scene.

These three seminal theories on vision have influenced recent approaches to depth cue integration, a problem at which increasing attention has been paid over the years since it has arisen also in computer vision, where it is referred to as depth fusion or sensor-fusion problem.

The unconscious inference hypothesis of Helmholtz has recently been revived in the so-called Bayesian studies of visual perception [Rao02]. Proponents of this approach consider that the visual system performs some form of Bayesian inference to derive a coherent perception from sensory data. Models based on this idea are effective since they allow to incorporate prior assumptions about the surface shapes and material properties and have been used to describe various visual subsystems, such as the perception of motion or the perception of depth.

Gestalt Psychology has influenced many strategies in the context of depth fusion, such as veto, disambiguation, accumulation, and cooperation [Bul88]. Veto and disambiguation are inspired from the Gestaltist's concept of masking. Veto implies that the depth computed from one cue determines the perceived depth, overriding the depths computed from other cues, whereas disambiguation refers to the use of one cue to locally disambiguate a representation derived by another cue, when both of them provide inherently ambiguous information. Accumulation and cooperation, are instead inspired in the Gestaltist's concept of collaboration. In accumulation, the depths computed from the various cues are combined additively to provide an overall perceived depth, whereas in cooperation, the effectiveness of one cue can be enhanced by information from another cue, resulting in a combined depth that is greater than the sum of the depths provided by each cue in isolation. The influence of Gestalt psychology is still present in the theory of weak and strong fusion proposed by Clark and Yuille [Cla90]. In weak fusion, the estimates of depth from each cue are linearly combined, whereas strong fusion allows for arbitrary nonlinear interaction between the cues.

The ecological approach of Gibson has inspired two of the most influential recent theories of object [Kel91] and depth perception [Fle04]. Due to the influence of these theories on the work developed in this Ph.D. dissertation, they will be detailed in next section.

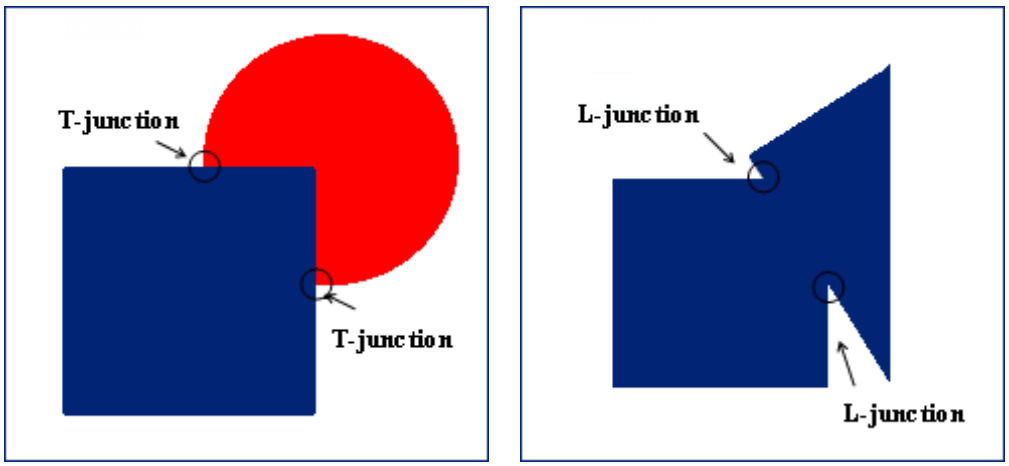
2.1.2 Depth and Object Perception

Depth and object perception are closely related. Consider the image in Figure 2.3 and try to identify and to order different levels of depth. Most readers will say that the *girl* is in front of the sea, that the *sea* is in front of the wood, that the *wood* is in front of the rock and that the *rock* is in front of the *sky*. This example illustrates that humans perceive the 3D scene not in terms of absolute depth map, but in terms of meaningful units which are spatially located. By meaningful units we refer either to surface or objects, referring by objects to bounded volumes

Figure 2.3: Example illustrating our tendency to organize pixels into objects and to place objects in certain spatial relations.

of matter, but at scales relevant to our thought and behavior [Kel98]. The idea of an intimate relation between depth and object perception is central in the Gibson's theory and it has been also more recently supported by experimental evidences by Koenderink et al. [Koe96, Koe98]. However, theories attempting to describe the principles governing the relation between depth and object perception have only very recently been formulated.

Shellman and Shipley [Kel91] proposed a framework for object perception that operates in a bottom-up fashion and starts from the detection of contrast discontinuities, commonly called *contours*, and singular points, commonly called *junctions*, given by the intersection of two or more object contours on the projected image plane. A first difficulty in achieving descriptions of objects from contours and junctions arise from the fact that not all contours in the image correspond to a boundary between two objects. Whereas luminance, chromatic and texture discontinuities may correspond to border lines between different materials or between different facets on a surface, depth discontinuities rarely arises in the absence of a boundary between objects, suggesting that they play a primary role in the detection of meaningful contours in the world. A second, more fundamental difficulty, may be labeled as *fragmentation problem*. Whereas the world and our representations of it contain coherent objects and continuous surfaces, the input from the world to our eyes is fragmentary. Most objects are partly occluded since their projections to the eyes are interrupted by parts of other objects. The process used by the visual



(a) (b)

Figure 2.4: Examples illustrating: (a) The phenomena of occlusion and amodal completion: although the circle is partially occluded by the rectangle, we can perceive it as a complete circle. (b) The phenomenon of modal completion: the rectangle matches the color of the triangle and, as a consequence, their contours are indistinguishable and T-junctions show up as L-junctions. However, we can easily perceive both shapes.

system to overcome fragmentation in the input and produce representations of complete objects is called *visual completion*. It is described in the next section.

2.1.2.1 Visual completion

Observation conditions may lead to partial object obscuration in two different ways. The first one is *occlusion*. In occlusion, an opaque object partly obscures the view of another object further away from the viewpoint (see Figure 2.4 (a)). In this case, the projection of the object boundaries partially hiding each other creates T-shaped junctions in the image plane. The second one is *camouflage*. In camouflage, the object closer to the viewpoint is rendered invisible by matching the color or the texture of another object further away to the viewpoint (see Figure 2.4(b)). In this case, the projection of the object boundaries creates L-shaped junctions in the image plane.

In both cases the visual system interpolates missing data by connecting visible contours across gaps caused by partial object obscuration to produce perceptual units that correspond more accurately to the actual objects in the scene. This process is known as *visual completion* or disocclusion. In the case of occlusion, the perceptual completion of partially occluded objects is referred to as *amodal completion*. In the case of camouflage, the perceptual completion of occluding objects is referred to as *modal completion*. The phenomenon of visual completion has been intensively studied by Kanizsa [Kan79]. Geometrically, it is a variation of the *good continuation principle* [Mon71], in the sense that the interpolated contour has, as much as possible, the same curvature as the pieces of contours it interpolates. Kellman and Shipley [Kel91]

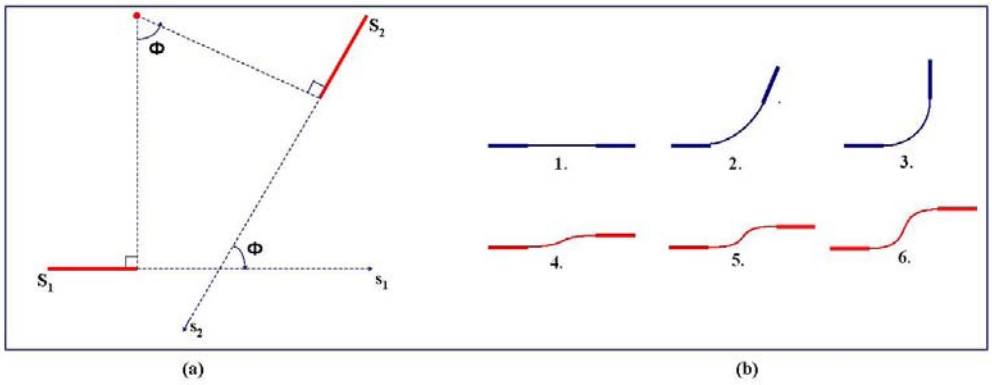


Figure 2.5: (a) Figure illustrating the geometry of relatability. (b) Figure illustrating the strength variations of relatability: the strength of the perceived connections decreases when the angle between two edges increases (1, 2, 3) and/or the offset between two parallel edges increases (4, 5, 6).

have investigated the geometric conditions under which visual completion occurs, developing a theory following which the visual system uses geometric relationships among visible contours to guide the process of interpolation across gaps and to derive shapes. They also observed that contours interpolated between two visible regions invariably begin and end at junctions, which they labeled tangent discontinuities (TDs). In fact, they presented a proof that all instances of occlusion produce TDs in the projected image plane and this invariant may be the reason for which the interpolation processes begin and end at TDs. The geometric relationships between visible contours are synthesized into the concept of *relatability*, whose definition is as follows.

Relatable contours: *Two contours are said relatable if the process of interpolation begins and ends at the points of tangent discontinuity of the contour, and their linear extensions meet in their extended regions (see Figure 2.5 (a)), forming an outer angle less than* $\pi/2$.

The processes of amodal and modal completion are crucial because they have important consequences not only for the perception of shaped contours but also for the global organization of depth. For instance, in Figure 2.4 (a) amodal completion leads to see a red circle *behind* a blue square instead of a polychromatic shape against a white background, while in Figure 2.4 (b) modal completion leads to see a triangle *behind* a square instead of a monochromatic polygon against a white background.

While Kellman and Shipley stated that the placement of completed contours in depth is separable from the process of performing completion, Anderson and Fleming [Fle04] demonstrated that these two processes are intimately and reciprocally bound together. The formation of completed contours and thus of perceptual units, is bounded to the placement of structures in depth and, reciprocally, the placement in depth has a considerable effect on what perceptual

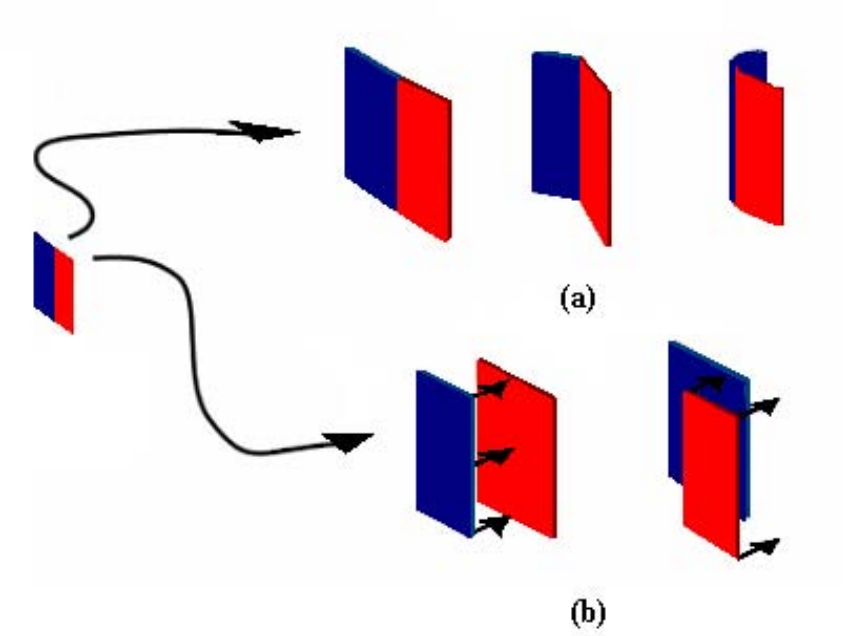


Figure 2.6: A contour which carries a depth signal is ambiguous. In (a) the depth assignment is ambiguous since it could belong to either the blue or the red surface. In (b) the contour could have been originated from an occlusion event. Nonetheless, in all configurations, both sides of the contour are constrained to be at least as far as the depth carried by the contour.

units are formed. They argued that depth perception mirrors the structural organization of the environment by tying its representation of depth to surfaces and objects, placed in certain spatial relations.

The means by which depth is organized into objects relies on visual completion, the Contrast Depth Asymmetry Principle (CDAP) and the Transmittance Anchoring Principle (TAP). The meaning of the last two principles are detailed in the next sections.

2.1.2.2 Contrast Depth Asymmetry Principle

The second principle used by the visual system to constrain all possible depth assignments is the Contrast Depth Asymmetry Principle (CDAP). It is closely related to the Figure/Ground organization, the task of assigning a contour to one of the two abutting regions.

Contrast Depth Asymmetry Principle (CDAP) [Fle04]: *The two sides of a contour with an associated depth value are constrained to either appear at the depth of the contour, or one side of the contour can appear more distant in depth.*

To understand this principle, consider that a local depth value has been assigned to a simple color discontinuity (see Figure 2.6) and that the only depth information present in the vicinity of

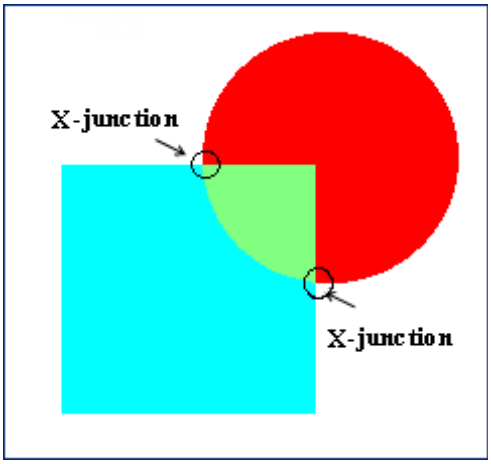


Figure 2.7: A more distant circle partially visible through a transparent square. The projection of the object contours in transparency create X-junctions on the image plane.

the contour arises from the contour itself. The visual system has to determine the surface events consistent with this local image data. If the contour arises from a change in reflectance, surface orientation or illumination change, then the two sides of the contour lie at adjacent positions in depth and therefore both have the depth of the contour. Instead, if the contour arises from an occlusion relationship and therefore corresponds to a depth discontinuity, then the two sides of the contour must be situated at different depths. Because the occluding surface owns the edge [Kof35], it is located at the depth of the contour. The only thing we known about the occluded surface is that it must be more distant in depth than the occluding contour. Thus occlusion introduces an asymmetry in the role of occluded and occluding surfaces: if the occluded surface is untextured, then it could be at any depth behind the occluding surface and the local image data (the local depth value of the contour) would remain the same. By contrast, if the depth of the occluding surface varies, moving closer to the viewpoint, the depth carried by the occlusion boundary must also change, because the occluding surface is responsible for the depth associated with the occlusion boundary.

By itself, this principle does not appear to provide a powerful constraint for image interpretation, because any single local signal discontinuity contains a variety of possible depth interpretations. However, when more than one contrast signal is present, this principle provides a strong constraint on how depth is assigned. Since the CDAP expresses a constraint on the relationship between image contrast and perceived depth, it can be applied to monocular images as well. The basic idea is to assign depth locally and to detect possible conflicting interpretations.

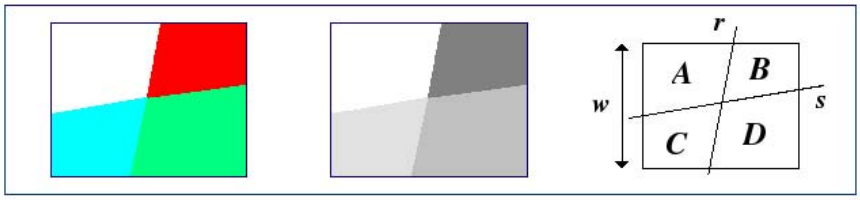


Figure 2.8: The polarity constraint tells us that s is the contour of the transparent object, since the polarity of the contrast between pairs of adjacent regions delimited by r ((A, B) and (C, D)), does not change when s is crossed.

2.1.2.3 Transmittance Anchoring Principle

The CDAP is not sufficient to determine whether a visible edge is seen in plain view or is partially occluded by a transparent surface. Transparency is a particular case of occlusion, which occurs when the occluding object is transparent and therefore the more distant object is visible through the less distant transparent one (see Figure 2.7). In this case, the projection of object contours creates X-shaped junctions in the image plane. Whereas the geometric characterization of T-junctions alone provides a local signature of occlusion, in the case of transparency a photometrical characterization is also needed. At points where transparency occurs, the photometric contributions of two distinct surfaces, a farther opaque surface and an occluding transparent one, collapse into a single luminance value. In order to represent both surfaces, the visual system has to separate a single luminance value into multiple contributions, a process known as *scission*. Scission poses to the visual system two principal problems. The first one is to identify when to a single luminance value correspond two distinct objects located at different depth: the transparent object and the opaque object. The second is to assign different values of luminance correctly at the transparent and the opaque objects. Metelli [Met74] derived two constraints on the photometric conditions required for perceptual scission. The first constraint is known as *magnitude constraint*: a transparent medium cannot increase the contrast of the visible structures. As a consequence, the contrast between the regions in transparency must be lower than the contrast between the regions in plain view. For instance, in Figure 2.8, the gray level contrast between the regions C and D is lower than the contrast between the regions A and B. Thus the regions C and D are candidate to be regions in transparency. The second constraint to be held is known as *polarity constraint*: a transparent medium cannot alter the contrast polarity of structures visible through it. As a consequence, the contrast polarity between the regions in transparency must be the same that the contrast polarity between the regions in plain view. For instance, in Figure 2.8, the contrast polarity between the regions C and D is the same than the contrast polarity between the regions A and B. Once scission has been identified, the problem of assigning surface properties (luminance value) correctly to the two depths is solved

by using geometrical principles which require good continuation of both the underlying and the transparent layer. All these constraints can be summarized in the following principle.

Transmittance Anchoring Principle (TAP) [Fle04] (see Figure 2.8): *When two continuous contours (such as the contours r and s) undergo a discontinuous change in contrast magnitude, but preserve contrast polarity, the regions with the highest contrast along this contour (A and B) will appear as a surface in plain view, whereas the regions with lower values of contrast along such contours (C and D) are decomposed into multiple layers.*

The magnitude of the contrast reduction is used to compute the transmittance of the overlying transparent layer.

Disocclusion, CDAP and TAP constitute the links between depth processing and perceptual unit information.

2.1.3 Depth Cues

A range of different cues are available to assist depth perception. A distinction can be drawn between cues that rely on multiple views, and those cues for which a single view is sufficient. The former category includes motion and binocular cues whereas the latter includes pictorial and configural cues. In the following, each family of cues is detailed.

2.1.3.1 Binocular Cues

It is now well known that stereoscopic vision provides a powerful source of information about the 3D structure of the world. One of the most important sources of stereoscopic depth information is provided by binocular disparity, the positional difference between the two retinal projections of a given point in space. The fact that retinal disparity contributes critically to depth perception was demonstrated by Wheatstone [Whe38] in 1838, when, thanks to the invention of the stereoscope, he could show conclusively that the brain uses horizontal disparity to estimate the relative depths of objects in the world with respect to the intersection point of the two optical axes, a process known as *stereopsis* (see Figure 2.9). Another source of binocular information about the absolute distance of objects in our immediate environment is derived from the physical sensation associated with the actions of muscles in the eyes. One source of information comes from *vergence*, the contraction, when viewing an object that is relatively near, of extraocular muscles aiming to the intersection of visual axes of the eyes at the distance of the object being viewed. The depth distance of an object can be derived mathematically, from the amount of inward turning of the eyes, together with the interocular distance, when both of these quantities are known [Kau74]. The process of maintaining focus on the back of the retina on a near object, through a continuous contracting force of the ciliary muscle, know as *accommodation*, also represents a

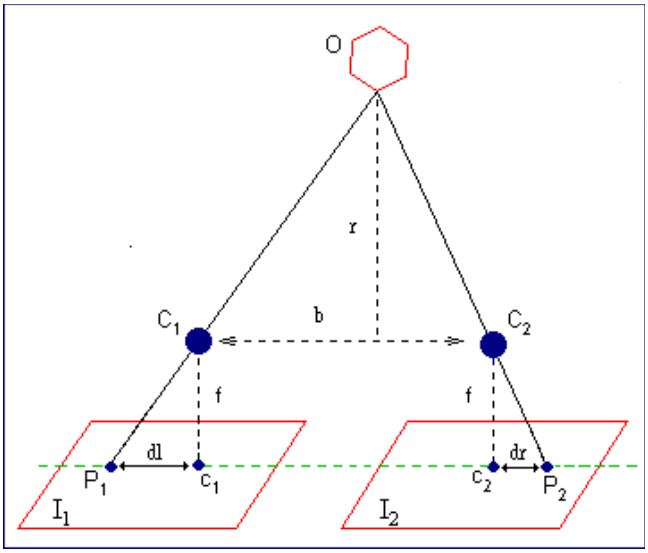


Figure 2.9: A figure displaying the stereo geometry. Two images (I_1 and I_2) of the same object O are taken from different viewpoints (C_1 and C_2). The distance **b** between the viewpoints is called *baseline*. The distance **f** from the center of the lens to the image plane is called *focal length*. In this example, the two camera image planes belong to the same plane and therefore the *correspondence problem*, that is the problem of finding a corresponding point viewed by one camera (P_1) in the image of the other camera (P_2) is greatly symplified: the search space for P_2 is reduced to the line parallel to the baseline and passing through P_1, called the *epipolar line* (dashed green line). In practice, when the two camera image planes do not coincide, the images are trasformed by a process called *rectification* [Har93, Har99], so that pairs of correspondent points lie on a single epipolar line. The horizontal difference from the image center (c_i, $i = 1, 2$)to the projected image point (P_i, $i = 1, 2$), called *absolute disparity*, is **dl** for the left image and **dr** for the right image. The difference between **dl** and **dr**, called *relative disparity*, is directly related to the distance **r** of the object normal to the image plane.

source of depth information. The degree of activation of the ciliary muscle provides a cue to distance [Kau74]. Convergence and accommodation are only effective at close distances and can only provide information about the distance of a single object in the visual field.

2.1.3.2 Motion Cues

Since Wheatstone's discovery, the vast majority of research on depth perception has focused on understanding the mechanism underlying the computation of disparity. Only nearly a century later, a depth cue comparable to stereopsis for its powerful and working principle was formally put forward by Helmholtz [Hel25]. Helmholtz noticed that objects at various distances from the observer move at different velocities on the retinal surface whenever there is translational motion. The capacity to extract 3D shape from motion is known as *motion parallax*. In term of retinal images, while stereopsis has its foundation in the correspondence of depth distances to binocular disparities, motion parallax come as a result of the displacement over time of projected point locations on the retina of a single eye. The similarity is in that both stereopsis and motion parallax depend upon the comparison of the perspective projections of the same scene from different viewing positions.

Motion generates also another kind of parallax field useful to impart a sense of depth, usually credited to Gibson [Gib50]. When an observer walks through a 3D world and looks straight ahead, a global *optical flow* is generated: the entire visual field appears to expands and flow out of the point of fixation of the observer. This pattern of optical flow provides a rich source of visual information about the three dimensional structure of the visual scene. However, motion parallax does not give any information about the depth ordering of independently moving objects. Instead, this information is provided by *dynamic occlusion*. When various objects are moving in the scene, motion discontinuities occur, which give the relative depth ordering between the moving object and the background.

2.1.3.3 Pictorial Cues

Nevertheless, the visual system is remarkably adept at getting a good impression of depth solely on the basis of a single monocular image. In this task, it is influenced to various extents by several factors, commonly referred to as pictorial depth cues because of their use by artists to convey a greater sense of depth in a flat medium. The majority of pictorial depth cues were described and categorized for the first time by Leonardo da Vinci [Vin05] in 17th century in his "Treatise on Painting" and it has been only in the second half of the last century that pictorial depth cues have become object of a new scientific interest, aiming at understanding the mechanisms underlying the so called "enigma of perception".

Considered by Kanizsa [Kan79] as major determinant of pictorial depth, the phenomenon

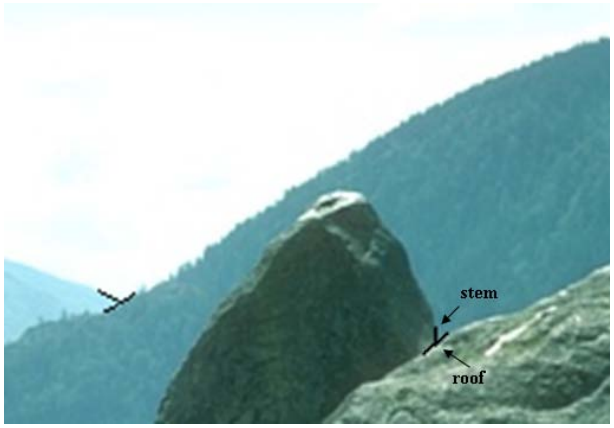


Figure 2.10: A photograph illustrating the effect of occlusion on the perception of relative depth. The key elements in conveying depth ordering are T-junctions.

of *occlusion* was first extensively studied by Chapanis and McCleary [Cha53] (1953). Occlusion occurs when an opaque object partly obscures the view of another object further away from the viewpoint (see Figure 2.10). In this case, as demonstrated by Kellman and Shipley [Kel91], the projection of the object contours partially hiding each other creates T-shaped junctions in the image plane. The geometrical configuration of T-junctions encodes relative depth information of the objects in partial occlusion: the stem of the T belongs to the partially occluded object and the roof to the occluding object. There are particular evidences that our visual system makes fundamental use of occlusion information to encode relative depths of superimposed surfaces at a relatively early stage in visual processing [Kan79].

However, the information conveyed by occlusion gives information about depth order rather than distance. Instead, a number of cues to distance is a direct consequence of the perspective projection. The first one is *linear perspective*. Those principles have been first demonstrated by the artist and architect Brunelleschi [Vas98] about four centuries ago. Lines which are parallel in the three-dimensional model will appear in the projected image to converge toward a single vanishing point as they recede into the distance. This phenomenon is illustrated in Figure 2.11. In general, this effect is emphasized when the lines originate close to the viewpoint and extend a considerable distance away.

A second consequence of perspective projection is that, as an object moves away from the viewpoint, it will subtend a smaller visual angle on the retina and therefore objects that are placed further away will have relatively smaller projected sizes. If the actual relative sizes of depicted objects are known, relative depths can be inferred from differences in the sizes of other projections. This phenomenon, called *relative familiar size* (see Figure 2.12), has been studied by Hittelson [Hit50], who provided experimental evidence of an intrinsic connection between our perception of relative familiar size and relative depth. Instead, Lehmann (cited by Metger

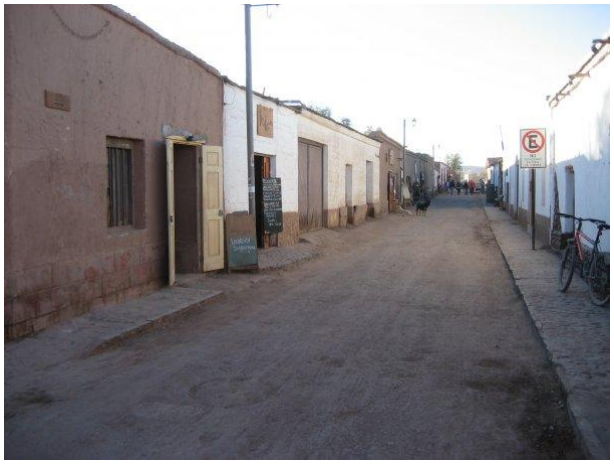


Figure 2.11: A photograph illustrating the phenomenon of linear perspective: lines known to be parallel in the three-dimensional scene appear to converge in the two dimensional projection.

[Met75]) demonstrated that relative familiar size has a subordinate role with respect to other cues of depth by showing that when alternative depth cues are present, they tend to dominate the perception.

A third consequence of perspective projection is that objects having bases below the horizon appear to be further away when they are higher in the field of view, whereas objects having bases above the horizon appear to be further away when they are lower in the field of view. This phenomenon, known as *relative height* has been experimentally illustrated by Gibson[Gib50], who demonstrated the tendency of the visual system to perceive an object as standing on the ground plane over which it is superimposed in a projected view (see Figure 2.13).

Another depth cue, which is a direct consequence of the projection, was introduced by Gibson[Gib50] in 1950. Gibson observed that as a large, planar surface recedes in depth, the projected size, under perspective projection of any unit area of the surface will decrease isotropically with the distance and anisotropically with the inclination of the plane away. If the surface is covered by markings, the projected shape and sizes of these surfaces markings will vary as a function of their distance from the viewpoint and the orientation of the plane. This phenomenon is called *texture gradient* (see Figure 2.14). On the basis of this observation, Gibson argued that humans perceive the visual space in terms of physical surfaces and hypothesized that distance was understood from the density of surface texture while the amount of surface slant was understood from the rate of change of texture density.

Other pictorial cues to depth rely more on color and luminance than on geometry and heavily depend on object properties and illumination conditions. These cues are *shading* and *shadows* (see Figure 2.15), which have been long exploited by artists to convey vivid illusion of depth in paintings. Following the distinction of Yonas [Yon79], shading refers to the luminance

Figure 2.12: A photograph illustrating the relative familiar size cue to depth. The relative distance of people can be inferred from their relative size in this image, because their actual relative sizes are known a priori.

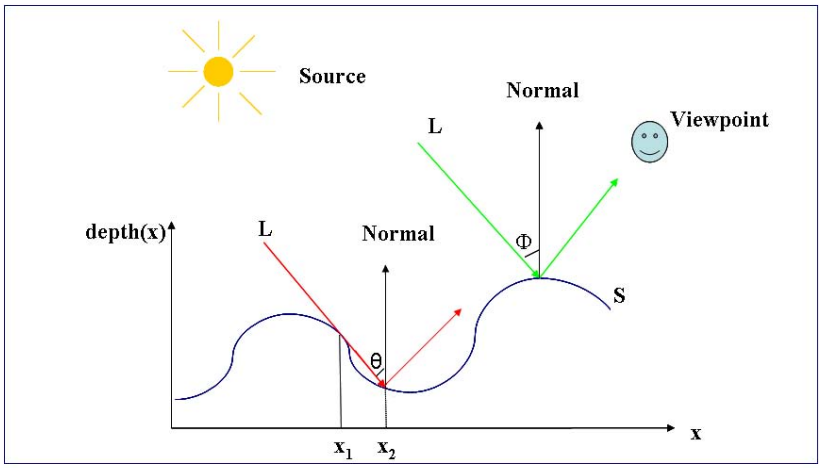
Figure 2.13: A photograph illustrating the phenomenon of relative height: the rectangle that is higher on the field of view is perceived as more distant.

Figure 2.14: A photograph illustrating the texture gradient cue of depth. The density variation of texture elements evokes a subjective impression of a flat surface receding in depth.

distribution on a surface due to illumination by a non-occluded light source, while shadows refers to luminance attenuation on a surface due to light source occlusion. The visual system uses patterns of shading to infer the three dimensional shape of a surface since the amount of light reflected to a point of observation depends on the surface's orientation relative to the direction of the light source. Instead, pattern of shadows are used to derive additional information about the shape of the surfaces that created the shadows as well as the shape of surfaces that the shadows lies on. Shading and shadows are primarily useful as cues to 3D shape more than distance.

Another pictorial cue of depth relying on luminance is *aerial perspective* [Kel77]. Objects that are further away look less contrasted than objects that are closer because images that fall in the retina go through more air and particles, from pollutants or from moisture (see Figure 2.16). O'shea et al.[O'S94] cite numerous psychological experiments confirming that stimuli having a lower contrast luminance are perceived to be more distant. Livingstone and Hubel [Liv87] reported evidence that most pictorial depth cues are difficult or impossible to be perceived under condition of equiluminance and highlighted the central importance of contrast as depth cue.

Apart from classical pictorial depth cues, digital images contain a source of information about depth that cannot be found in artistic depictions of natural scenes: different image regions are often blurred by different amounts, because of *depth of focus* limitations (see Figure 2.17). Pentland [Pen85] provided experimental evidences that the visual system is able to interpret relative differences in focus as indicating a relative depth.



(a)

(b)

Figure 2.15: Depth from shading and shadow.(a) S is a surface, the axe x represents all points of the surface, the axe $depth(x)$ represents the depth of the surface points x_i, L is the light incident on the surface, θ and Φ are the angles that the incident light forms with the normal at the surface on the incidence point. For all the points of the surface between x_1 and x_2 there is a luminance attenuation because the light is not directly incident at these points. Instead, for all remaining points, the depth can be obtained from the shading if the reflectivity function and the position of the light sources are known. (b) A photograph of coins illustrating how the relief is known through shading and shadow.

Figure 2.16: A photograph illustrating the phenomenon of aerial perspective. The effect is observable in the decreased visibility of distant buildings, in contrast with the buildings on the immediate foreground.

Figure 2.17: A photograph illustrating how an impression of distance can be intuitively understood from difference in relative focus. The fact that the leaves are closer to the viewpoint than the waterfall is immediately obvious, although the amount of depth distance is not known.

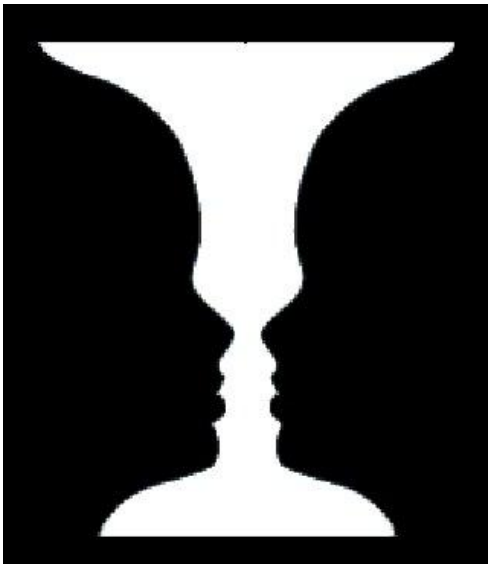


Figure 2.18: Rubin's vase illustrating the Figure/Ground assignment problem. We perceive that each boundary belongs to one, but not both, of the two abutting regions. The figure side has a shape and the ground side is shapeless, extending behind the figure.

2.1.3.4 Configural Cues

Configural cues are commonly described as cues of shape and certainly they fall under that category. However, we decided to discuss them in this section because, as demonstrated also by recent investigations [Bur05, Pet93], they play a prominent role in Figure/Ground organization [Rub21]. Figure/Ground organization is the task of assigning a contour to one of the two abutting regions. It is commonly thought to follow region segmentation and it is an essential step in forming our perception of surfaces, shapes and objects, as vividly demonstrated by the picture in Figure 2.18. This picture is highly ambiguous and we may perceive either side as shaped entities, or *figures*, which are separated from adjacent regions by their bounding edges. However, we cannot perceive both sides simultaneously as shape but always perceive one side as being shapeless and continues behind the figure as background; hence, it is called *ground*. Configural cues are a kind of cues that predict which of two adjacent regions in the visual field will appear to be configured, or shaped (to be figure) versus unshaped (to be ground). By allowing to assign contours to one of the abutting regions, configural cues provide information about relative depth.

The configural cues, illustrated and described in Figure 2.19, were introduced and demonstrated empirically by Gestalt psychologists [Kof35, Kan76, Kof58] and include convexity, symmetry, small area, and surroundedness. In absence of other cues, convex regions are more likely to be seen as figures than adjacent regions that are concave (see Figure 2.19 (a)); small area regions tend to be perceived as figure with respect to large area regions (see Figure 2.19 (b)); regions involving a symmetry with respect to a given axis tend to be perceived as figure with

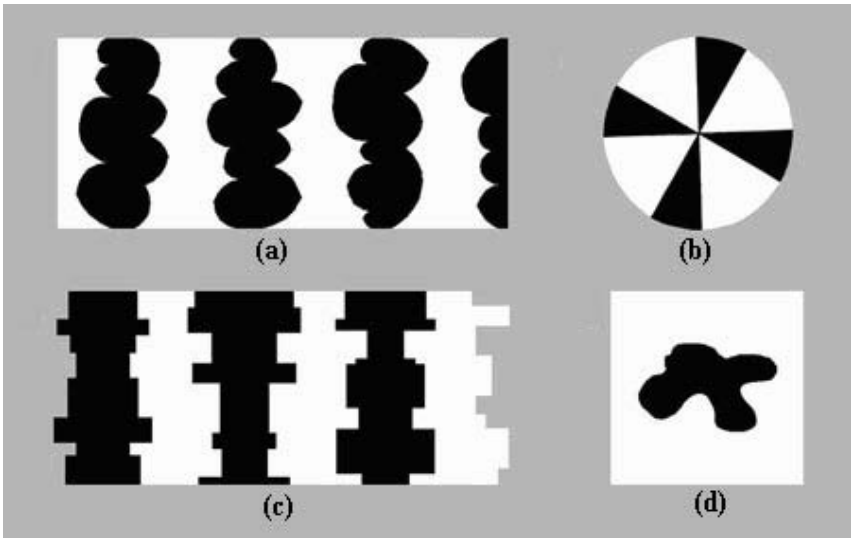


Figure 2.19: (a) Convexity: the black shapes are perceived as figures because, contrary to white regions, they are limited by piecewise convex boundaries. (b) Area: the black triangles are perceived as figures because their are smaller than the white ones. (c) Symmetry: the black shapes are perceived as figure surrounded by the white one. (d) Surroundedness: the black shape is perceived as a figure because it is surrounded by the white one.

respect to adjacent regions that are asymmetric (see Figure 2.19 (c)); regions that are enclosed by another region (see Figure2.19 (d)) are more likely to be seen as figures.

2.1.4 Monocular depth cues versus binocular and motion cues

Binocular, motion and monocular cues have been ranked by Cutting and Vishton [Cut95] according to their effectiveness in determining ordinal depth relationships. Depending on the distance from the viewpoint, they defined three spatial ranges: the personal range, within two meters; the action space, within thirty meters; and the vista space, beyond thirty meters. They found that the effectiveness of a particular cue depends largely on the spatial range. Whereas within the personal space the five most important cues correspond to motion and stereo, in the action as well as in the vista space the top five most important cues are monocular. This result suggests that, although binocular and motion cues have been extensively studied in the past, a large amount of research work is necessary in the fields of both human and computer vision about monocular cues, which are critical for spatial understanding.

2.2 Depth Estimation in Computer Vision

Section 2.1 has reviewed all the factors that convey information about depth, ranging from depth cues which rely on multiple views to those cues for which a single view is sufficient. It has also

pointed out the importance of monocular cues with respect to binocular and motion cues. In this section, we overview how the problem of depth estimation has been approached in computer vision literature, focusing on currently available techniques in the single image scenario, which is the main objective of this Ph.D. dissertation.

Pioneering contributions towards the problem of depth estimation in single images concentrated on high level methods based on a 2D representation of objects called *line drawing*. A line drawing is defined as the set of lines defining the boundaries of the objects and their surfaces (see Figure 2.20). The first contribution on 3D interpretation of objects from a line drawing is due to the system for object recognition proposed by Roberts [Rob65]. Given an object taken from a certain set of a finite number of object prototypes, the Robert's system identifies the object by first extracting a line drawing from the image and next searching for a prototype whose projection coincides with the line drawing. The Robert's system requires a set of strong assumptions such as that objects are isolated in the images, that they are taken from a finite number of prespecified prototypes, and that line drawings can be extracted completely. However, it represents a sound starting point for a 3D prototype-based interpretation of objects from a line drawing and it has been further developed by Falk [Fal72] and Grape [Gra73] so that imperfect line drawings and/or partially occluded objects can also be dealt with.

The problem of a 3D "prototype-free" interpretation of objects from a line drawing was first addressed in 1968 by Guzman [Guz68], in the field of Artificial Intelligence. Guzman tried to find a systematic way of decomposing a line drawing of a pile of objects into regions so that each region may correspond to one object. In his method, configurations of lines at junctions are used as keys for the region decomposition. Though his method was not based on any theoretical foundation but only on a collection of *ad hoc* rules, it showed that it is possible to separate scene into the constituent objects exclusively on the basis of monocular geometric properties, without any knowledge about human everyday experiences, by a relatively simple mechanism based on configurations of lines at junctions. This approach has been further developed in a more theoretical manner by Huffman [Huf71], Clowes ([Clo71]) and Waltz [Wal75], who independently proposed a junction dictionary for categorizing possible combinations of lines at junctions. It has also been extended to handle curved objects by Malik [Mal87] and it has been modeled in algebraic terms by Sugihara [Sug84] and through a Markov Random Field (MRF) by Saund [Sau05].

However, a common limitation of all above mentioned methods is that they rely on a high level representation of images, which is commonly not available and not easy to extract from image data.

A low level approach relying directly on image data has instead been proposed by Mumford and Nitzberg [Nit90]. The goal of their work was not a 3D interpretation of objects but a layered representation, consisting of a set of depth ordered regions. They modeled this problem,

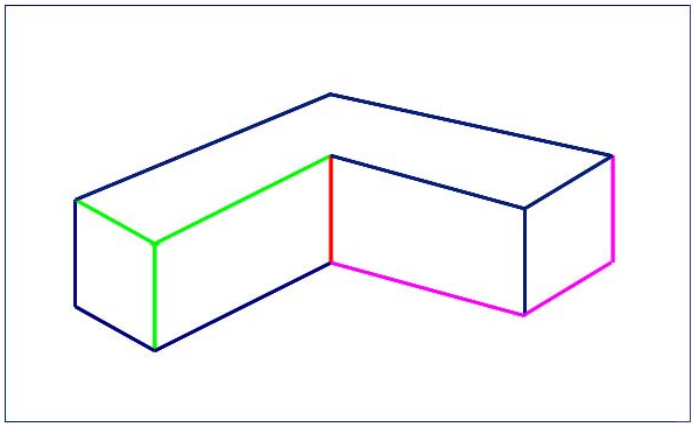


Figure 2.20: Examples of line drawing. We naturally interpret 2D line drawings as planar representations of 3D objects. We interpret each line as being either a convex (angled toward the viewer), a concave (angled away from the viewer) or an occluding contour in the actual object. In this figure, examples of convex, concave and occluding contours are depicted in red, green and pink respectively.

(a) (b) (c)

Figure 2.21: Example of processing by the variational method in [Thi07]: (a) Original image. (b) Initialization: the two objects are manually marked. (c) Result: the contours of both the occluding and occluded objects are reconstructed.

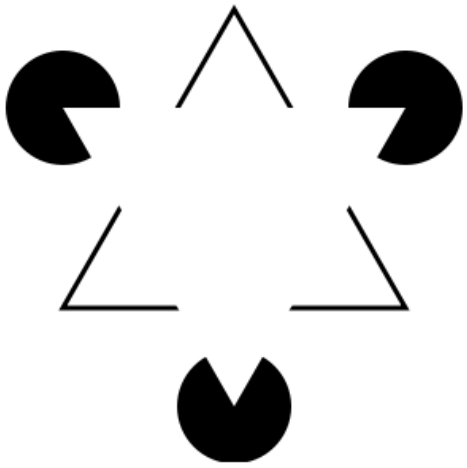


Figure 2.22: Kanizsa triangle. This figure comprises three black pac men approaching each other and three black angles on a white background. But many observers see a white triangle on top of three black disks and an outline triangle. The white triangle appears brighter than the white background and shows a contour even in regions where there is no luminance change in the image.

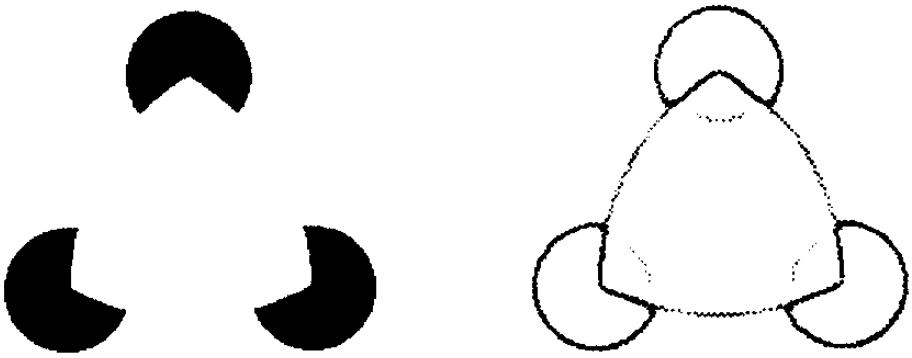
commonly called *segmentation with depth* problem, through a *level sets formulation*, consisting of a variant of the Mumford and Shah's segmentation model [Mum89], allowing regions to overlap. The overlapping of regions gives in a sense the most primitive depth information: if a region occludes another one, one of the objects is in front of the other. Therefore the proposed *2.1 sketch* is an intermediate kind of image representation between the merely 2D segmentation and the 2.5-D sketch of David Marr [Mar82], which aims at a faithful quantitative reconstruction of the spatial environment. Mumford and Nitzberg first compute edges and T-junctions and then minimize the functional combinatorially with respect to all possible ways of connecting the T-junctions by new edges that is consistent with a given ordering hypothesis. The main weakness of this work is that edges, and in turn T-junctions, are computed only based on their intensity profile and, as a consequence, they are likely to be missed in correspondence of occlusion/low-contrast regions. Beside its limitations, this work has inspired more recent theoretical investigation, addressing the issues of the numerical minimization of the functional [Ese03] and the computational complexity [Thi07]. In [Thi07], the computational complexity is drastically reduced by incorporating the treatment of the spatial order estimation within a single energy. In addition, the problem of a more effective computation of occlusion boundaries is solved by introducing prior shape information into the segmentation scheme. However, the range of applicability of the level set framework is still limited to very simple scenes, where occlusion boundaries can be easily detected by their intensity profile or, as in [Thi07], where the shape of objects involved in the scene is known a priori (see Figure 2.21).

Motivated by advances in psychophysic pointing out the importance of monocular depth factors in the interpretation of illusory contours, a number of relevant algorithmic frameworks

for monocular depth perception appeared at the beginning of the nineties and were conceived as computational models for Figure/Ground segregation. The main idea that all these works attempted to demonstrate was that illusions such as the Kanizsa triangle shown in Figure 2.22, are due to the modification of the image according to the depth interpretation at the higher level vision, which is in conflict with physical data of the input. Computational models implementing this idea were in most cases developed by psychologists and have been tested only on a set of very simple image scenes built from flat surfaces of uniform reflectance, called *Colorforms*. They presented solutions based on two different perspectives: the contour processing and the region processing perspective.

From the contour-processing perspective, the formation of a global percept from local cues has been modeled as an optimization process with a contour interpretation mechanism. Under this perspective, illusory contours arise as an extension of visible contours. Williams [Wil90] proposes an algorithmic framework aiming to construct a complete and consistent representation of Colorforms from the incomplete and fragmentary image evidence provided by depth discontinuities, also called occlusion contours. The idea exploited in this work is that whether or not an occlusion contour is completed depends on a non-local process with knowledge of surfaces and occlusion. The Williams's system operates in two stages: a problem posing stage and a problem solving stage. In the problem posing stage, the image evidence, given by measurable image intensity gradients, is collected and incorporated in a contour graph, where each node of the graph corresponds to a point in the image and each edge of the graph connecting two nodes corresponds to an occlusion contour in the image. The mechanics of occlusion of one surface by another are described by a set of integer linear constraints for each node and edge in the contour graph. These constraints, which assure the physical consistency of a contour grouping process with the image evidence, generate an integer linear program. During the problem solving stage, the optimal feasible solution of the integer linear program if found by using the Branch and Bound algorithm [Lan60] and the Simplex algorithm [Bur90]. As a result, a boundary graph is generated, whose edges are labeled with a sign of occlusion and a depth index. The main contribution of this work is the identification of constraints deriving from occlusion and junctions as well as the demonstration that they are relevant to grouping occlusion contours. The main weakness is that constraints are applied only locally and are not able to propagate globally.

Saund [Sau99] proposed a solution to this problem based on the use of an annealing-style optimization framework allowing locally derived constraints to propagate around a junction graph, leading to a global interpretation. The nodes of the junction graph represent occlusion contours, L- and T-junctions, and the edges represent connections between two aligned occlusion contours or between junctions and the occlusion contours contributing to their formation. The shortcoming of this approach is that the conditions guaranteeing the convergence of deterministic simulated annealing-style optimization are not meet in practice and therefore it may fail also in



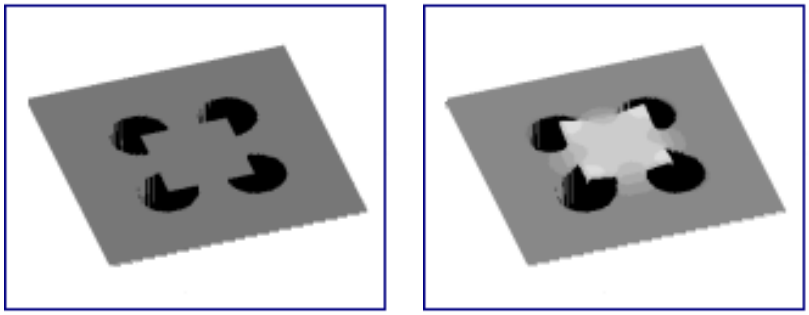
(a) (b)

Figure 2.23: Example of processing by the contour-based method in [Hei93]: (a) Original image. (b) Illusory contours: also the contours of the circles on the background are completed.

simple cases of Colorforms.

Singh and Huang [Sin03], proposed a computational model of transparency, able to detect transparent overlays, to depth order them, and to compute their surface brightness. First, the image is segmented using active contours and X-junctions are detected using a formalization of the Fleming and Anderson's TAP. The influences of local X-junctions are then propagated by searching for circuits of mutually consistent X-junctions in a graph representation. These circuits constitute the boundaries of candidate transparent overlays, and the transparency interpretation is verified in the interior of the region delimited by these boundaries. This model works well when a perfect segmentation is given or can be perfectly computed by active contour method.

Inspired by physiological experiments on the neural mechanisms of monkeys, Heitger et al. [Hei93] proposed a computational model of neural contour processing. Occluding contours are obtained as local maxima in a local neighborhood of the response of a contour operator that sums statistically a representation of 1D signal variations, given by contrast discontinuities, and a representation of 2D signal variations, such as T-junctions, corners and line ends, called *key points*. Key points are obtained by a grouping scheme consisting in convolutions with a set of orientation selective impulse responses followed by nonlinear paring operations. The grouping scheme is selective in the sense that it produces occluding contour signals only if the configuration of key-points is consistent with the interpretation of occlusion. The resulting contour representation includes an indicator of Figure/Ground direction. The main contribution of this model is that it is able to represent occluding contours in absence of consistent contrast, a task which is crucial for explaining illusory contours and segmenting images meaningfully. However, this method cannot resolve possible ambiguities and tends to generate a complete representation also of occluded contours (see Figure 2.23) making impossible a depth ordering.



(a) (b)

Figure 2.24: Example of processing by the diffusion-based method in [Kog02]: (a) Original image. (b) Depth map.

From the region-processing perspective, the formation of a global percept from local cues has been modeled as an optimization process with a surface diffusion mechanism. Under this perspective, illusory contours arise from the surface boundaries. Geiger et al.[Gei96] proposed an algorithmic framework able to select which contours are seen among all possible ones. First, occlusions are detected by identifying local occlusion cues such as junctions, corners and line endings. Some of these cues, as corner and line-endings, have multiple occlusion interpretations in the sense that the specification of which region delimited by the lines forming the cue is the figure and which is the ground, cannot be done locally. Thus, initially, a set of local surface interpretations in terms of figure and ground is assigned in the form of surface-states. Then, a Bayesian linear diffusion algorithm that prevents the diffusion of coefficients at intensity edges is applied. The best image organization, that is the set of contours that are seen, is selected based on a coherence measure between pairs of junctions. This method has been tested only on simple Colorforms, where local cues are easily detected and a simple linear diffusion allows a correct propagation inside regions. In addition, it does not handle simple sorting in presence of multiple depth layers.

Karlsson [Kar04] proposed a model in which depth discontinuities arisen from T-junctions are iteratively extended to the entire image domain by a Gaussian diffusion. Border conditions for avoiding flow of depth across edges are introduced as a sort of artificial depth jumps. This allows to handle simple sorting but the resulting depth map produces surfaces that are only bent around the points where the initial depth discontinuities are located.

Biologically inspired approaches have been proposed by Kogo et al.[Kog02], and Mordohai et al.[Mor04]. Kogo et al. [Kog02] proposed a feedback model, in which the activity occurring in lower areas of the visual perception process are used to modify activity occurring in higher areas (the perception of the lightness of the image). The activity occurring in lower areas consists in analyzing local properties and to extract relative depth information, represented by the differentiated form of the signal. The global integration of the signal determines the perceived depth

which, in turn, modifies the activity in higher areas consisting in the image lightness perception (see Figure 2.24). In illusory contours, such as the Kanizsa triangle, this model creates a central surface and extends the contours from the L-junctions. The relative depths are determined by convolution of Gaussian derivative based filters, while the feedback loop is conducted through an anisotropic diffusion equation [Pro99], which constitutes the link between lightness perception and depth perception. This method only works on Colorforms, for which relative depth information can easily be extracted from the image gradient and the feedback loop can be properly driven by the Proesmans's anisotropic diffusion equation.

Mordohai et al.[Mor04] proposed a *tensor voting framework* for automatic selection between modal and amodal completion on binary images. This work is based on a data representation formalism that uses second-order symmetric non-negative definite tensors and an information propagation mechanism termed tensor voting. Based on the information propagated via tensor voting, the saliency of completed contours is computed. Saliency indicates the quality of a feature to be perceptually important. The input of the algorithm is a set of tokens, which in principle may correspond to any type of local primitives but usually corresponds to bright discontinuities. Each token is represented through a second order tensor, or equivalentely, through an ellipse whose shape encodes the preferred orientation of the token and whose size encodes the saliency of the information encoded along with its normal and tangent orientations. All tokens are initially represented as unoriented ball tensors. A polarity vector is also associated to each token. The magnitude of the polarity vector indicates the difference between the number of neighbor tokens that lie on the two sides of the token under consideration and it is therefore very high in the case of line ends (see Figure 2.25). The direction of the polarity vector indicates the direction of the completed contour having the token under consideration as potential boundary. The information encoded by the tensors is propagated to their neighbors via first and second order voting. During the voting process, each token communicates its preference for structure type and orientation to its neighbors in the form of votes, which are also tensors that are cast from tensor to tensor. Each vote has the orientation the receiver would have if the voter and the receiver were part of the same smooth perceptual structure. If the tokens align to form a curve, then the accumulation of votes will produce high curve saliency and an estimation for the tangent at each point. The main advantage of the propagation mechanism is in that it avoids premature decisions based on local operators: a preliminary label, based on the result of first and second order voting is assigned only if both the ball saliency and the polarity are high. The final labeling occurs only after all the completion possibilities have been examined.

Maradarasmi et al.[Mad94] proposed a MRF formulation: assuming that all surfaces in the scene are piece-wise constant, the problem of finding a piece-wise smooth segmentation of the image into surfaces and providing a relative depth ordering between the regions of the partition, is equivalent to the problem of assigning a discrete depth value to each image pixel. Initially,

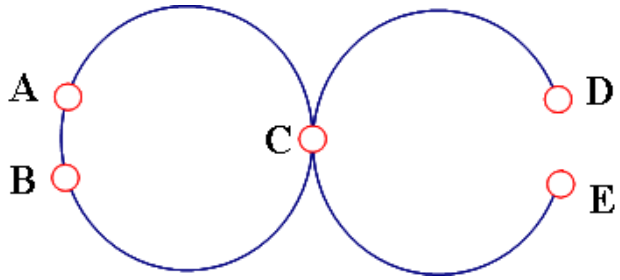
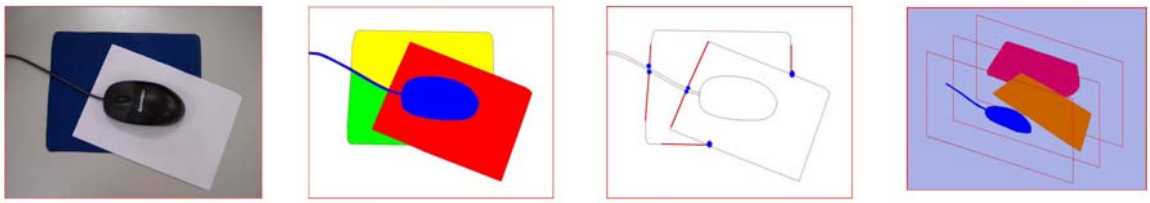


Figure 2.25: The endpoints D and E are also smooth continuations of their neighbors and therefore they cannot be qualitatively discriminated from points A and B using second order properties only. The polarity vector captures the fact that D and E are terminations of the curve while A and B are not. In fact, the polarity vector has a magnitude very close to zero for A and B, which are surrounded by others curvels, and locally maximum for D and E, which receive strong support from one side of the contour only.

image features such as T-junctions, corners, convex contours and line terminations are detected and a local cost, quantifying the probability that a given pixel belongs to an occluding surface, is assigned based on these features. At each pixel, an initial random depth value is assigned. Data constraints arising from hypothesized local occlusions as well as prior constraints arising from the assumption of piece-wise smooth image, are embedded into an energy cost which is associated to each image pixel. The best depth label assignment corresponding to the maximum a posteriori probability is then determined by minimizing globally the assignment cost. Due to the assumption of piece-wise constant surfaces, this method only works on Colorforms.

Stella et al.[Ste01] extended Maradarasmi's work by proposing a hierarchical MRF, which introduces a topology-dependent multiscale hierarchy to facilitate long range propagation of local occlusion. The model incorporates explicit decision rules that assert continuity of depth values along contours and within surfaces, and discontinuity of depth values across contours. The parameters that encode the relative importance of these decision rules are estimated by a learning method proposed by the same authors, which does not assure that the learning will actually converge. The algorithm has been tested only on images of simple shapes for which the input edge maps are assumed to be made of closed contours.

Recently, Gao et al. [Gao07] proposed a Bayesian inference framework which unifies the contour-based and the region-based perspectives. First, atomic regions are computed manually by interactive operations. Then, T-junctions are computed on atomic regions and broken into terminators (see Figure 2.26). A terminator corresponds to the point of intersection of the T-junction stem with the occlusion contour. A graph representation is constructed consisting of two types of nodes: atomic regions and their corresponding terminators. With this graph representation, the problem of depth ordering the atomic regions and reconstructing the contours of the occluded regions is formulated as graph partition problem. The partition problem consists



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

Figure 2.26: Example of processing by the hybrid method in [Gao07]: (a) Original image. (b) Atomic regions. (c) Computation of terminators (in blue) from T-junctions. (d) Final layer representation.

of assigning to each atomic region a depth layer from a discrete set of depth variables, and to each terminator the depth layer that corresponds to it in amodal completion and that is constrained to lie on the same depth layer. The inference is based on the Swendesen-Wang Cut algorithm [Smi05].

Having been conceived to understand the computational mechanism underlying visual perception rather than for being extensively used in machine vision applications, these works have been tested only on a small set of Colorforms [Wil90, Sau99, Gei96, Mad94, Kog02, Mor04, Thi07] or on real images previously segmented by interactive methods [Ste01, Gao07]. As a consequence, these methods avoided to address the problem of reliable depth cues detection in real images, which represents one of the major limitations for segmentation with depth applications. These characteristics made the extension of their field of applicability to natural images not straightforward and in most of cases impossible.

During the last decade, the computer vision community has been increasingly aware that to deal with natural images, grouping processes cannot depend merely on low-level signal processing but should rely on specific knowledge about objects and their relationships to the environment. This awareness has been accompanied by enormous progresses in machine learning which have provided a new learning-based paradigm in machine vision. This paradigm has inspired a number of works on depth estimation in single images [Ren06, Sax07, Hoi07, Rot09]. All these works are based on the use of a large database of ground-truth images annotated with human-marked or collected by a laser scanner ground-truth. Ground-truth images are used to collect and incorporate as much information as possible on the structure of the visual world. All collected information is used to gradually infer the most likely depth assignment on unseen test images.

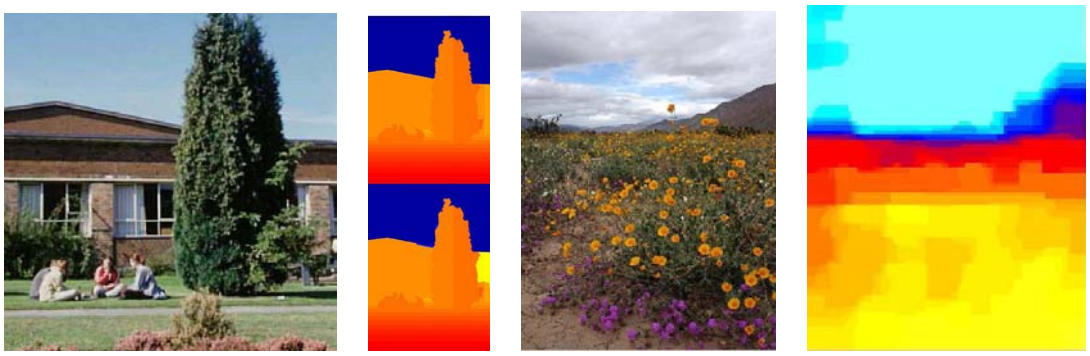
Following this approach, Delag et al. [Del06] presented an algorithm for predicting depth from monocular image features on indoor images. Their approach is based on a set of strong assumptions such as the scene consists of ground/horizontal planes and vertical walls, that the position as well as the calibration matrix of the camera are known, and that the image contains a set of vanishing points. First, vanishing points are extracted using standard techniques. Then,

the location of the floor boundary is estimated in every column of the image using a trained Dynamic Bayesian Network (DBN), which uses a set of 50 local image features, including standard multi-scale intensity gradients, local color samples, and a similarity measure between local color samples and the floor chroma. Finally, by applying perspective geometry, the 3D geometry of the scene is reconstructed. When the image is generated under perspective projection and contains only a floor and vertical walls, this approach gives a complete 3D reconstruction.

Using the same assumption that the scene consists of ground/horizontal planes and vertical planes but focusing on outdoor images, Hoiem et al. [Hoi07] proposed to recover occlusion boundary and depth ordering by learning models of occlusion based on a large set of both 2D perceptual cues and 3D surface depth cues. Many of the relevant cues, such as texture histograms, perspective, or relative depth, require good spatial support to be useful. As a consequence, the authors start from an initial oversegmentation and then, by incorporating additional information as it becomes useful from learned models, they reduce the number of regions to gradually infer the occlusion boundaries and occlusion relationships, reasoning together about 2D boundaries and 3D surface. The inference is performed on a Conditional MRF allowing the join inference over both boundary and surface labels.

These methods do not apply to the scenes that are not made up only of vertical surfaces standing on an horizontal floor. Rother and Sapiro [Rot09] have very recently proposed a probabilistic framework for 3D object reconstruction from a single image which exploits prior knowledge about the class of object being involved. The 3D reconstruction is the result of the interplay between the 3D prior shape and the background probability computed from a single input image using a background model. The 3D prior shape is a structure that encodes the probability that a given portion of the 3D space is occupied by an object of a given class, when an object of the given class is placed at the reference position. This approach only applies to a given class of objects.

Saxena et al.[Sax07] proposed a framework able to estimate the 3D structure of an unstructured environment, which tries to model the fact that the environment is reasonably structured and that humans are usually able to infer a nearly correct 3-D structure using prior experience. They assumed that the environment is made up of a small planes and starts from an over-segmentation of small regions, called superpixels, each of one is assumed to lie entirely on only one planar surface. For each small patch in the image, a MRF is used to infer the set of parameters that capture the 3D location and 3D orientation of the patch. The MRF is trained with supervised learning to learn the relation of local monocular cues to the 3D structure. For example, their model is able to learn that green patches are more likely to be the grass on the floor as well as that green patches have a very different texture when viewed close up than when viewed far away and that therefore different depths values can be assigned to different green patches depending on their texture variations. Local image cues alone are insufficient to infer the



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

Figure 2.27: Example of results obtained by the learning-based methods in [Hoi07] ((a) and (b)) and in [Sax07] ((c) and (d)): (a) Original image. (b) Minimum and maximum depth estimates, obtained ranging the parameters (red=close, blue = far). (c) Original image. (d) Corresponding depth map.

3-D structure since, for example, a blue patch may be a part of a blue object and not of the sky. To account for ambiguities like these, depths are determined by looking at the overall organization of the image. This is achieved by modeling the relations between various parts of the image through the MRF and by learning these relations basing on a set of training images, in which the ground truth depths were collected using a laser scanner. This approach does not make any assumption about the structure of the scene, such as the assumption by Delage et al.[Del06] and Hoiem et al.[Hoi07], nor on the class of objects being involved such as in [Rot09]. However, its strength is heavily based on photometrical properties of natural outdoor environments while the inference for occlusion boundaries is soft and typically not accurate.

Ren et al. [Ren06] formalize the concept of familiar configuration by constructing prototypical local shapes, called shapemes, from image data. Based on shapemes, they train a logistic classifier to locally predict Figure/Ground labels by using a conditional MRF to enforce global Figure/Ground consistency at T-junction which is approximately optimized using loopy belief propagation [Fre00]. Given a perfect segmentation, their methods produce impressive results on natural images, but performance drops dramatically without perfect segmentation, suggesting that the main difficulty is in finding occlusion boundaries, rather than labeling them.

Summarizing, the most successful approaches on real images are learning-based. However, they rely on strong assumptions on the image structure [Hoi07, Del06] or on the image content [Rot09] and in all cases do not produce accurate occlusion boundaries [Ren06, Del06, Sax07, Hoi07, Rot09] (see Figure 2.27).

In the next chapter, we introduce a general approach to monocular depth estimation, which can be in principle applied to any kind of images, without any restriction on the structure nor in the content and that produces accurate occlusion boundaries.

2.3 Approach Overview

From section 2.1 two important issues have emerged. The first is that the processes of object and depth perception interact with each other simultaneously suggesting that the search for object contours should go together with the search for occlusion boundaries. The second is that one of the most powerful monocular depth cue is occlusion and that, in absence of occlusion, configural cues play a crucial role in identifying the foreground objects. These results suggest the idea of treating image segmentation and depth segregation under an unified framework, by incorporating depth ordering information provided by occlusion and configural cues into a grouping process.

In section 2.2 we have reviewed some recent works [Hoi07, Sax07], developed in parallel to this dissertation, that produce a segmentation due solely to occlusion boundaries. In both of these works, most of the cues considered by the authors for computing occlusion boundaries require a meaningful spatial support, that usually consists of small patches resulting from an oversegmentation. Each patch is assumed to lie entirely on only one planar surface and thus to belong to a single object. However, the unreliability of the gradient profile in a neighborhood of T-junctions makes this assumption difficult to hold in correspondence of these singular points, compromising the accuracy of the segmentation. During the last decade, the awareness that junctions are crucial for the inference of 3D scene properties seem to have been accompanied by a lost of faith in the ability of bottom-up grouping strategies. The new paradigm is to do minimal bottom up processing, without even assuming that linked contours or junctions can be extracted, and to use prior experience as primary engine for parsing an image.

In this Ph.D. dissertation, we take a different approach in which depth cues are computed on a pixel-based representation of the image and used for constraining a grouping process. Our methods for computing depth cues do not rely on any assumption on the image structure nor on the class of objects being involved and can therefore be applied to any kind of image. The advantage of this approach is that it allows to obtain an accurate segmentation in term of occlusion boundaries, which is one of the main objective of the present work. We propose two different frameworks for monocular depth cue integration: a diffusion-based framework and a region-based reasoning framework. The former consists in iteratively propagating local depth information provided by depth cues to the entire image domain by using a nonlinear filter, which allows to recover simultaneously the shape and relative depths of depicted objects. The second approach consists in constructing a hierarchical region-based representation of images, which preserves previously detected T-junctions. By pruning the tree representation, a partition is obtained having the property that pairs of neighboring regions belong to different levels of depth. Depth relationships between the regions of the partition are then encoded in a graph, which allows to easily detect and solve possible conflicting interpretations, leading to a global

depth ordering. The obtained image tree representation, in which each region has been labeled with a relative depth value, enables the implementation of depth-oriented pruning strategies, allowing to remove regions following a depth criterion and to replace them with a visually plausible background.

Chapter 3

Monocular Depth Cue Detection

In this chapter, we turn our attention to the problem of automatically detecting inherent depth cues on real images. The challenge is to develop detectors reliable enough to be exploited in segmentation and filtering applications. Our focus is not only the localization precision, which is important for accurately recovering occlusion boundaries, but also a correct semantic interpretation, which becomes crucial for analyzing depth relationships between regions in the global image context. Section 3.1 is devoted to the depth cues of occlusion. A review on junction detection is done, evidencing the limitations of currently available techniques. Two different approaches for overcoming the drawbacks of state-of-the-art algorithms are investigated and a comparative evaluation of their performances is discussed. Section 3.2 focuses on transparency: an algorithm for detecting surfaces in transparency and determining their depth order and medium brightness is proposed. Section 3.3 concentrates on visual completion: both amodal and modal completion are considered and an algorithmic translation of the relatability concept is described. Section 3.4 puts emphasis on the configural cue of convexity: a local approach is proposed and experimental results are detailed. Finally, section 3.5 summarizes the main findings of this chapter.

3.1 Occlusion

As discussed in chapter 2.1, geometric signatures of occlusion and transparency are respectively T- and X-junctions. They are the projections of points where the contours of two objects respectively in occlusion or in transparency meet. The piece of each contour that emanates from the junction point is defined as a *branch*. The orientations of the branches encode the semantic of a junction. As a consequence, junction detection should involve two sub-tasks: localization of junction points and branch extraction. In the literature, the search for T- and X- junctions in natural images has been addressed as the more general problem of junction detection. The following section reviews current techniques, omitting the ones that limit themselves only to corners [Kit82, Har88, Smi97, And00, Mik02, Mik04, Low04] or that rely on the use of multiple

(a) (b) (c)

Figure 3.1: Example extracted from [Bie87]. (a) A line drawing with different kind of junctions. (b) When all junctions are removed from the line drawing it is difficult to recognize objects. (c) When all straight or smooth lines are removed it is still easy to recognize objects.

images [Bau00, Fav03, Apo06] since they are beyond the goal of this work.

3.1.1 Related work

Junction detection is a classic problem in computer vision, which has focused the interest of the community for over thirty years. This long time research effort is justified by the fact that junctions provide crucial information about 3D geometric properties of surfaces and occlusions. A number of psychological experiments have shown the importance of junctions at all levels of visual perception: beside shape recovery and object recognition from a single 2D image, junctions intervene in contour extraction as well as in contour initialization and tracking [Bie87]. A clear idea of the power of junctions for visual perception can be obtained by looking at a simple line drawing (see Figure 3.1(a)): in the complete absence of any other feature, the presence of different types of junctions, their psychophysical interpretation as well as the interactions between them lead to a vivid three dimensional impression. In addition, as demonstrated by Biederman [Bie87] through psychological experiments, removing junctions from the image impedes perceptual recognition (see Figure 3.1(b)) while removing most of the straight edges does not (see Figure 3.1(c)).

Commonly, it is believed that junctions can be detected without any knowledge of physical phenomena and objects in the world that caused them. Therefore, they are considered as local features, that is as local image patterns that differ from their immediate neighborhood by their structure. This explains why most of currently available techniques act by analyzing a local neighborhood of each image pixel.

Early junction detectors rely on a convolution-based approach, which includes gradient-based and edge-based methods. Both are based on the scale-space theory, following which image features are considered at a given location and at a given scale. The scale quantifies the amount of smoothing performed on the image before computing the feature in order to suppress high

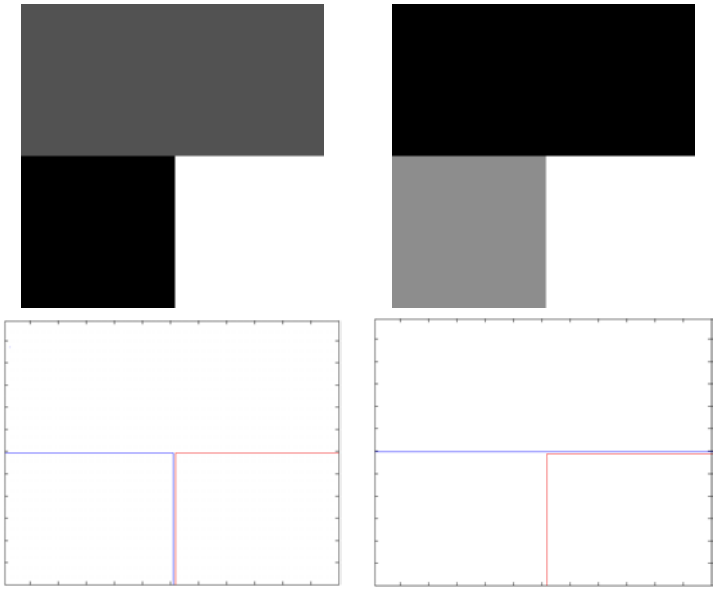


Figure 3.2: Possible configurations of level lines at T-junctions depending on the gray level order. (a) Two level lines change direction abruptly. (b) One level line changes direction abruptly.

spatial frequencies due to noise and/or texture. Roughly speaking, the locality refers to the fact that an image feature at position x is computed by using information from a small neighborhood of x. The importance of localization is related to the interposition problem: digital images are made of a superposition of different objects partly hiding each other or belonging to the ground. A convolution with a spatially large impulse response would confuse the gray level values of pixels belonging to different objects and would remove all sharp edges. Consider the classical Gaussian smoothing operators defined by

$$G_t(x) * u(x) = \int_{\mathbb{R}^2} G_t(\mathbf{y}) u(x - \mathbf{y}) d\mathbf{y}, \tag{3.1}$$

where $u : \Lambda \subset \mathcal{R}^2 \rightarrow \mathcal{R}$ is a real image, $x \in \Lambda$, and $G_t(x) = (1/4\pi t)e^{-\|x\|^2/4t}$. If $t > 0$ is small, then the Gaussian G_t is well localized around zero and $G_t(x) * u(x)$ is essentially an average of the values of $u(x)$ in a small neighborhood of x. Instead, if t is not small, the image gets blurred by the convolution.

Gradient-based methods hypothesize junction locations by analyzing local gradient and level line curvature [Koe88, Rom91]. Level lines, also called *isophotes*, are the boundaries of isolevel sets, that is the family of sets $\{x|u(x) = \lambda\}$, where u is a real image which takes real values in the set $[0, 255]$, x is a point of the image domain, and $\lambda \in [0, 255]$. Looking at the family of level lines of an image, also called *topographic map* [Cas96b], T-junctions can be defined as points where two level lines that are going together take opposite directions. Indeed, depending on the order of gray levels, one or both level lines change direction abruptly (see Figure 3.2) and

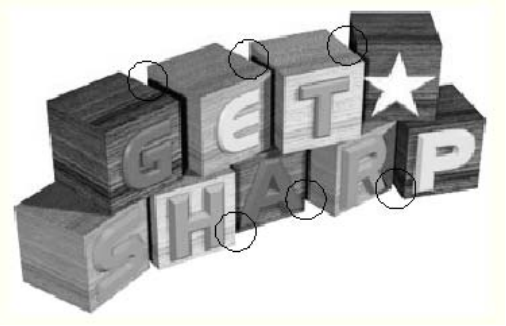


Figure 3.3: Example of result obtained performing the method proposed by Romeny et al. [Rom91].

thus the level line curvature is always high in correspondence of T-junctions. Relying on this geometric reasoning, Koenderink et al. [Koe88] detected T-junctions as extrema in the level line curvature multiplied by the gradient magnitude raised to some power. Instead, Romeny et al. [Rom91] detected T-junctions using the gradient of isophote curvature (see Figure 3.3). Indeed, isophotes show a large change in curvature at T-junctions over a relatively small spatial scale. In all these methods, the signal is smoothed by convolution with Gaussian impulse responses of different width before computing local gradient and curvature, which are tracked through different scales to localize the junction point.

To handle the problem of the scale selection from the scale-space representation, Lindenberg [Lin94] proposed a two-stage approach with detection at coarse scales followed by localization at finer scales. Both scale levels are detected automatically through general heuristic principles based on the study of the evolution of properties over scales of the image filtered by certain non-linear combinations of normalized Gaussian derivative. Initial hypotheses about interesting scale levels are generated from scales where the normalized level curve curvature rescaled by the gradient magnitude assume maxima over scales. Based on this scale information, a more refined processing stage is used to determine the point x that minimizes the perpendicular distance between x and all tangent lines in a neighborhood of a candidate point over scales. This type of approach increases significantly the complexity of both the detection and localization steps since it implies to compute derivative at all scales.

Due to the use of a Gaussian convolution all above mentioned methods are not well localized except at fine scales. Furthermore, they do not deliver any semantic interpretation of the junction in terms of branch orientations.

A more complete characterization of junctions in terms of branch orientations is instead accomplished by edge-based methods that detect junctions as intersection of edges [Har88]. The problem with this strategy is that the edge-detection step is performed by the Canny's edge detector, which lets the edges vanish in a neighborhood of the junction center. Indeed, the

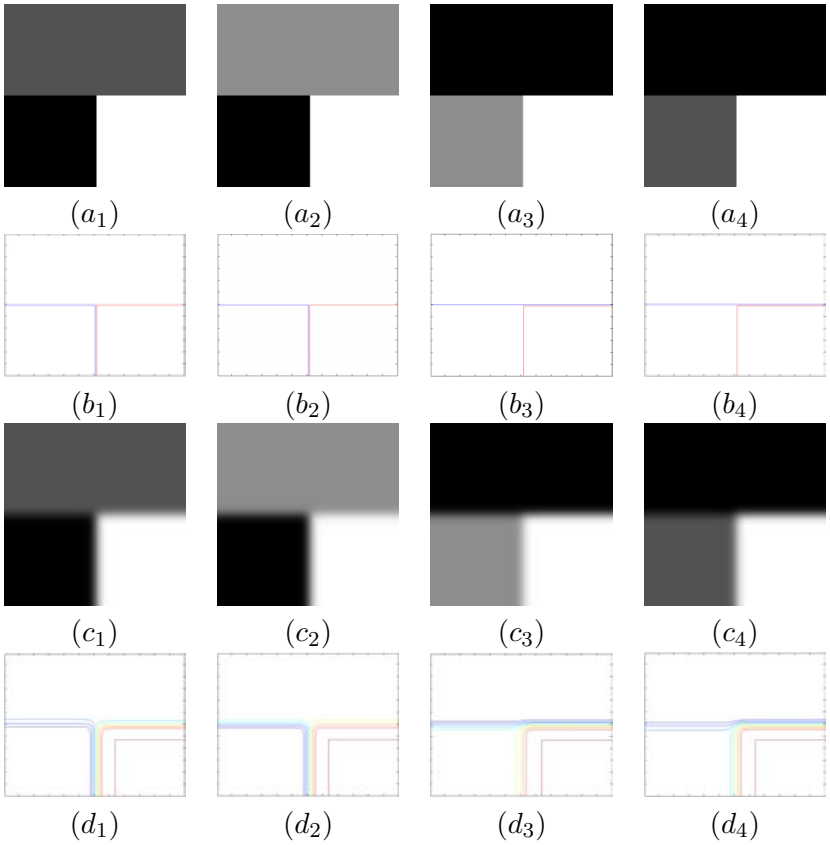


Figure 3.4: Illustration of the importance of contrast invariant smoothing for T-junctions. The images (a_1) and (a_2) as well as the images (a_3) and (a_4) differ by a monotone contrast change, while the images (a_1) and (a_3) as well as the images (a_2) and (a_4) differ by a nonmonotone contrast change. For each image (a_i), the image (b_i) shows its level lines, represented with different colors, the image c_i is the result of applying a linear scale-space (Gaussian) to the image (a_i) and the image (d_i) shows the level lines of the smoothed image c_i. If the linear scale space were contrast invariant, the evolution of figures that differ by a monotone contrast change would be the same. This is not the case since all four T-junctions give different evolutions under the Gaussian smoothing, since the evolution depends on the gray level values instead of their order.

Canny edge detector identifies edge points by finding the local maxima in the magnitude of the gradient vector. In the usual implementation of the Canny edge detector, these local maxima are found by suppressing all points where the magnitude of the gradient is not locally maximal in the direction of the gradient at the point in question. At some points close to the junctions, depending on the relative gray levels of the image, the gradient direction becomes rotated and the magnitude of the gradient vector is no longer maximal. As demonstrated by Li et al. [Li89], false non maxima suppression always occurs on the less contrasted edge of a T-junction. Therefore, such methods necessitate, after the edge detection, a subsequent following up of the edges to restore the junctions, leading to an inaccurate detection of the junction centers.

Another common limitation of all above mentioned methods is in that the Gaussian smoothing rules out the fundamental principle of Mathematical Morphology stated by Wertheimer in

Figure 3.5: Example of result obtained performing the morphological-based method proposed by Caselles et al. [Cas96a].

1923, that is *contrast invariance* [Wer23]. Intuitively, contrast invariance refers to invariance with respect to contrast changes, usually produced by illumination conditions and by the response of the sensor used to capture the image. In order to be robust, image analysis must be invariant with respect to contrast changes since they are not relevant for the human visual system. Formally, an image operator T is said to be contrast invariant if it commutes with all nondecreasing functions g, that is, if $g(T(u)) = T(g(u))$, where u is an image. The function g models the contrast changes. If g is strictly increasing, then the contrast invariant relation ensures that the filtered image $T(u) = g^{-1}(T(g(u)))$ does not depends on g. The significance of contrast invariance for smoothing T-junctions is illustrated in Figure 3.4.

To tackle the problems of convolution-based methods, Caselles et al. [Cas96a] proposed a morphological junction detector which localizes junction points before performing any smoothing. Their work formalizes in a computational-mathematical program several salient aspects of the phenomenological description of Gaetano Kanizsa [Kan96]. According to the Kanizsa's theory, visual perception tends to remain stable with respect to the basic operations of occlusion, transparency and contrast changes, by detecting junctions. To formalize such a theory, Caselles et al.[Cas96a] suggested that image analysis should start from the identification of basic mathematical objects, simple to handle, into which any image can be decomposed and from which it can be reconstructed and that are stable with respect to the three basic operations. They proposed as basic objects pieces of level lines joining junctions. According to this interpretation, their approach for detecting T-junctions relies on the use of level lines, which are curves directly provided by the image itself and that are invariant with respect to contrast changes. Beside the contrast invariance property, the interest of level lines for junction localization is that they are accurate at occlusions: they are in fact everywhere normal to the gradient as edges are, and thus object contours locally coincide with some isophotes. Contrary to previous approaches also based on level line analysis, junctions are computed before performing any smoothing. More precisely, junctions are detected as any meeting point of two level lines which is not the result of a quantization effect. This approach has the advantages of being contrast invariant and of allowing

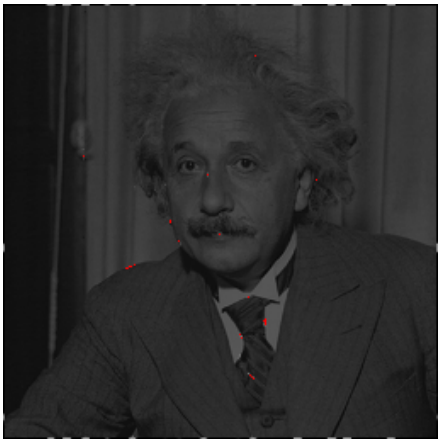


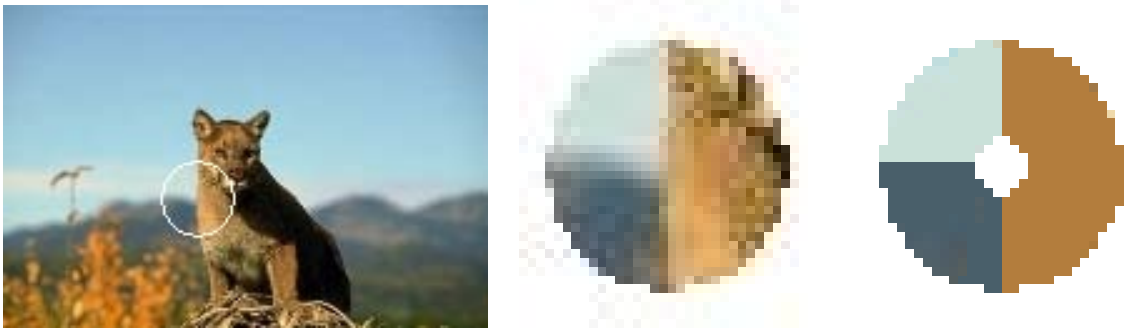
Figure 3.6: Example of result obtained performing the steerable filter-based method proposed by Freeman [Fre91].

a precise localization of junctions. However, it leads to over-detection along object boundaries. Indeed, the blur due to the image acquisition process generates, in correspondence of object contours, many parallel level lines, which collapse at many points generating tiny T-junctions (see Figure 3.5). In addition, this method does not address the problem of the branch extraction.

A different strategy to hypothesize junction centers is adopted by model-based template matching techniques [Bey89, Big94, For94, Hue71]. In this approach, the characterization of the junction is added to the localization criteria. These methods assume that a suitably small local neighborhood is sufficient to detect a junction. Detection window are typically larger with respect to detection window of convolution-based methods and their analysis is more involved. The basic idea is to fit a junction-model to the input signal in a neighborhood. This involves minimizing an energy function which gives the measure of the distance between the junction-model and the input signal.

In [Bey89, Big94, For94, Hue71] junctions are modeled as local structures with multiple intrinsic scales and orientations. The signature characterizing the junction is obtained by applying a filter at different scales and orientations. This has been achieved by the principle of steerability that enables the decomposition of a filter into a linear combination of bases functions [Fre91]. However, in order to achieve high resolution in orientation, a very large number of basis functions is needed. Consequently, the computational complexity is the main drawback of steerable-filter based methods. In addition, no characterization of junction types is attempted (see Figure 3.6).

In [Der94], junctions are modeled as a superposition of adjacent corners characterized by a large number of intrinsic parameters such as gray level intensities, position, orientation, as well as a parameter related to the blurring effect due to the image acquisition system. The corners forming a junction are constrained to have the same position and blur parameter. The



(a) (b) (c)

Figure 3.7: (a) An image where a T-junction is marked by a white circle. (b) Zoom on the circular neighborhood of the T-junction center. (c) Template model of the T-junction: a small neighborhood around the junction center is omitted and the regions delimited by the branches, the wedges, are approximated as constant luminance or constant color regions.

minimization strategy requires an initial good estimate of the different parameters to reduce the number of iterations needed to achieve the convergence as well as the probability of failing in a local minimum. The initial estimate is performed by minimizing an energy term related to the variance in gray level intensities within the considered region. The problem of this kind of modeling is that the minimization strategy suffers from lack of robustness since it usually achieves local minima.

Köthe [KÖ3] proposed a method based on the structure tensor [For86]. He demonstrated that standard methods for structure tensor calculation violate the Shannon's sampling theory causing aliasing and the loss of small features. To avoid aliasing, he represented the elements of the gradient tensor with half the sample distance of the original image. To prevent the loss of small features, Köthe proposed a non-linear averaging filter shaped as hour-glasses. Köthe also demonstrates that any positive semi-definite second order tensor can be decomposed into two parts, one encoding the edge strength and orientation and the other dealing with junction strength. Thus, the detected junctions and edges arise from a decomposition of the same original tensor representation which leads to much fewer errors in the junction detection results. No comparisons with previous methods was attempted.

During the last decade, hybrid methods retaining both the edge-based and the template-based approaches have received much attention. Typically, hybrid methods model junctions as piecewise constant regions called *wedges* emanating from a central point, omitting a smaller disk centered at this point (see Figure 3.7). The radial partition of the template is generally obtained by minimizing an energy function which measures the distance of the junction-model from the input signal. Candidate partitions are found by detecting edges and grouping them around the central point. The task is to find the minimum number of wedges that best describes the junction. The center identification is generally based on a local operator, while many different

minimization strategies and many different criteria to characterize edges have been proposed.

Parida and Geiger [Par98] formulated the junction detection problem as the one of finding the parameter values that best approximate the local data. Parameter values provide attributes of the detected junctions: center location, radius of the disk, intensity within each edge, number of radial line boundaries. An energy function is proposed which includes a gradient term in the junction model to find edges by a grouping mechanism. The local minima of the error are declared as junctions. As energy minimization strategy, they use dynamic programming that, from the point of view of the efficiency, represents the bottleneck of this method. The complexity of the search increases with the angular resolution, increasing also the quality of the results. However, this increase of complexity is already significant for modest resolution. Furthermore, results are good in term of number of estimated branches but they show inaccuracy in localizing the junction center as well as in estimating the branch orientations. Localization inaccuracy is mainly due to the use of the Forstener and Gulch's interesting point operator [For87] for detecting candidate points, whereas inaccuracy in estimating branch orientations is attributable to the use of the gradient information for grouping points to form edges.

Inspired by the junction detector of Parida and Geiger, better known as *Kona*, and aiming to develop a detector that is computationally more efficient, Cazorla and Escolano [Caz03] proposed a Bayesian framework for wedge identification. They analyzed two different approaches for finding the wedges: a region-based and an edge-based approach. The region-based method relies on a parametric junction model like the one proposed by Parida and Geiger. Under the assumption of piecewise constancy, wedges are mapped to homogeneous circular sectors. A circular sector is considered homogeneous when its intensity values are consistent with a Gaussian probability distribution. Junction detection is formulated as the problem of segmenting the intensity profile into circular sectors with homogeneous intensities. This is achieved by minimizing an energy function, which corresponds to the cost of coding, for each possible angle, all pixels of a given circular sector with the same probability distribution. The energy minimization is performed by the gradient descent algorithm [Avr03], which is susceptible to local minima. The intensity distributions are estimated inside each provisional circular sector, adjusting the provisional radial edges consequently. The edge-based method finds wedges by identifying edges emanating from the junction center. Edges are identified by measuring the response of edge filters and performing a statistical test, based on the logarithm of the ratio between the probability of being or not being an edge. Such distributions are estimated empirically by gathering and quantizing the frequencies of the filter responses in both cases. The edge-based method shows better performances than the region-based one in terms of strictly correct detections. Both methods are computationally more efficient and robust than the Kona method without loosing its reliability, but still suffers from inaccuracy due to the use of a local operator for center identification.

Another attempt to reduce the computational complexity of the Parida and Geiger method

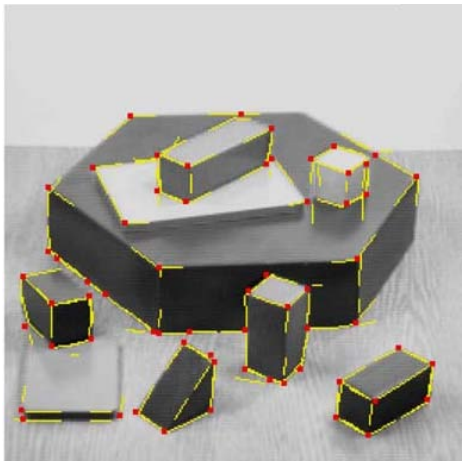
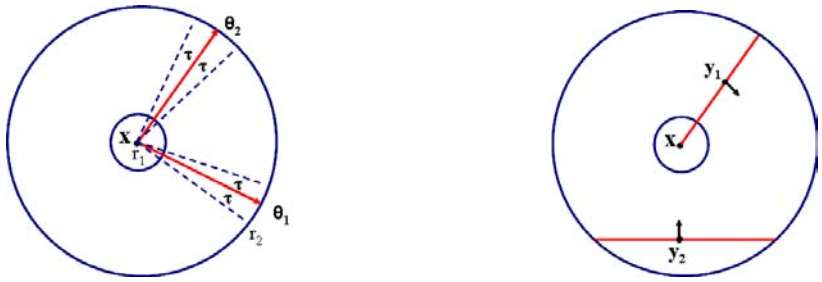


Figure 3.8: Example of result obtained performing the hybrid method proposed by Bergevin [Ber04].

has been made by Lagarniere et al. [Lag04]. The authors proposed a junction detector which operates in the gradient domain. In the gradient domain, junction centers correspond to points where image ridges (local maxima in the gradient direction) converge. The class of junction to be detected is restricted to these such that the junction branches can be approximated as straight lines in the vicinity of the point of convergence. These junction radial lines are found from the identification of the so called *circumferential anchor points*, that correspond to the extremities of potential radial lines for the hypothetical junction, located on the circumference of a circular neighborhood. Circumferential anchor points are detected as local directional maxima of the intensity gradient values located on the circumference of the circular neighborhood. The reduction of the computational cost is accomplished by operating on two binary edge maps: the first one, let say B is obtained by thresholding the gradient image, whereas the second one, let say B^+ contains the points of B that are ridges. Each point in B is considered as candidate junction point to be analyzed. Radial lines in a neighborhood of a given candidate point have to be such that the value of the gradient along all points of the radial lines is always greater than a threshold. A strength value is assigned to each radial lines and to each junction and it is used to perform a non-maxima suppression post processing phase to eliminate clusters of junctions. This method is computationally more efficient than the Kona method, but still suffer from localization inaccuracy.

The problem of a more precise localization has been addressed by Bergevin et al. [Ber04]. Candidate points to be characterized and validated are localized by applying the interest operator proposed by Rosenthaler et al. [Ros92]. This operator produces energy contour map, which yields a representation of strong 2D intensity variations, by applying a number of oriented filters to each image pixel. Based on the energy contour map, interesting regions are selected. Typically, a low threshold on the energy is applied to include all the true junctions in the selected interest regions.



(a) (b)

Figure 3.9: (a) Junction model used by the hybrid method in [Sin08]: (a) A local circular crown centered at P and having r_1 and r_2 as internal and external radius respectively. It is divided into wedges by radial edges emanating from P, whose orientation is identified by the angle θ_i that the edges form with the horizontal direction. (b) Points y_1 and y_2 all occurs in a neighborhood of point x , but only y_1 occurs along an edge that intersects x. Therefore, y_1 should provide a stronger edge measure than y_2.

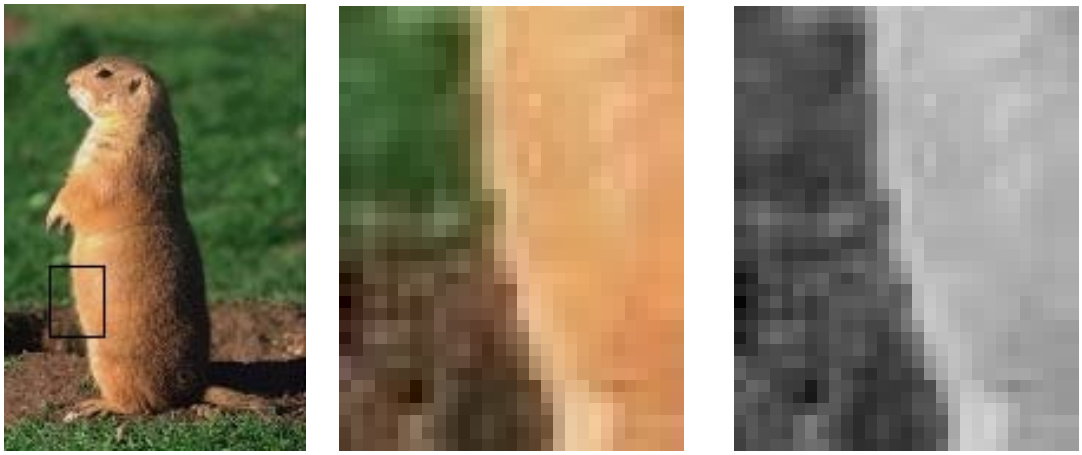
Then, the local edge operator proposed by Heitger [Hei95] is applied. This operator permits to extract edge orientations in a neighborhood of a candidate point by applying a number of oriented filters at each image pixel. The branch extraction is then performed by grouping edge points having similar energy and orientation and it is completed by approximating its branches as constant curvature primitives. The validation relies on topological criteria as well as on the density of the contour map. The main advantage of this approach is that all types of junctions are described and extracted uniformly using the same generic process. Also the localization accuracy is improved with respect to the Kona method (see Figure 3.8), but at the expenses of a very high computational complexity due to the use of oriented filters.

The issue of increasing the robustness of the Kona method, has been addressed by Ruzon and Tomasi [Ruz01] by using color information. The authors modeled the neighborhood of candidate point as a distribution of colors in the CIE-L*a*b color space [Wys82]. Image neighborhoods are represented by their color histogram. To obtain a more compact signature, quantization is performed. Computing the perceptual distance between the representation of adjacent image neighborhoods requires measuring the distance between two color signatures. This is a sub-problem of measuring the distance between probability density functions. To this goal the Earth Mover's Distance (EMD) [Rub98] is used. The maximum value of the EMD over all the minima is called *abnormality*. When the abnormality is high, it indicates a complete lack of symmetry in the image that usually corresponds to a junction. This method has shown an improvement with respect to earlier methods based only on luminance. However, this improvement is obtained at the expenses of a very high computational cost due to the quantization of each image window and the computation of the EMD many times.

Recently, a new hybrid-based approach has been proposed by Sinzinger [Sin08]. The algorithm involves three main steps. The first step consists in the identification, in a local neigh-

borhood of each image pixel, of the minimal set of radial edges emanating from the pixel under consideration. The identification is based on the minimization of a so called *angular energy* through a greedy technique. The angular energy includes a term called *edge strength*, which depends on two factors: the *pixel edge strength* and *angle edge strength*. The pixel edge strength depends not only on the image gradient at each pixel y of the potential edge, but also on the interaction of the image gradient to the potential edge xy, where x is the center of the circular neighborhood under consideration (see Figure 3.9 (a)). In fact, the contribution of the pixel y to the pixel edge strength should be strong only when the image gradient at x is aligned to the edge xy. This is the case for the pixel y_1 but not for the pixel y_2 (see Figure 3.9 (b)). The angle edge strength depends upon the pixel edge strength of all pixels in the region $R(\theta)$ defined by $R(\theta) = y(\zeta, r) : \zeta \in (\theta - \tau, \theta + \tau), r \in (r_1, r_2)$, where τ is an arbitrary threshold and r_1 and r_2 are respectively the internal and external radius of the circular crown centered at x (see Figure 3.9 (a)). The angle edge strength provides an approximate measure of how evenly distributed are the edge points in $U(\theta)$ are with respect to the edge identified by θ. The second step consists in determining the junctions by minimizing a second energy function, called *junction energy* that includes a term which models the wedges. Wedges are assumed to be relatively homogeneous and as a consequence they are modeled by the standard deviation within the wedge. The complete junction energy for a given point should have strong edge strength and small standard deviation for all wedges. Finally, the third step consists in discarding junctions which are not the result of the intersection of straight lines. While this method shows good results in term of recall, the precision is very limited specially in presence of textured regions.

In a very recent paper [Mai08] published by Maire et al., a new approach for junction detection has been proposed, which do not fit any of the three classes of methods analyzed above. The authors present an unified framework for contour and junction detection: junctions are detected as intersection of contours and contours themselves are detected by combining the use of local information derived from brightness, color and texture, with global information obtained from spectral partitioning [Chu97]. More precisely, contours are detected by using the method in [Mar86]. This method predicts the posterior probability of a boundary at each image pixel by measuring the difference in several features on the two halves of a disc of radius σ centered at the pixel under consideration. The features of brightness, color and texture are considered at three different scales and are combined linearly in a single multiscale signal. Global information is introduced by using an affinity matrix [Shi00]. The entries of the affinity matrix encode the similarity between pixels, given by the maximum value of the a posteriori probability of a boundary along a line connecting two pixels. Once contours have been detected, junctions are found by an Energy Minimization algorithm. The optimal junction locations are estimated by minimizing its distance from the contours, weighted by a function of the total contrast of the contour. This approach has shown leading results compared to the Harris contour detector [Har88] on the Berkeley Segmentation Dataset Benchmark [Fow]. However, in this kind of ap-



(a) (b) (c)

Figure 3.10: (a) An example of T-junction in a real image. (b) A zoom of the T-junction: looking at a local level it is more difficult to perceive the T-junction because of the blur in correspondence of object boundaries. (c) When looking at the gray level version of the image, it is even more difficult to perceive a T-junction since the gray levels of the regions delimited by the stem are very similar.

proach the detection of junctions follows the global image segmentation and therefore it is not compatible with the main objective of this Ph.D. dissertation, which conversely aims to exploit depth information provided by T- and X-junctions for segmentation purpose.

Summarizing, until recently, there have been three predominant approaches for junction detection: the convolution-based approach, the model-based approach and the hybrid approach. The hybrid methods are more reliable. However, they appear to have limitations in term of localization accuracy and robustness, specially in realistic situations including textured and high contour density regions.

In the following, we explore two different approaches for addressing the above mentioned limitations. The first approach detects junctions as intersection of line segments. Contrary to classical edge-based methods, line segments are computed by processes that use rather global information. The advantage of this strategy is in that it avoids the difficulty of dealing with the confusing intensity profile in the vicinity of a junction. Instead, the second approach aims to overcome this difficulty by the use of a region-merging strategy that, contrary to classical hybrid approaches founded on region-merging, employes a statistical modeling of regions at pixel level and a merging criterion based on information theory statistical measures.

3.1.2 Junction detection by intersection of line segments

From the previous section, it has emerged that the major difficulty of the junction detection problem in single images arises from the unreliability of the gradient profile near the junction

center. Indeed, the observed junctions correspond to depth or orientation discontinuities and are the result of a smoothing of the original photon flux introduced by the image acquisition system. This smoothing makes a mix of the gray levels belonging to surfaces having different depth or orientations, causing the failure of methods based on a 2D variation in the intensity signal (see Figure 3.10). In this section, we propose a method for junction detection which takes advantage of segments that are incident at a junction. The segments themselves are detected by processes that use rather global approaches. This strategy is also motivated by psychophysical experiments [McD04] suggesting that although some junctions in real images are locally defined and can be detected with simple mechanisms, a substantial fraction necessitates the use of more complex and global processes. Our method is built on the response of the Line-Segment Detector (LSD) proposed by Grompone et al. [Gio08], which computes segments as outliers of an unstructured background model in a linear time with respect to the image size. Junctions are then detected as intersection of segments and classified according to their geometric configuration. The proposed approach leads to a fast algorithm for junction detection which represents a good trade-off between qualitative results and computational cost.

In the following, the line segment detector used in this approach is presented and the algorithm for line segment-based junction detection is described.

3.1.2.1 Line-segment detector

LSD is a method for efficient and robust segment detection recently proposed by Grompone et al. [Gio08]. The perception of segments is related to the grouping law of constancy of direction (alignment), which is a special case of continuity of direction. LSD puts together two well known state of the art algorithms for segment detection: Burn's segment detector [Bur86] and the meaningful segment detector developed by Desolneux et al. [Des00]. The objective is to keep the advantages of the "a contrario" framework proposed by Desolneux improving its accuracy and efficiency. The Desolneux et al. "a contrario" framework is based on a mathematical translation of the Helmholtz's principle [Hel25], which is as follows:

Helmholtz principle: *Gestalts are sets of points whose (geometric regular) spatial arrangement could not occur in noise.*

Informally, the Helmholtz principle states that there is no perception of conspicuous structures in white noise, whose samples are identically distributed independent random variables. This means that geometric structures such as segments cannot be perceived in a white noise image. Therefore, if segments are perceived, it is because they represent outliers with respect to the noise. Contrarily to classical methods that find a statistical model for the object to be detected, a contrario methods act by finding target objects as outliers of a non-structured background model represented by a white noise image. The main advantage of this strategy is that

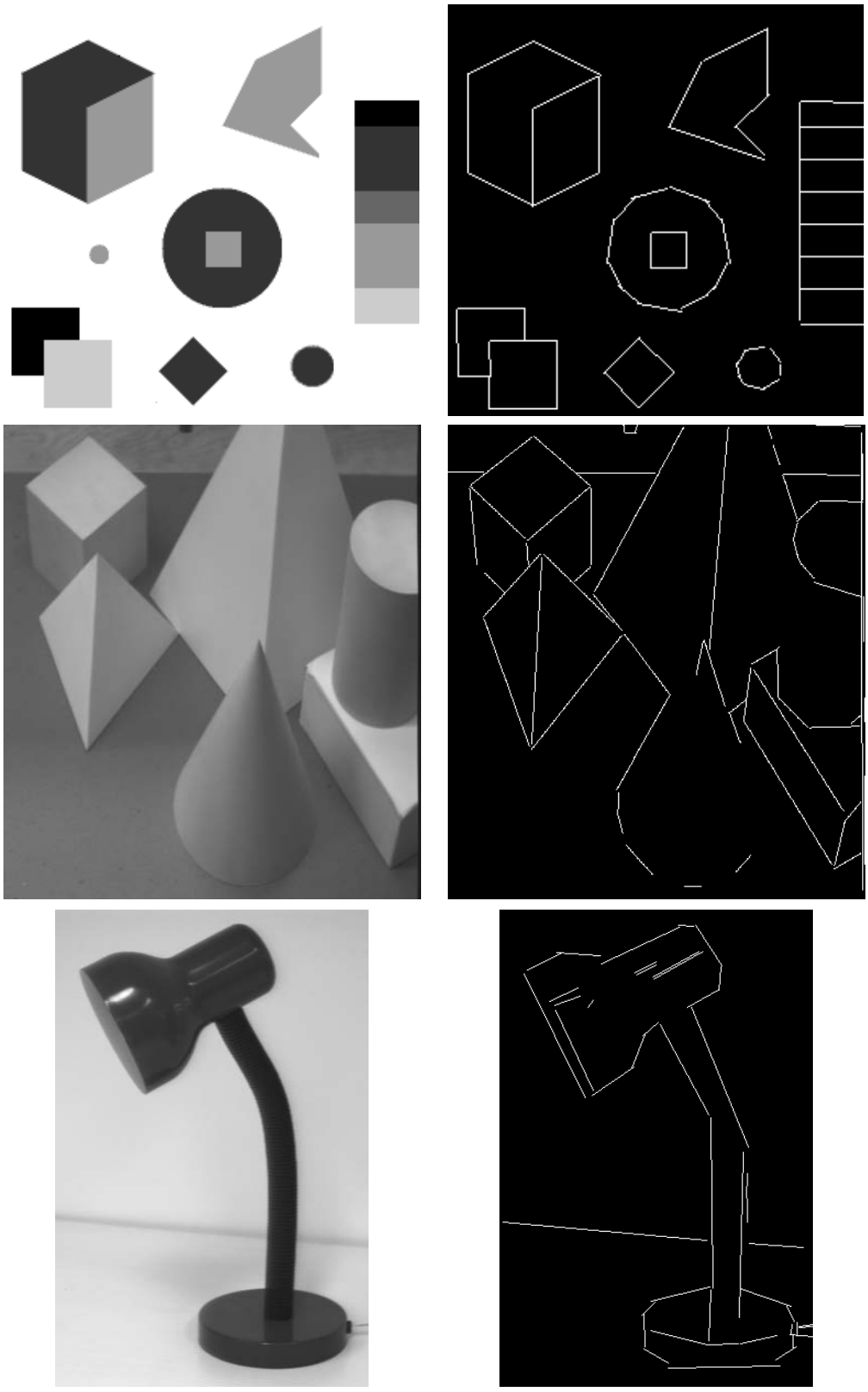
the thresholds of the detection algorithm can be defined in order to control its expected number of false detection under the background model. Although the segment detector of Desolneux definitely tackles the problem of false positive and false negative detections, it appears to have a limitation in terms of detection accuracy. This is due to the fact that it looks for sets of aligned points on the whole image. Grompone et al. solved this problem by using Burn's strategy for partitioning the image into line-support regions, which correspond to a group of connected pixels that share the same gradient angle up to a certain tolerance. The medium orientation is computed for each support regions, leading to a more accurate result. In addition, the use of a previous line support-region detection step speeds up the computation leading to a line segment detector able to process images in linear time relative to the number of pixels. Another advantage of this approach is that curves are approximated by segments of different lengths and orientations, since the size of the line support regions varies with the orientation. In Figures 3.11 and 3.12, there are some examples of processing by the LSD. As can be observed, segments give a good description in term of the geometrical structures of the image. The above detailed characteristics of this algorithm make it highly suitable for junction detection purposes. In the next subsection we explain how we use line segments detected by LSD to localize and classify junctions.

3.1.2.2 Line segment-based junction detector

Beside the advantages stressed in the previous section, the main interest of the LSD for junction detection purpose is in that it leads to an easy visualization of junctions, even if the junction center is often missing for the detection. In fact, in this case the visualization of junctions is the result of an interpolation process driven by the good continuation principle [Mon71]. Straight lines are extended and junctions are detected as intersection of straight lines, avoiding the problems of contour detection in the neighborhood of a junction. According to the number and the orientation of intersection segments, junction points are classified. A tolerance is used on the segment extremity position and the orientation of the junction branches (see Figure 3.13).

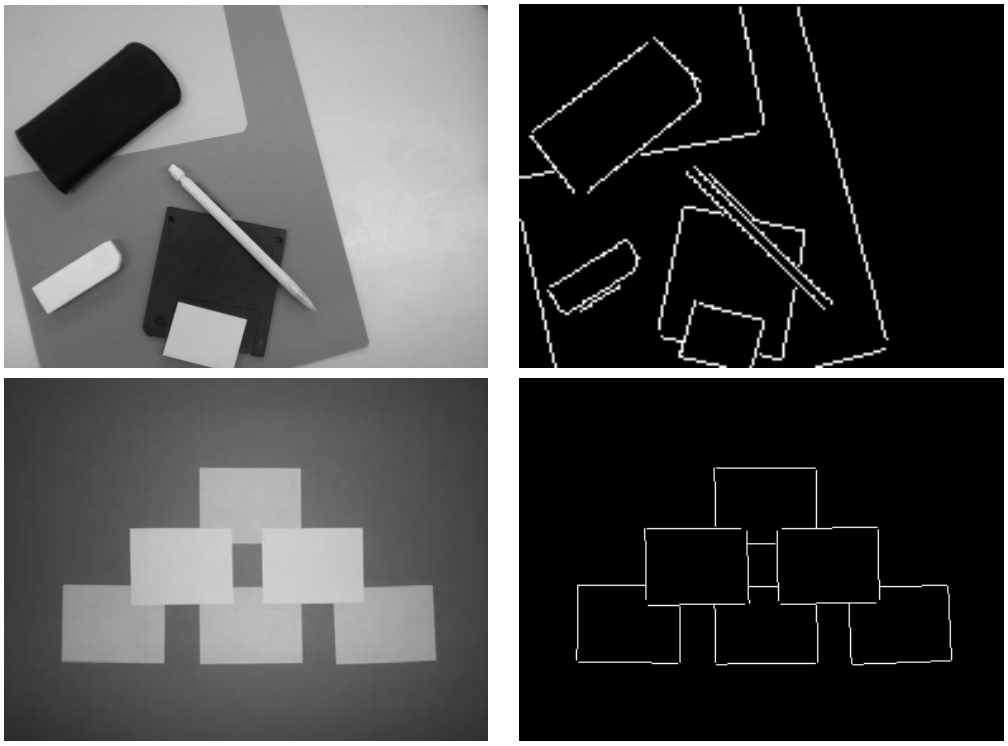
More precisely, the intersection of two segments may lead to an X-junction, a T-junction or an L-junction (see Figures 3.15(a), (b) and (c) respectively), depending on the distance of the tips ouside the tolerance neighborhood Ω with respect to the intersection point P. In the following, the condition on the tolerance on the center location is assumed to be satisfied. By this we mean that when the segments do not intersect, the extremities of the segments closer to the hypothetical intersection point fall in the tolerance neighborhood. Under this assumption, junctions are classified depending on the segment orientation as well as on the distance of the tips ouside the tolerance neighborhood with respect to the intersection point P.

In the case of two segments, when all four tips outside Ω are at a sufficiently large distance



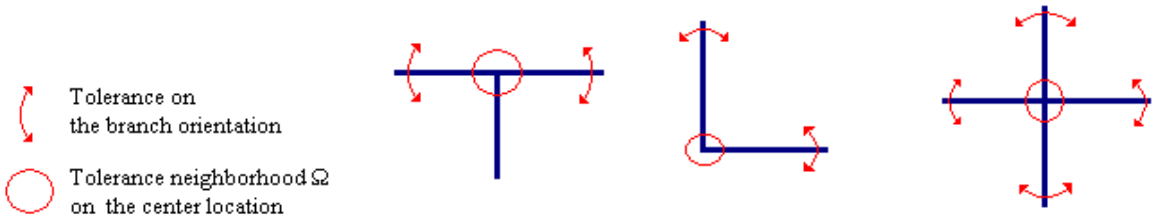
(a) (b)

Figure 3.11: Example of processing by LSD: (a) Images. (b) Segments detected by LSD.



(a) (b)

Figure 3.12: Example of processing by LSD: (a) Images. (b) Segments detected by LSD.



(a) (b) (c)

Figure 3.13: (a) T-junctions. (b) L-junctions. (c) X-junctions.

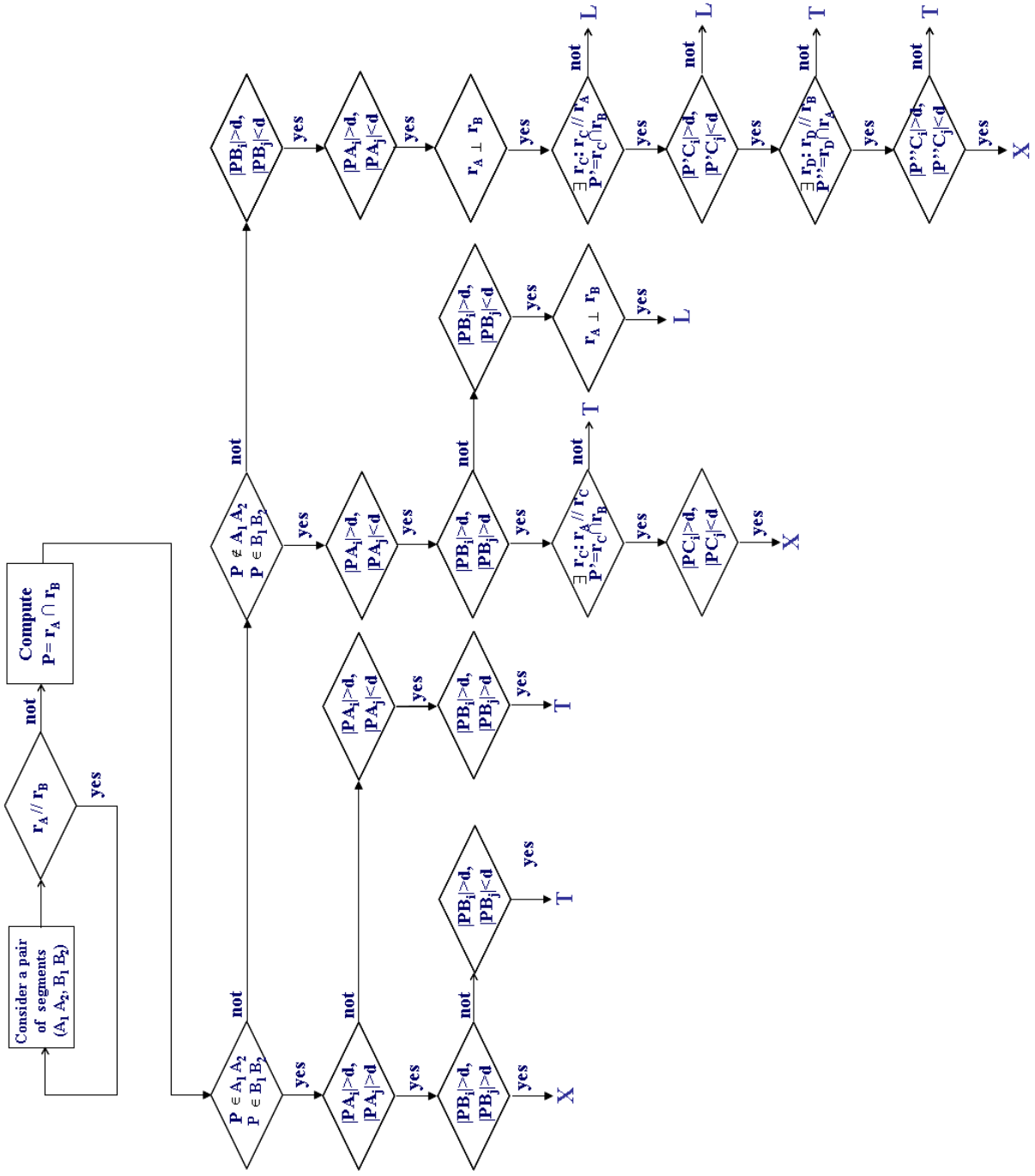


Figure 3.14: Block diagram of the T-junction detection algorithm by intersection of segments. A_1A_2 , B_1B_2, and C_1C_2 are the segments delimited respectively by the pair of points (A_1, A_2), (B_1, B_2), and (C_1, C_2). r_A, r_B, and r_C are the lines the segments, respectively, A_1A_2, B_1B_2, and C_1C_2 lie on. $|PA_1|$ is the euclidean distance between the points P and A_1. The relations of parallelism ($//$) and perpendicularity ($\perp$) are computed with precision $\frac{\pi}{n}$, where $n = 16$.

from P, they convey the perception of an X-junction (see Figure 3.15(a)). Instead, if this number is three or two, the intersection of two segments leads to the perception of a T-junction (see Figure 3.15(b)) or an L-junction (see Figure 3.15(c)).

The intersection at point P of three segments s_1, s_2, and s_3 may lead to the perception of an X-junction, a T-junction or an Y-junction (see Figures 3.15(d), (e) and (f) respectively). When the number of tips outside Ω is four (three) and all them are at a sufficiently large distance from P, if two of them, say s_1 and s_2, are aligned, an X-junction (T-junction) is perceived (see Figures 3.15(d), (e)). Otherwise, if the number of tips outside Ω having a sufficiently large distance from P is three and there is no pair of aligned segments, an Y-junction is perceived (see Figure 3.15 (f)).

The intersection of four segments at a point P leads to see an X-junction when the number of tips outside Ω is four and the segments are two by two aligned (see Figure 3.15(g)). Figure 3.14 summarizes the junction detection and classification algorithm through a block diagram. In Figures 3.16 and 3.17 some examples of results obtained by using the proposed method are shown. T-junctions are marked with pink circles around the junction centers and visualized as vectors that emanate from the junction centers and point to the region closer to the viewpoint. Instead, L-junctions are marked with pink circles and Y-junctions with a pink rectangles. As can be observed, the line segment based junction detector gives satisfactory visual results.

In the next section, we proposes an alternative method for detecting T-junctions. A comparative evaluation of the two proposed strategies in terms of T-junctions detection, will be given in section 3.1.3.9.

3.1.3 T-junction detection by region merging

In this section, we propose an algorithm for T-junction detection based on region-merging. Contrary to classical hybrid approaches founded on region-merging, the merging process we propose here is a tree-based segmentation inspired from the work of Salembier and Garrido [Sal00]. The authors proposed an image tree representation, called Binary Partition Tree (BPT), consisting of a structured representation of a set of hierarchical partitions in which the finest level of detail is given by the initial partition of image pixels. The nodes of the tree are associated to regions that represent the union of two children regions and the root node represents the entire image support. Starting from the initial partition of all image pixels, pairs of neighboring regions are iteratively merged until a termination criterion (usually the number of regions of the final partition) is reached. The order in which neighboring regions are merged depends on a similarity measure between the region models. In this seminal work, the regions are modeled deterministically by their mean color value and the order in which regions are merged is determined by a similarity measure based on color difference between the region models. Relying on

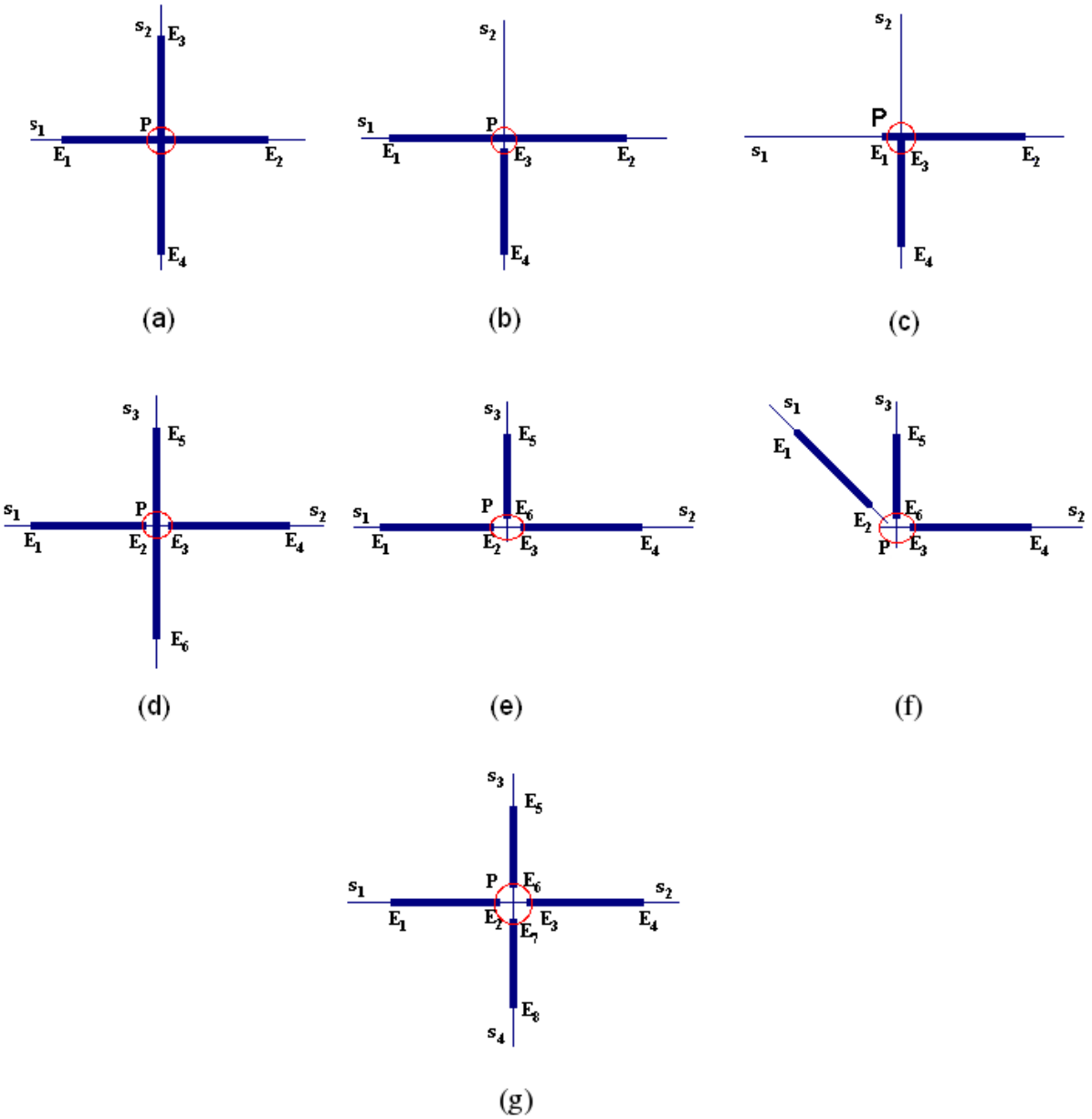


Figure 3.15: Possible configurations of segments conveying the perception of a X-junction ((a),(d),(g)), L-junction (c), T-junction ((b),(e)) and Y-junction (f).

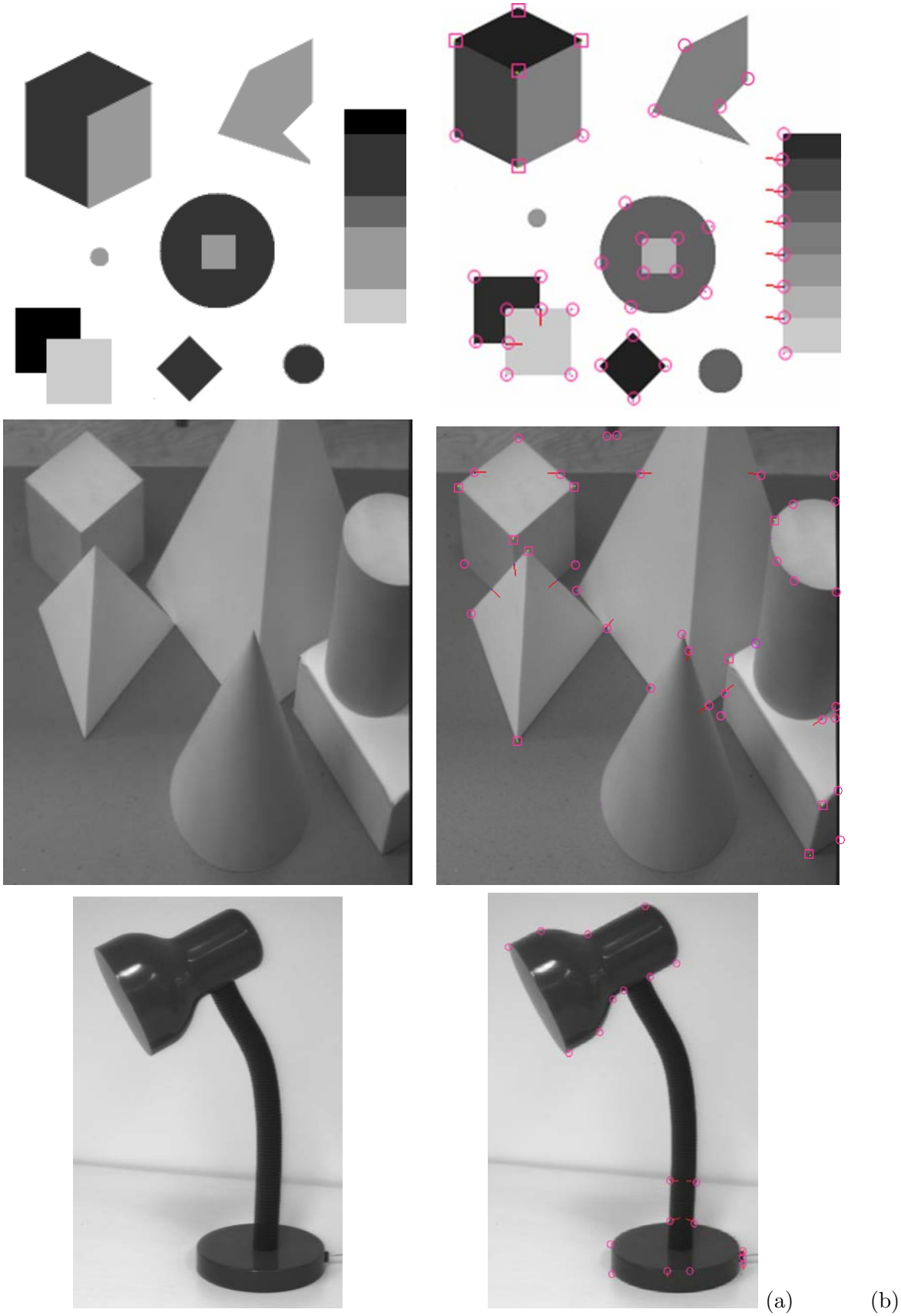
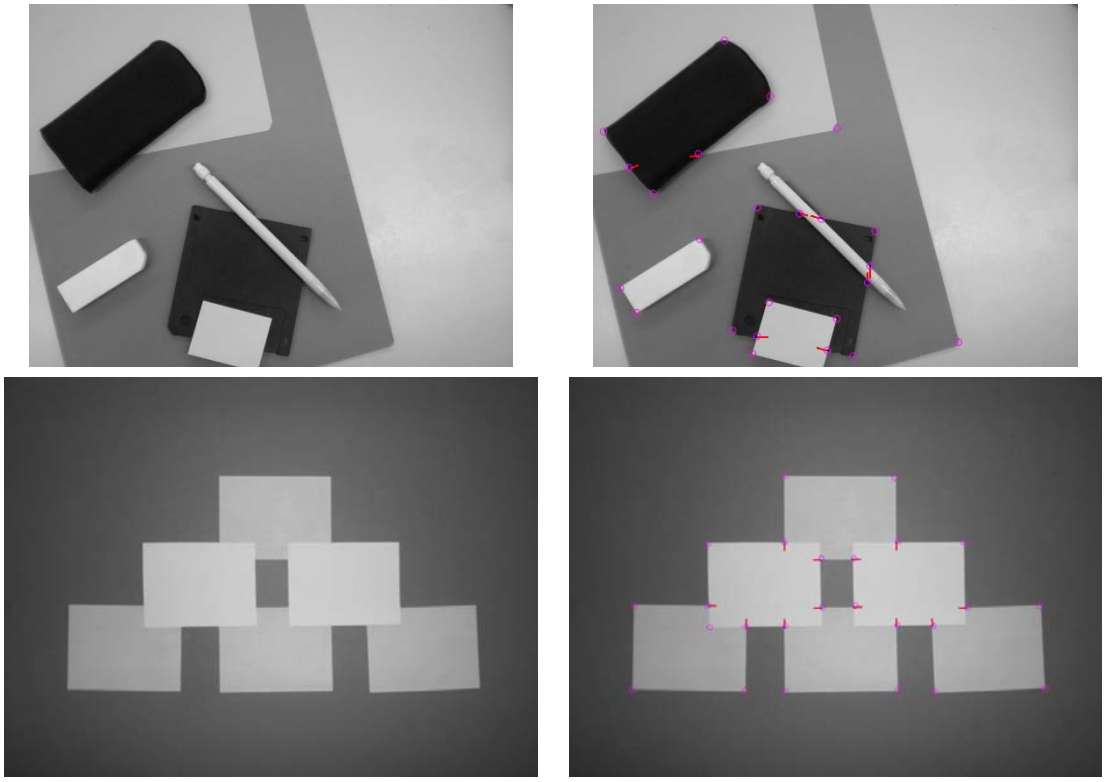


Figure 3.16: (a) Original images. (b) Results of applying the segment-based junction detector.



(a) (b)

Figure 3.17: (a) Original images. (b) Results of applying the segment-based junction detector.

the BPT representation, Calderero and Marques [Cal08], have recently presented a new region model based on color histogram and a new family of statistical similarity measures between the region models based on information theory. This kind of modeling has demonstrated a noticeable improvement with respect to first order statistical models where mean or median color values are used as region model since they do not assume that regions are homogeneous in color nor texture. However, the merging process starts by considering that each pixel is a single region, which is modeled still deterministically by its color value and therefore the effect of the statistical modeling become really important only in the late stages of the merging process. In the context of T-junction detection, since the region to be segmented is a small neighborhood of a given candidate point, the importance of a good modeling in early stages of the merging process becomes crucial. We propose to solve this problem by modeling each pixel statistically by a probability distribution, instead of deterministically by its color value. The probability distribution of a given pixel is obtained by exploiting non-local self-similar structures, which can be detected by patch comparisons. By *self-similarity*, we mean that every small patch in a natural image has many similar patches in the same image. As can be appreciated in Figure 3.18, most objects in the real world have a self-similar or periodic structure: different parts of the same object show the same statistical properties at many different locations. The fact that natural images have such self-similarity property is a kind of stationarity assumption, actually more general and more accurate that any existing image statistics since it does not rely on a subjacent model but directly on the data itself. This assumption has been proved to be sound by the works of Efros and Leung [Efr99], and Levinia [Lev02] and it has been successfully used in the seminal work of Efros [Efr99] for texture synthesis and then in [Bua06] and in [Bua08] for image and video denoising. To the best of our knowledge, it is the first time that self-similarity is exploited for segmentation purpose. Neverthless, the idea of modeling a single pixel statistically by a probability distribution in the context of image segmentation is not a novelty. It has also been very recently proposed by Chan et al.[Cha08, Ni09], who modeled each pixel by a normalized histogram of the pixel intensities in a neighborhood of the pixel.

The algorithm we have developed involves three main steps. A first selection step provides candidate points that represent potential T-junctions to be characterized and possibly validated in a second step. The characterization, namely the branch extraction, is performed on a close surrounding of candidate points, omitting a small neighborhood centered at them. The obtained branches are then propagated inside the omitted domain according to the "good continuation principle" [Mon71] and constrained to meet at the candidate point. This procedure is also supported by psychophysical experiments [Wue96] suggesting that junctions are detected even when the center is occluded. To each validated T-junction, a graduate measure of *junction likelihood* is assigned relying on the regularity of its branches. This measure is used in the third step, devoted to the reduction of clusters of validated points.

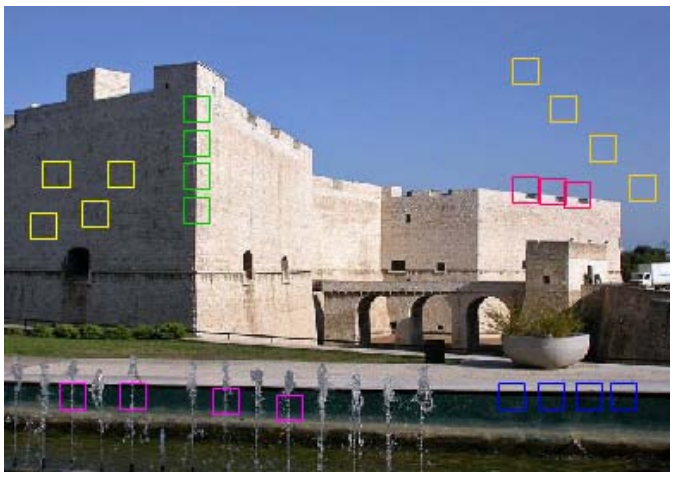


Figure 3.18: Figure illustrating the phenomenon of non-local self-similarity. Similar patches are delimited by windows of the same color.

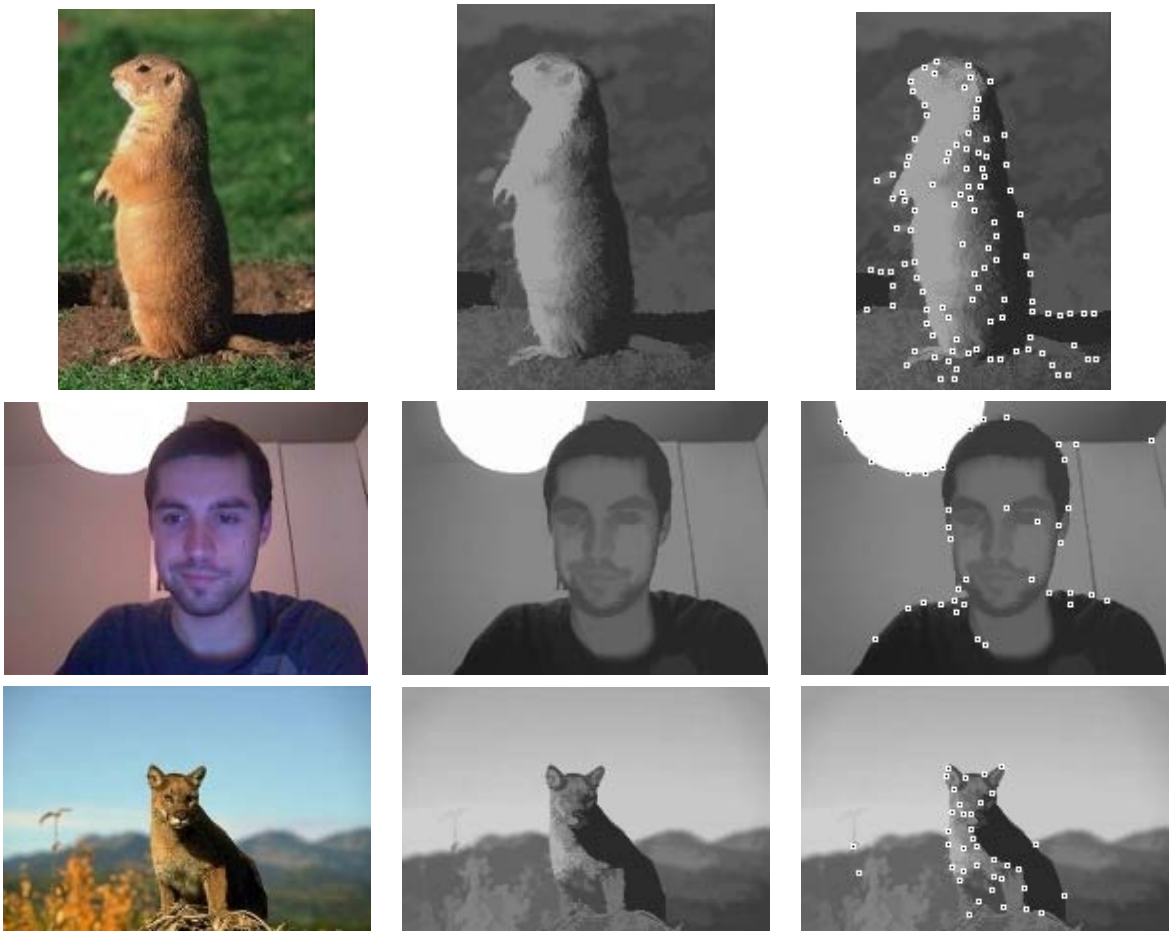
In the following, the working principle and the algorithmic implementation of each step are discussed and detailed.

3.1.3.1 Candidate points selection

Let us consider the classical discrete representation of the image on a grid A of size $M \times N$. The discrete image $U : \Gamma \to \mathcal{Z}$ is given by the union of the squares centered at the position (i, j), where $0 \leq i \leq N, 0 \leq j \leq M$. The brightness in each square, called pixel, is constant and equal $u(i, j)$.

Since T-junctions are structural features, the search for candidate points is performed on the structural part of the image, also called *cartoon component*, obtained by a simplification of the original image with a hierarchy of leveling [Mar07], based on Gaussian scale-space markers. More precisely, at each scale k the "cartoon component" U_k of the image U is obtained as $U_k = \Lambda(M_k|U_{k-1})$, where $M_k = U * G_{\delta k}$ is the marker obtained by convolution of U with a Gaussian kernel G of standard deviation δk, U_{k-1} is the reference image with $U_0 = U$ and Λ is a leveling [Mey97]. In all experiments, we used $\delta k = 3$ and $k \in \{1, 2, 3\}$.

The selection of candidate points relies on the observation that T-junctions can be thought as a superposition of two adjacent corners. As a consequence, corners are good candidates points for T-junctions. Corners are localized by SUSAN [Smi97], a local and fast nonlinear filter relying on an homogeneity principle. SUSAN stands for Smallest Univalue Segment Assimilating Nucleous. Rather than evaluating local gradients, which might be noise-sensitive and computationally more expensive, this method takes a different approach. At each point (i, j), a circular neighborhood of the fixed radius of three pixels around it is considered. The center pixel is referred to as



(a) (b) (c)

Figure 3.19: Examples of candidate point selection: (a) Original images. (b) Result of applying a hierarchy of leveling. (c) Result of applying the SUSAN filter to the images in (b): candidate points are marked in black and are surrounded by a white square.

nucleus and its intensity value is used as reference. All other pixels in this circular neighborhood are partitioned into two categories depending on whether they have similar intensity value as the nucleus or different. In this way, a local area of similar brightness, called USAN, is associated to each image point. The relative size of the USAN contains information about the structure of the image at that point. Near edges, this ratio drops to 50 and near corners it decreases further about 25. Hence corners can be detected as points where the number of pixels with similar intensity value in a local neighborhood reaches a local minimum and is below a predefined threshold (a good value is 20). Local minima of the SUSAN (Smallest USAN) are then selected from the remaining candidates.

To take into account the localization inaccuracy of the SUSAN filter, coordinates of candidate points are allowed to vary on a small neighborhood. In practice, we apply a dilation with a square structuring element of size 5×5 on the mask of candidate points obtained by applying SUSAN. The mask resulting from the dilation defines the set of candidate points and the branch extraction is performed for each candidate point.

Although the proposed strategy selects corners as candidate points, it allows a fast and significant reduction of the number of image pixels to be processed while keeping all real T-junctions. Figure 3.19 shows some results of the candidate point selection step. As can be observed, the ratio between the number of selected candidate points and the number of image pixels is very small in all shown examples. However, in the case of very textured regions such as near the paw of the lynx or in the image on the first row, this ratio increases.

In the rest of section 3.1.3 we shall use for the discrete image representation a grid with integer coordinates. Every image pixel is associated to a point x with integer coordinates (i, j). The branches of a T-junction correspond to lines passing between image pixels and therefore the T-junction center does not have integer coordinates. For each candidate point x with coordinates (i, j), we consider a squared neighborhood W of size $w \times w$ centered at the right down point x_{dc} with coordinates $(i + \frac{1}{2}, j + \frac{1}{2})$, where the branch extraction is restricted to and a 4×4 squared neighborhood Ω of x_{dc}, where the photometrical profile is considered unreliable (see Figure 3.20). Our strategy consists in first extracting the branches in $(W - \Omega)$ and then in extending them in Ω until the candidate point is reached. The next section details how to perform the branch extraction in $(W - \Omega)$.

3.1.3.2 Branch extraction in $(W - \Omega)$

The branch extraction in $(W - \Omega)$ is performed on the original image by a tree-based statistical region-merging algorithm. For clarity of exposition, a detailed description of the tree construction process has been postposted to section 4.2, where this algorithm will be used for image segmentation purpose. Instead, this section focuses on the pixel modeling, which is a key ele-

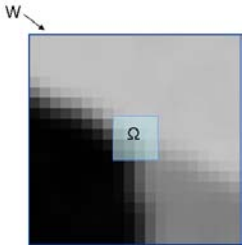


Figure 3.20: A window W centered at a candidate point: Ω is the neighborhood we consider unreliable.

ment in the performances of the branch extraction step by region-merging. As anticipated above, the key assumption behind the pixel modeling is that the image is a fairly general stationary random process. For images, the stationarity condition means that as the size of the image grows, for every small patch in an image, it is possible to find many similar patches in the same image. Under the stationarity assumption, the probability distribution of a single pixel can be computed as follows. Let U be an image, x a pixel of the image domain and let $\mathcal{N}(x)$ be a square image patch centered at x which does not include x. Let us assume that the probability distribution of $U(x)$ depends only on the values of the pixels in $\mathcal{N}(x)$ and it is independent of the rest of the image (markovian model). Then, the probability distribution of $U(x)$ given the pixel values of its neighborhood $\mathcal{N}(x)$, can be estimated by computing the set:

$$\Gamma(x) = \{y : \frac{d(\mathcal{N}(x),\mathcal{N}(y))}{d(\mathcal{N}(x),\mathcal{N}_{best})\}} < (1+\epsilon)\},$$

where $d(\mathcal{N}(x),\mathcal{N}(y))$ is a distance between a patch centered at x and a patch centered at another pixel y of the image domain, $\mathcal{N}_{best}$ is the patch that gives the best patch match and ϵ is a small constant. The histogram of all pixel values in $\Gamma(x)$ gives an estimation of the probability distribution of the value of x given the values of its neighborhood $\mathcal{N}(x)$ [Efr99]. More precisely, assuming that the pixel values range from 1 to L, the histogram is constructed by adding one to the value $U(y)$ for each $y \in \Gamma(x)$ and then normalizing so that the histogram integral is equal to one. However, setting a hard threshold $(1+\epsilon)$ to defines the set $\Gamma(x)$ leads to be able to estimate the probability distribution only of pixels for which similar patches can be found. Indeed, if it is not the case, the hard threshold strategy would leave the set $\Gamma(x)$ empty. To overcome such a problem, Buades et al. [Bua05] proposed to use an exponential function, which allows a more continuous distribution. More precisely, the probability distribution of a pixel x conditioned to its neighborhood $\mathcal{N}(x)$, can be computed by computing for each pixel y the quantity:

$$w(x,y) = \frac{1}{Z(x)} e^{\frac{-d(\mathcal{N}(x),\mathcal{N}(y))}{h}}, \tag{3.2}$$

where $Z(x)$ is the normalizing factor:

$$Z(x) = \sum_y e^{\frac{-d(\mathcal{N}(x),\mathcal{N}(y))}{h}}, \tag{3.3}$$

and

$$d(\mathcal{N}(x),\mathcal{N}(y)) = \sum_z \frac{(U(x-z) - U(y-z))^2}{K(z)}, \tag{3.4}$$

is the similarity between pixel values of a patch centered at x and a patch centered at y. The variable z indicates the displacement on the patch with respect to x, and $1/K$ is a Gaussian-like function decaying from the center of the patch to its boundary. More precisely, $K(z) = (2 * \|z\| + 1)^2$ acts as a weight function of the euclidean distance between two patches. The goal of the function K is to give more importance on the patch to pixels closer to the reference pixel. Indeed, since we would like to compare local structure as accurately as possible, the error for nearby pixels should be greater than that of distant pixels. The parameter h controls the decay of the exponential function, and therefore of the function w. Due to the fast decay of the exponential term, large euclidean distances lead to nearly zero weights acting as an automatic threshold. To reduce the computation cost, the search for similar patches is restricted to a search window of size $S \times S$. The histogram corresponding to the probability distribution of x is obtained by adding, for each pixel y the value of the function $w(x, y)$ to $U(y)$.

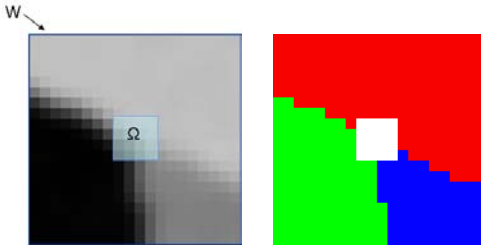
Summarizing, the function $w(x, y)$ associated to the pixel x, depends on the similarity $d(\mathcal{N}(x),\mathcal{N}(y))$ between the corresponding pixel values of a patch centered at x and a patch centered at $y \in S \times S$ and satisfies the conditions $0 < w(x, y) < 1$ and $\sum_{y \in S \times S} w(x, y) = 1$. For every pixel x, $w(x, y)$ converges to the conditional expectation of x given its neighborhood $\mathcal{N}(x)$. Indeed, as demonstrated in [Bua05], $w(x, y)$ can be seen as an instance for the exponential operator of the Naradaya-Watson estimator [Wat64, Nar64], which estimates the conditional expectation of a random variable.

Once pixels, which are considered as initial regions have been modeled by their pdf, the region merging algorithm starts to iteratively merge pairs of neighboring regions in $W - \Omega$ following the statistical *Kullback-Leibler merging order* (KL) [Cal08], until three regions are obtained (Figure 3.21). The KL merging order gives a measure of the probability that two regions have been generated by the same statistical distribution. Indeed, it is based on measuring the similarity between the empirical distributions of two regions and the empirical distribution of their merging, weighted by the area of the regions. Formally, it is stated as:

$$KL_{area}(R_i, R_j) = -n_i \cdot D_{KL}(P_i \parallel P_{i \cup j}) - n_j \cdot D_{KL}(P_j \parallel P_{i \cup j}), \tag{3.5}$$

where R_i and R_j are two adjacent regions with size n_i and n_j and empirical distribution P_i and P_j respectively, whose union would generate a new region, $i \cup j$, with empirical distribution

$$P_{i \cup j} = \frac{n_i}{n_i + n_j} P_i + \frac{n_i}{n_i + n_j} P_j \tag{3.6}$$



(a) (b)

Figure 3.21: (a) Image to be segmented. (b) Branch extraction in $W - \Omega$.

and D_{KL} is the Kullback-Leibler divergence operator (not symmetric version) between two statistical distributions [Kul51], given by:

$$D_{KL}(P_i \parallel P_{i\cup j}) = P_i \log \frac{P_i}{P_{i\cup j}}. \tag{3.7}$$

More precisely, P_i is the color histogram in the color space (YUV) and $D_{KL}(P_i \parallel P_{i\cup j})$ is given by

$D_{KL}(P_i \parallel P_{i\cup j}) = \alpha \cdot D_{KL}(P_{Yi} \parallel P_{Yi\cup Yj}) + (1-\alpha) \cdot (D_{KL}(P_{Ui} \parallel P_{Ui\cup Uj}) + D_{KL}(P_{Vi} \parallel P_{Vi\cup Vj}))$. We set $\alpha = \frac{1}{2}$.

3.1.3.3 Setting the parameters for the local segmentation

The above described region merging algorithm involves a set of parameters that need to be fixed:

- *Size w of the local neighborhood W*: this parameter addresses the scale issue. The right scale depends on both the image resolution and the viewing distance.
- *Size n of the similarity window (patch size)*: this parameter has to be as small as possible to take care of the image details and fine structure, being at the same time robust to noise. Therefore, its value should increase with the amount of noise in the image.
- *Size S of the search window*: theoretically this parameter should be the same as the image size, but in practice a search window of five times the size of the similarity window guarantees good results with a reduced computational cost.
- *Value h of the filtering parameter*: this parameter controls the decay of the exponential functions. When the standard deviation of the noise is known, the value of h should depend on it [Bua06]. For a small h, the similarity function would not be robust to noise since very similar neighborhood could give small value of the similarity. Nevertheless, by increasing the value of h, very different neighborhood could give large value of the similarity.

- *Number of bins used to construct the histogram*: the number of bins determines how accurately the probability distribution is represented. The highest the number of used bins is, the more accurate would be the representation. However, by increasing the number of bins, the estimation of the divergence between histograms becomes inaccurate since there might be a very small number of samples per bin. In all experiments, we have fixed the ratio between the number of bins used for the luminance component (Y) and the number of bins used for the chroma components (U and V) to 3:1:1 in the YUV color space.

In order to fix these parameters we proceeded as follows. We considered two values of the similarity window: 3×3 (see Figures 3.22, 3.24 and 3.26) and 5×5 (see Figures 3.23, 3.25 and 3.27). For each value of the similarity window, we ranged the test values of the filtering parameter h between 30 and 110 with intervals of 20 and the values of the number of bins of the luminance component between 5 and 250 with intervals of 50. As can be observed, a smaller similarity window gives in general more accurate segmentation results. For a similarity window of 3×3 and a search window of size 15×15, good values for the filtering parameter and the number of bins for the luminance component are respectively 70 and 150.

3.1.3.4 Candidate point validation in $W - \Omega$ before branch extraction in Ω

The validation of candidate points is based on the topology of the branches, as well as on the geometrical and photometrical profile of wedges. The region merging strategy does not guarantee neither that the three final regions will reach Ω (see Figure 3.28 (a)) nor that all the regions will intersect the boundary of W (see Figure 3.28 (b)) with at least a minimum number of pixels equal to half the window length. In both cases the candidate point is discarded. To guarantee the visibility of each wedge, we impose a threshold on the minimum gray level difference between the mean gray level of each pair of wedges and on the minimum color difference between the mean color of each pair of wedges. If the minimum gray level or color difference is below a given threshold (t_{gray} and t_{color} respectively), the point is discarded (see Figure 3.28 (c)). In most cases corresponding to object boundaries or textured regions, one wedge is composed of a very small number of pixels or looks like a narrow band. We then use a "size criterion" that is as follows: if at least one region completely disappears after an erosion (binary) with a square structuring element, then the candidate point is discarded (see Figure 3.28 (d)). The size s of the square structuring element is related to the length w of W. To keep T-junctions whose contours converging at the junction center form a small angle, s is taken as: $\frac{w}{6}$.

3.1.3.5 Branch extraction in Ω

As explained above, the photometrical profile is not reliable in Ω, and thus the use of a region merging algorithm would be misleading. Instead, the extrapolation of the branches inside Ω is

Figure 3.22: Size of the similarity window of 3×3.

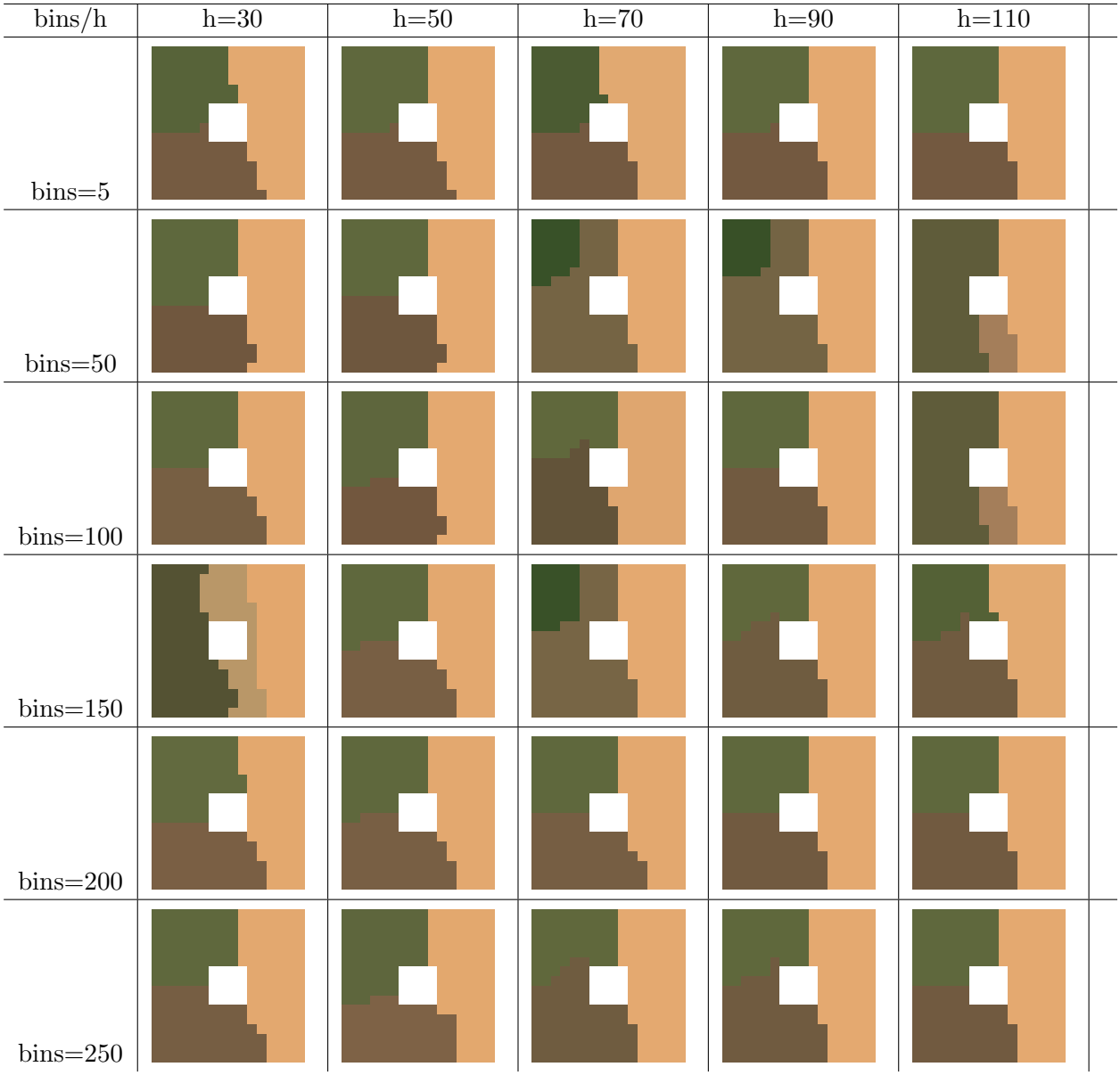


Figure 3.23: Size of the similarity window of 5×5.

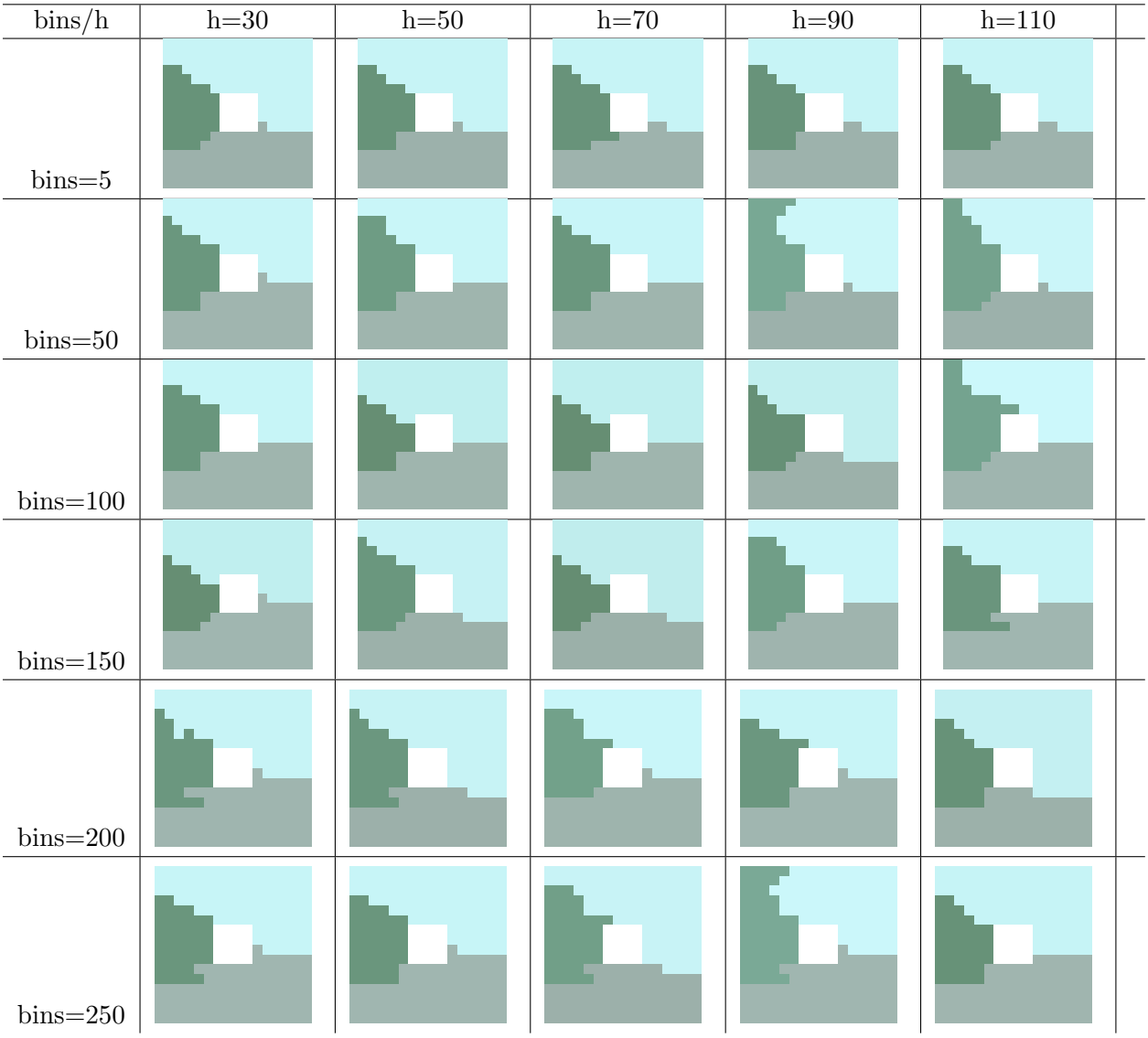


Figure 3.24: Size of the similarity window of 3×3.

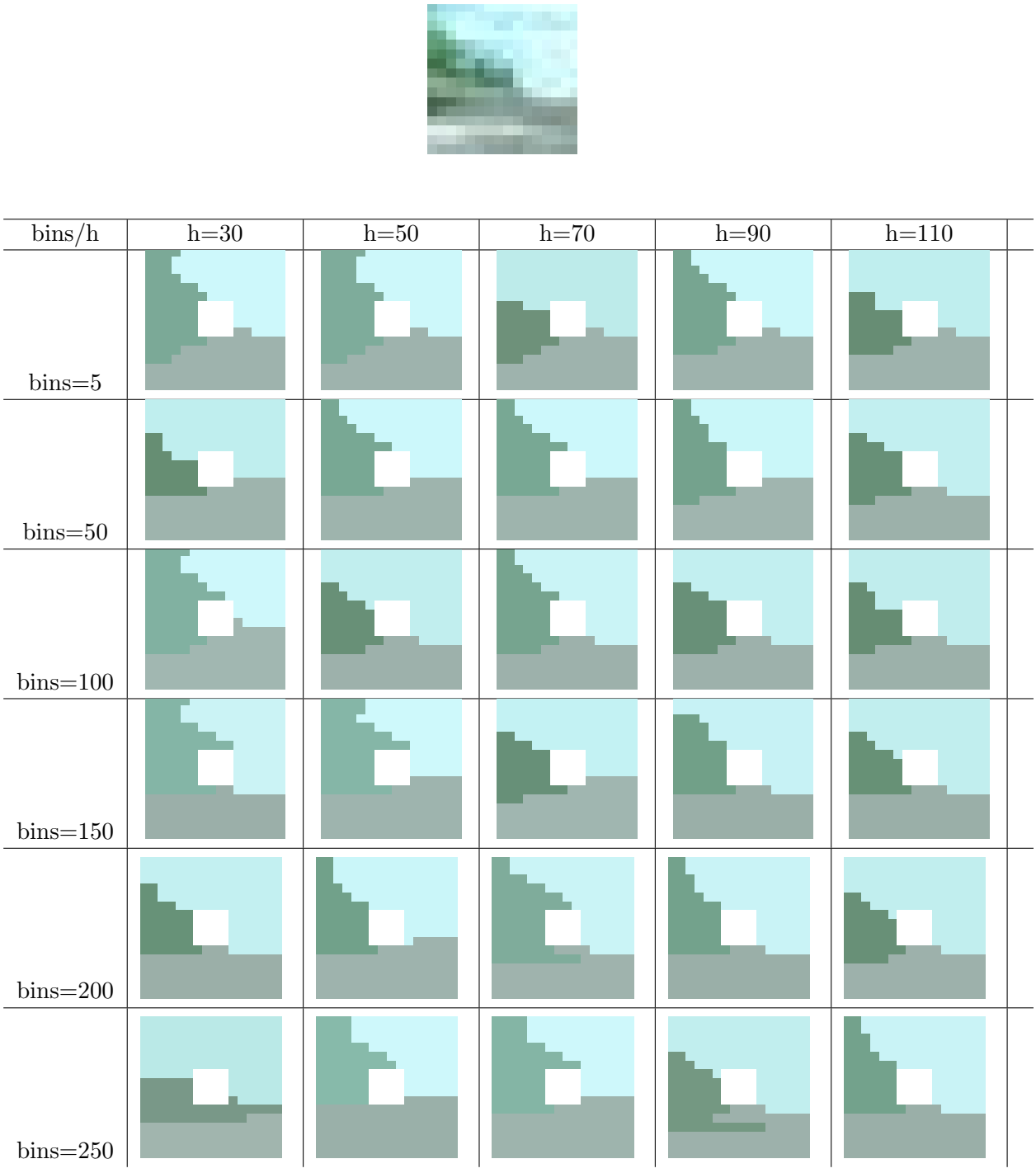


Figure 3.25: Size of the similarity window of 5×5.

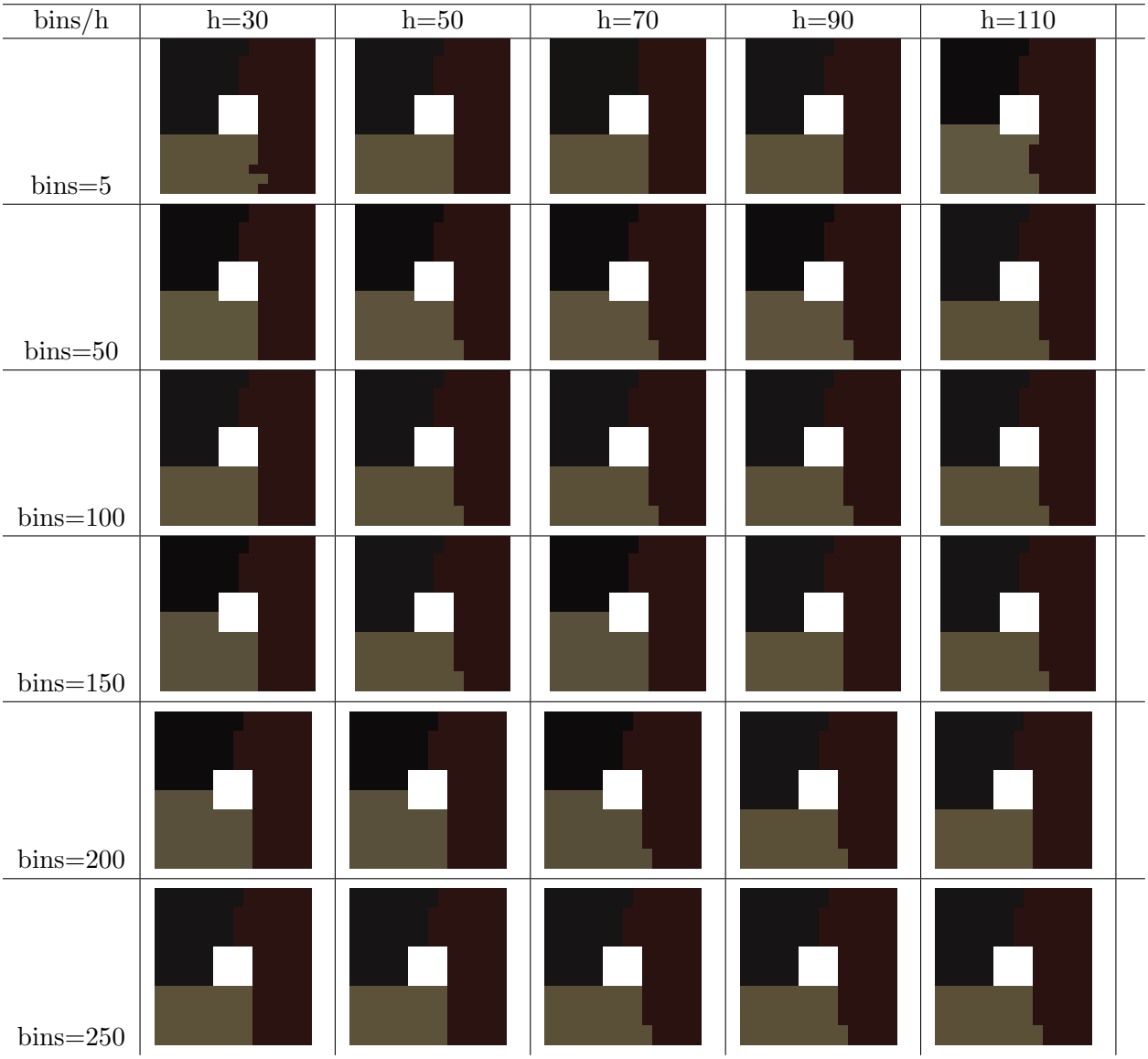


Figure 3.26: Size of the similarity window of 3×3.

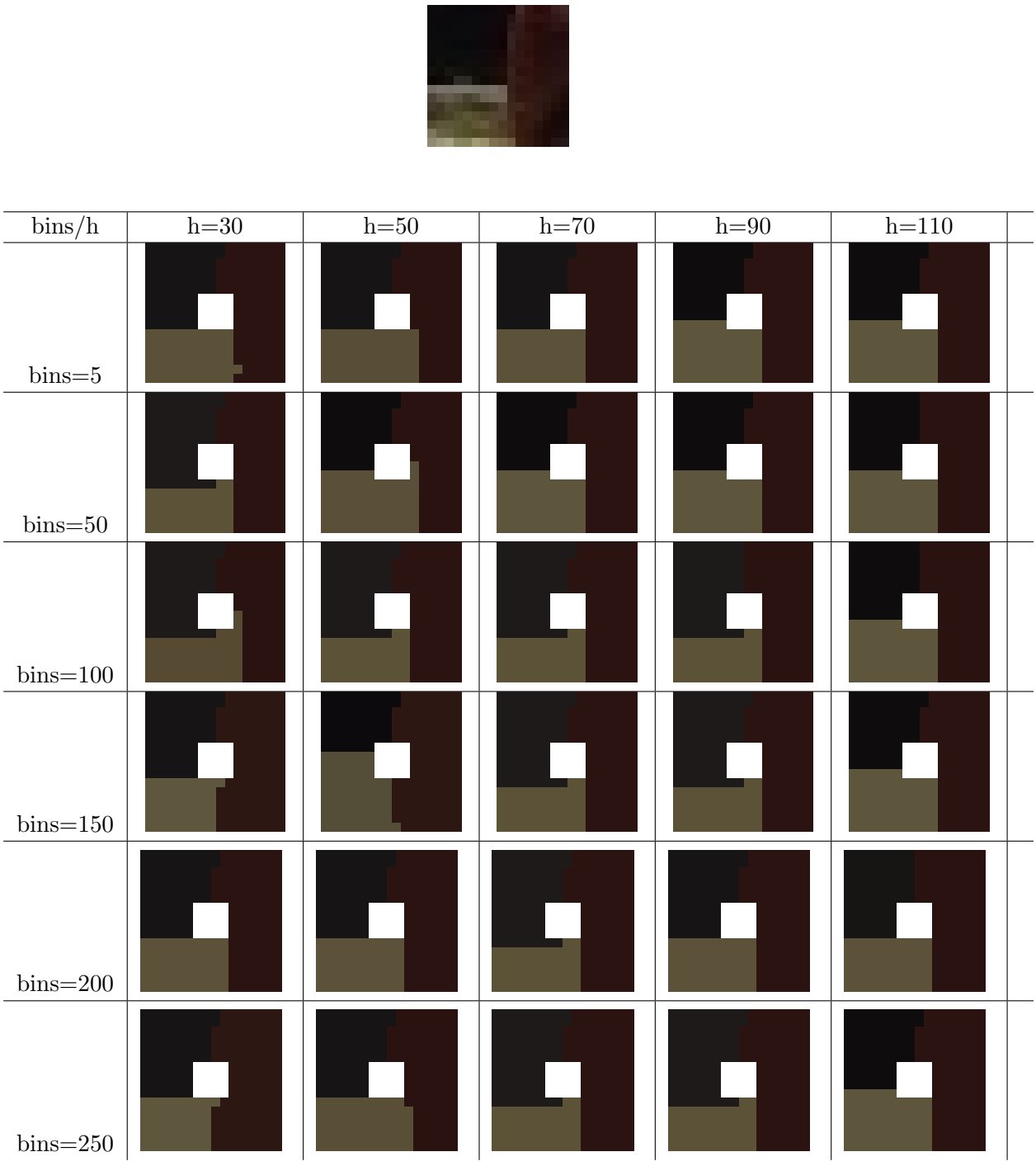


Figure 3.27: Size of the similarity window of 5×5.

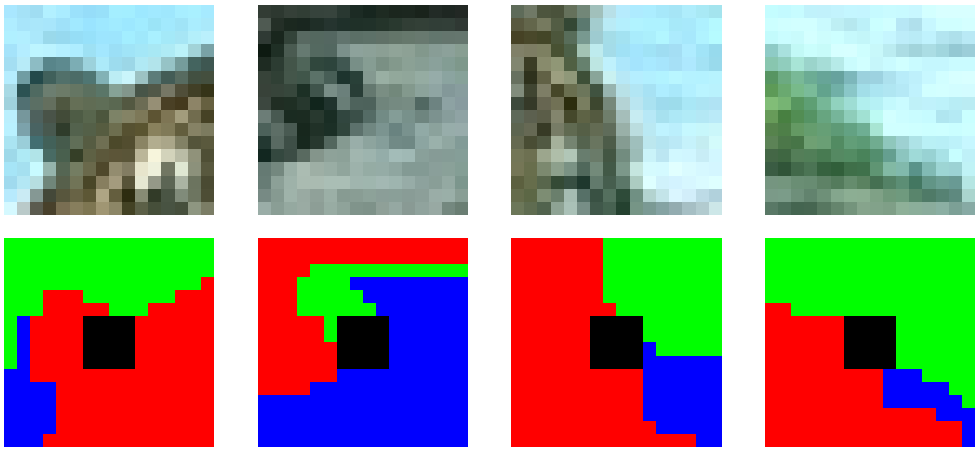


Figure 3.28: Examples of candidate points discarded because: (a) Not all three regions join the boundary of Ω. (b) Not all three regions join the boundary of W with a minimum number of pixels. (c) The medium color difference between two regions is too small. (d) One of the three regions is too small.

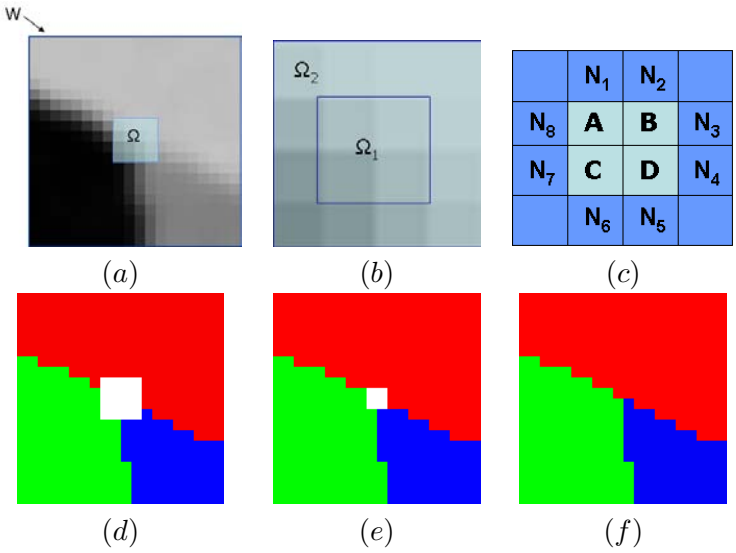


Figure 3.29: (a) Image to be segmented. (b) Partitioning of $\Omega = \Omega_1 \cup \Omega_2$. (c) Labeling of Ω_1 and Ω_2. (d) Branch extraction in $W - \Omega$. (e) Branch propagation in Ω_2. (f) Branch propagation in Ω_1.

made according to the "good continuation principle" [Mon71] and it is achieved in two steps. Let $\Omega = \Omega_1 \cup \Omega_2$ (see Figure 3.29(b)).

The first step consists in assigning to each pixel in Ω_2 the value of its adjacent pixel (in connectivity 4) outside Ω with the constraint that all three labels (red, green, blue) have to be assigned to at least one of the pixels N_i (see Figure 3.29 (c)). This guarantees the propagation of contours as straight lines (see Figure 3.29 (e)). In the second step, the branch extrapolation in Ω_1 is achieved using a geometric criterion that minimizes the sum of the absolute value of the curvature at the new branch points created by the hypothetical labeling, with the constraint that branches meet at the candidate point. The curvature at the candidate point is computed eliminating the stem of the hypothetical T-junction.

Let P be the set of pixels to be labeled, N the set of neighbors, and L the set of labels to be assigned (see Figure 3.29 (c)).

$P = \{A, B, C, D\}$
$N = \{N_1, N_2, N_3, N_4, N_5, N_6, N_7, N_8\}$
$L = \{red, green, blue\}$ In 4-connectivity, each pixel in P has two neighbors in N, whereas any pixel in N has only one neighbor in P.

To fulfill the constraint that branches meet at the candidate point, each label has to be assigned at least once at pixels in P. This goal is pursued in two stages: the first one is devoted to perform all label assignments that are mandatory (to fulfill the constraint). The second one, is devoted to label the remaining pixels.

The assignment of the label L_m of N_j to its neighbor in P, say P_i, is said mandatory if:

- N_j is the only pixel of N labeled with L_m such that its neighbor in P is still to be labeled
- there is no pixel in P labeled with L_m

Let us suppose that the label L_m of N_j is assigned to P_i. Then the other neighbor in N of P_i, say N_k has become useless since its unique neighbor in P has already been labeled. As a consequence, a new mandatory assignment may have been generated if the occurrence of the label of N_k, say L_n, is 2 before the mandatory assignment. Since there are only three labels for eight pixels, this "cascade effect" may only occur once. According to the above considerations, the algorithmic structure of the first step is as follows:

1. For each label L_m , compute how many neighbors belonging to N are labeled with label L_m, that is C_{red}, C_{green}, C_{blue}, where C_m is the number of neighbors N_i having label L_m.
2. If C_m is equal to 1:
   (a) search the N_i having value L_m

(b) assign L_m to P_j such that P_j and N_i are neighbors

(c) consider the neighbor of P_j: one is N_i , the other one is N_k. Decrement C_k by one.

3. go back to 2

The second step is a propagation (possibly with constraints) that minimizes a cost function. The possible constraints correspond to labels that still have not been assigned. The optimization cost is defined as the sum of the absolute value of the curvature of the level line at the new branch points created by the hypothetical labeling. The algorithmic structure of the second step is as follows:

- If there is a P_j whose two neighbors in N have the same label L_m, then P_j has to be labeled with L_m.
- If there are P_j that still have to be labeled, then compute all possible assignments and their costs, and among all assignments that satisfy the constraints, choose the one having the minimum cost.

The final result is shown in Figure 3.29 (f).

3.1.3.6 Candidate points validation after branch extraction in Ω

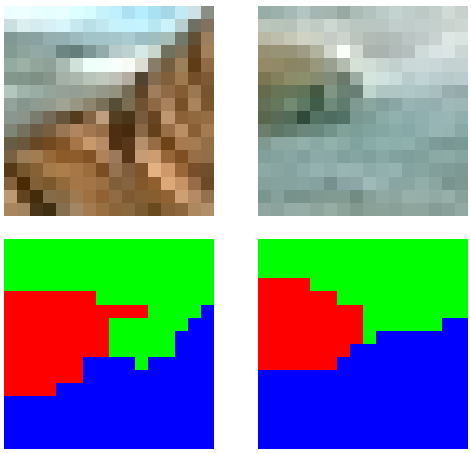
After the branch extraction is completed, three validation criteria are checked. The first validation criterion consists in measuring the "smoothness" of the branches. Since object contours are smooth, T-junction branches are expected to be smooth. A single branch is considered to be smooth if the value of the integral of the absolute value of curvature on the three branches is below to a given threshold t. The value of this threshold depends on the window size w and it is computed as follows: $k = cw/2$, where c is a small value (see Figure 3.30 (a)).

The second validation criterion deals with the branches orientation and it is necessary in order to distinguish from other junction types. For each branch, we first compute the vector that represents its medium orientation in W by averaging the orientation vector of each point x of the branch, which is given by:

$$Orient(x) = \frac{1}{\|Du(x)\|}(-\frac{\partial u(x)}{\partial y}\frac{\partial u(y)}{\partial x}) \tag{3.8}$$

Then, we compute the angles between each pair of vectors. We say that a candidate point represents a T-junction if there is a pair of vectors such that the angle between them is equal to π with precision $\frac{1}{n}$ (see Figure 3.30 (b)). In all experiments we fixed $n = 4$.

The third validation criterion deals with clusters of points obtained as a result of the first two validation criteria (see Figure 3.32 (a)). Clusters are due to the locality of the branch



(a) (b)

Figure 3.30: Examples of candidate points discarded because: (a) The sum the curvature of each point of a branch is too large. (b) There is no pair of branches with an angle of $\phi + -\phi/4$.

extraction strategy: the shape of wedges of adjacent candidate points varies slightly and if a candidate point is validated, its neighbors have a high probability of being validated too. As a consequence, isolated points have an high probability of being spurious detection and therefore they are removed (see Figure 3.31).

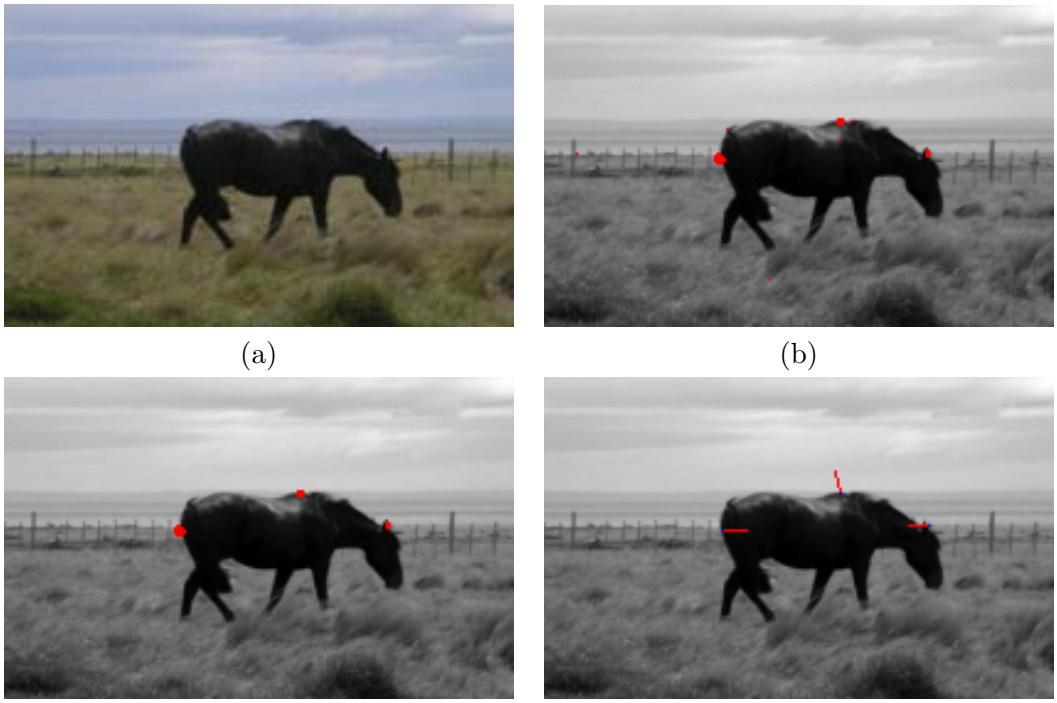
3.1.3.7 Cluster reduction

As result of the validation step, we obtain a set of clusters (see Figure 3.31 (c)). For each cluster, the choice of the point that best represents the cluster is based on a graduate measure of T-junction likelihood related to the smoothness of the branches. More precisely, the point that best represents the cluster is the one which has the minimum value of the sum of the absolute value of the curvature of the level line at each point of the branches (see Figure 3.31 (d) and Figure 3.32). The points of the branches are points with half-integer coordinates. The computation of the curvature at these points is based on an interpolation at the center of the 2×2 window made of pixels (i, j), $(i + 1, j)$, $(i, j + 1)$ and $(i + 1, j + 1)$.

Figure 3.33 summarizes the complete T-junction detection algorithm through a block diagram.

3.1.3.8 Parameter setting

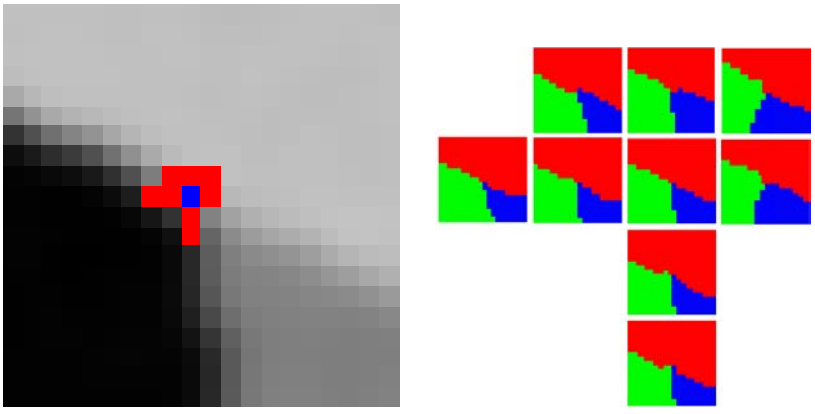
The validation process involves a set of parameters to be adjusted. Figures 3.36, 3.37 and 3.38 show the Precision/Recall (P/R) curves corresponding to the images in Figures 3.34 and 3.35, for which the ground truth has been fixed beforehand manually. In a feature detection task, the



(a) (b)

((c) (d)

Figure 3.31: Examples of cluster: (a) Original image. (b) Before removing isolated points: there are some spurious T-junctions. (c) After removing isolated points. (d) After the cluster reduction step: the vectors point to the region closer to the viewpoint. As can be observed, the roofs of the T-junctions have been correctly computed.



(a) (b)

Figure 3.32: Example of cluster reduction: (a) Validated points are marked as color pixels. The point having the smallest value of the sum of the absolute value of curvature of the level line at each point on each branch is marked in blue. (b) Local segmentations corresponding to each validated point: as can be observed, the pixel marked in blue in (a) is the one having the most regular branches.

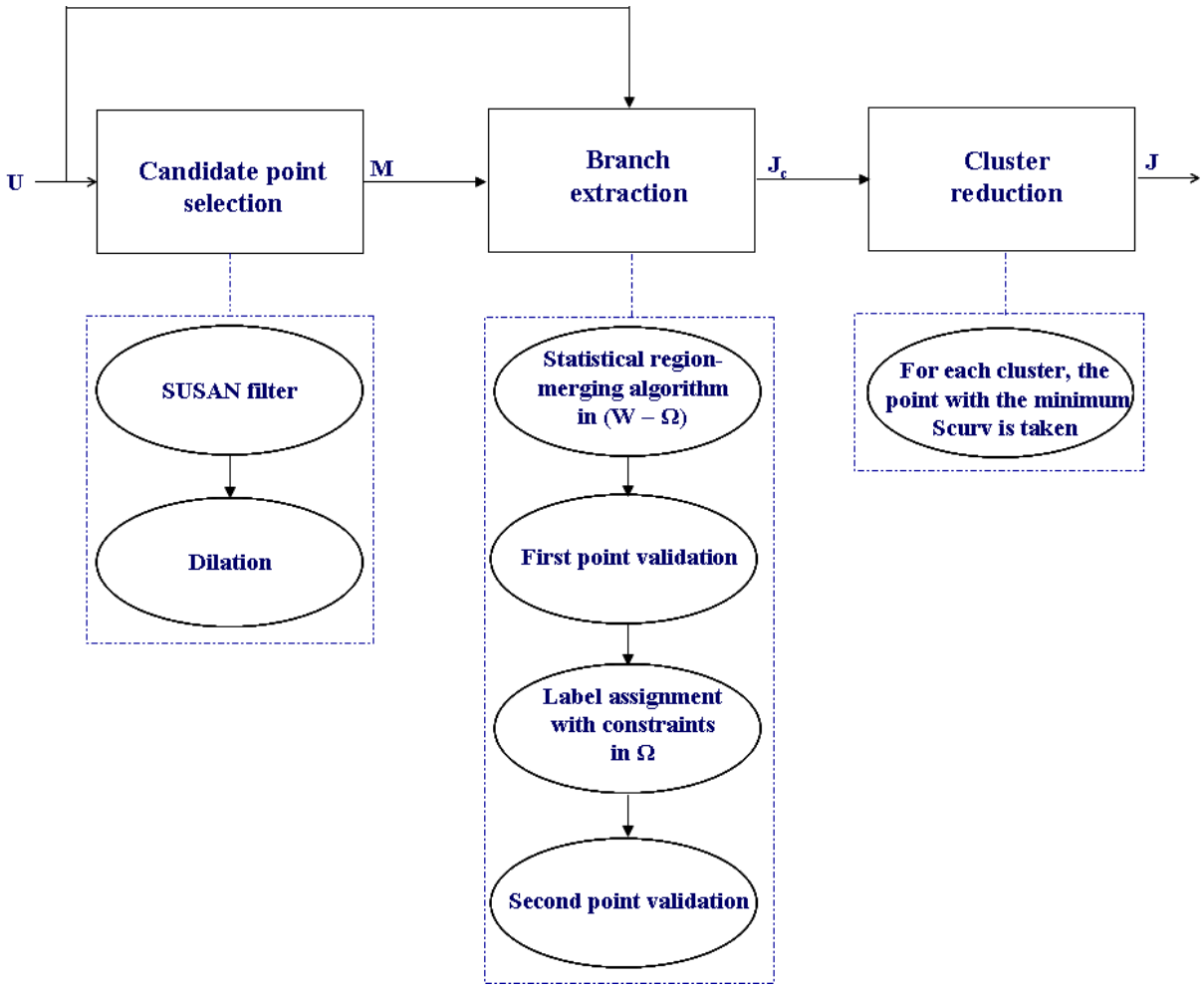


Figure 3.33: Block diagram of the T-junction detection algorithm by region merging. U is the original image, M is the mask marking the candidate points to be analyzed, J_c is the image marking the points that have been validated, J is the image marking the center of the detected T-junctions, S_{curv} is the sum of the absolute value of the curvature of the level line at each point of each branch.

Precision is the number of true positive detection divided by the total number of feature detected. Recall is defined as the number of true detection divided by the total number of features that actually have to be detected. Therefore, Precision can be seen as a measure of exactness of the detection, whereas Recall is a measure of completeness of the detection. The image set forming the benchmark for parameters optimization has been chosen to minimize the subjectivity of any hand-marked ground truth and to maximize the variety of natural scene. Our goal here is not to evaluate the performance of the algorithm but just to study the performance of the T-junction detection algorithm as a function of the parameters. The parameters that need to be fixed are the following:

- *Maximum value of the sum of the absolute curvature of each point of a branch*: the curves in Figure 3.36 have been obtained by assigning to this parameter a value in the set $\{0.3, 0.4, ..., 1.6, 1.7\}$. As can be observed, when the parameter value is small the recall is low because many T-junctions having curved branches are missed and the precision is high because very few T-junctions are detected and there are very few false positive. By increasing the value of the parameter the recall increases and at the same time increases the possibility of detecting some spurious T-junctions and therefore the precision decreases.

- *Minimum color difference between two wedges*: the curves in Figure 3.37 have been obtained by assigning to this parameter a value in the set $\{6, 8, ..., 52, 54\}$. As can be observed in Figure 3.37, when this value is small more T-junctions are detected and therefore the precision is low because many spurious T-junctions might be detected while the recall is high because most real T-junctions are detected. By increasing the value of this parameter, less T-junctions are detected and therefore the precision increases but the recall may decrease because some T-junctions could be missed.

- *Minimum gray level difference between two wedges*: the curves in Figure 3.38 have been obtained by assigning to this parameter a value in the set $\{2, 4, ..., 48, 50\}$. As can be observed in Figure 3.38, this parameter has the same behavior as the minimum color difference: for small values of minimum gray level difference, the precision is low and the recall is high, whereas when it is increased the precision increases and the recall decreases.

The P/R curves have been computed taking 0.8 as maximum sum of the absolute curvature for each branch, 10 and 5 respectively as minimum color difference and minimum gray level difference between wedges. However, looking at the curves a recommendable set of values for the three above mentioned parameters are 0.7, 20, and 20.

Figure 3.39 shows some examples obtained by using this set of parameter values. Each detected T-junction has been visualized as a vector whose origin (in blue) represents the T-junction center, whose direction is given by the direction of the stem and whose orientation

points to the top. The overall performances are in general convincing.

3.1.3.9 T-junction detection by LSD versus T-junction detection by region merging

In this section, the performances of the line-segment based junction detector and the region-merging based junction detector are compared. The comparison is limited to T-junctions. Table 3.1 shows the value of precision and recall obtained performing the two methods on the images in Figure 3.34 and Figure 3.35. As can be observed, the performances of the region-merging approach are much better specially for the images in Figure 3.35. The poor performances of the line-segment based junction detector for these images is mainly due to the fact that the LSD does not use color information. Indeed, looking at the gray level version of the images in Figure 3.41 (c) and (d), it is difficult to perceive object boundaries without using prior experience. The results of applying the line segment-based junction detector and the region merging-based junction detector are shown in Figure 3.43 and Figure 3.44 respectively.

Image	Method	Precision	Recall
3.40 (a)	Meth1	0.67	1.0
3.40 (a)	Meth2	1.0	0.8
3.40 (b)	Meth1	0.715	0.625
3.40 (b)	Meth2	1.0	0.75
3.40 (c)	Meth1	0.167	0.09
3.40 (c)	Meth2	0.89	0.73
3.40 (d)	Meth1	0.125	0.1
3.40 (d)	Meth2	0.78	0.7

Table 3.1: *Evaluation in terms of precision and recall of the performance of the line segment based junction detector (Meth1) and the region-merging based junction detector (Meth2).*

3.1.3.10 Statistical local segmentation versus statistical global segmentation and deterministic local segmentation

Before ending section 3.1.3, we consider useful to answer to some questions that at this point could have arisen in the mind of some reader. In section 3.1.3.2 we have proposed a region merging algorithm that, starting from the initial partition of the pixels iteratively merges pairs of neighboring regions until a termination criterion is reached. Each pixel is modeled by a probability distribution and the order with which regions are merged depends on a similarity measure between the region models. The first question is: would it be possible to apply this region merging process to the entire image domain to detect T-junctions as intersection of the contours delimiting the regions of the final partition? That is a good question, considering that the work in [Mai08], reviewed in section 3.1.1, as well as the method proposed in 3.1.2,

(a)

(b)

Figure 3.34: Ground truth images used to fix the parameters. T-junctions are marked as a black point surrounded by a white ring.

(a)

(b)

Figure 3.35: Ground truth images used to fix the parameters. T-junctions are marked as a black point surrounded by a white ring.

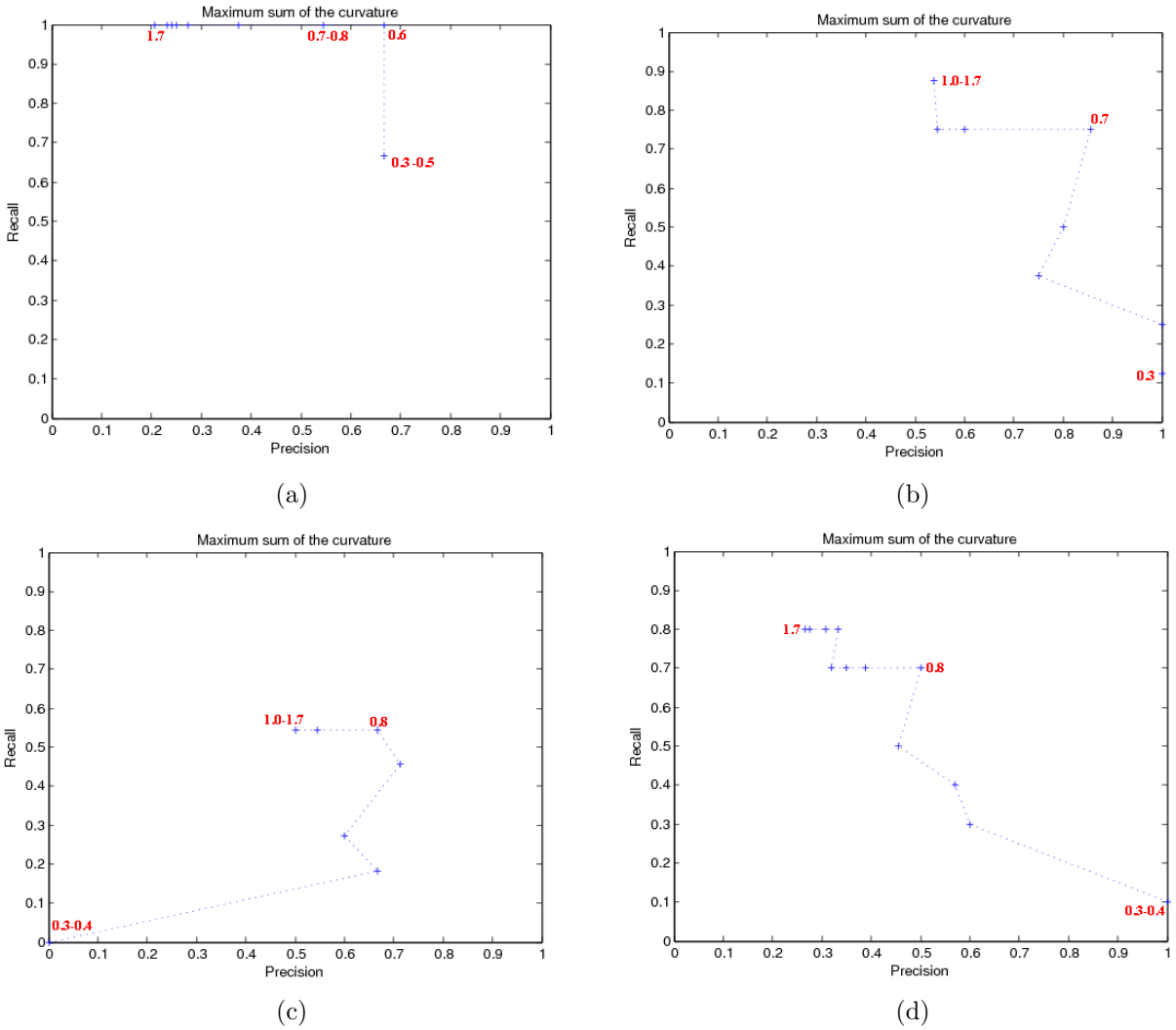


Figure 3.36: (a) Precision and Recall as a function of the maximum sum of the absolute curvature for each branch. The dynamic range is $\{0.3, 0.4, ..., 1.6, 1.7\}$.

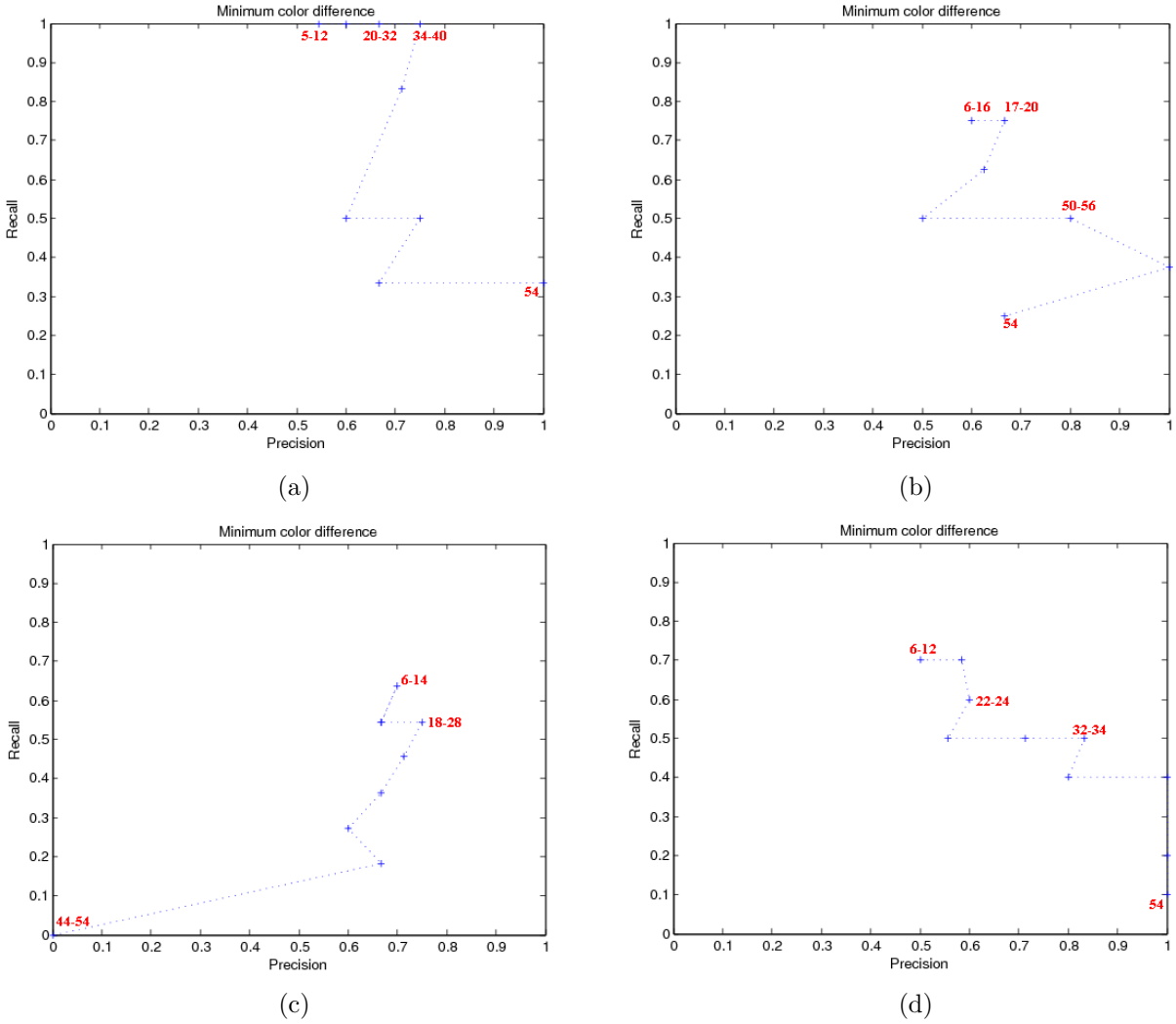


Figure 3.37: (a) Precision and Recall as a function of minimum color difference. The dynamic range is $\{6, 8, ..., 52, 54\}$.

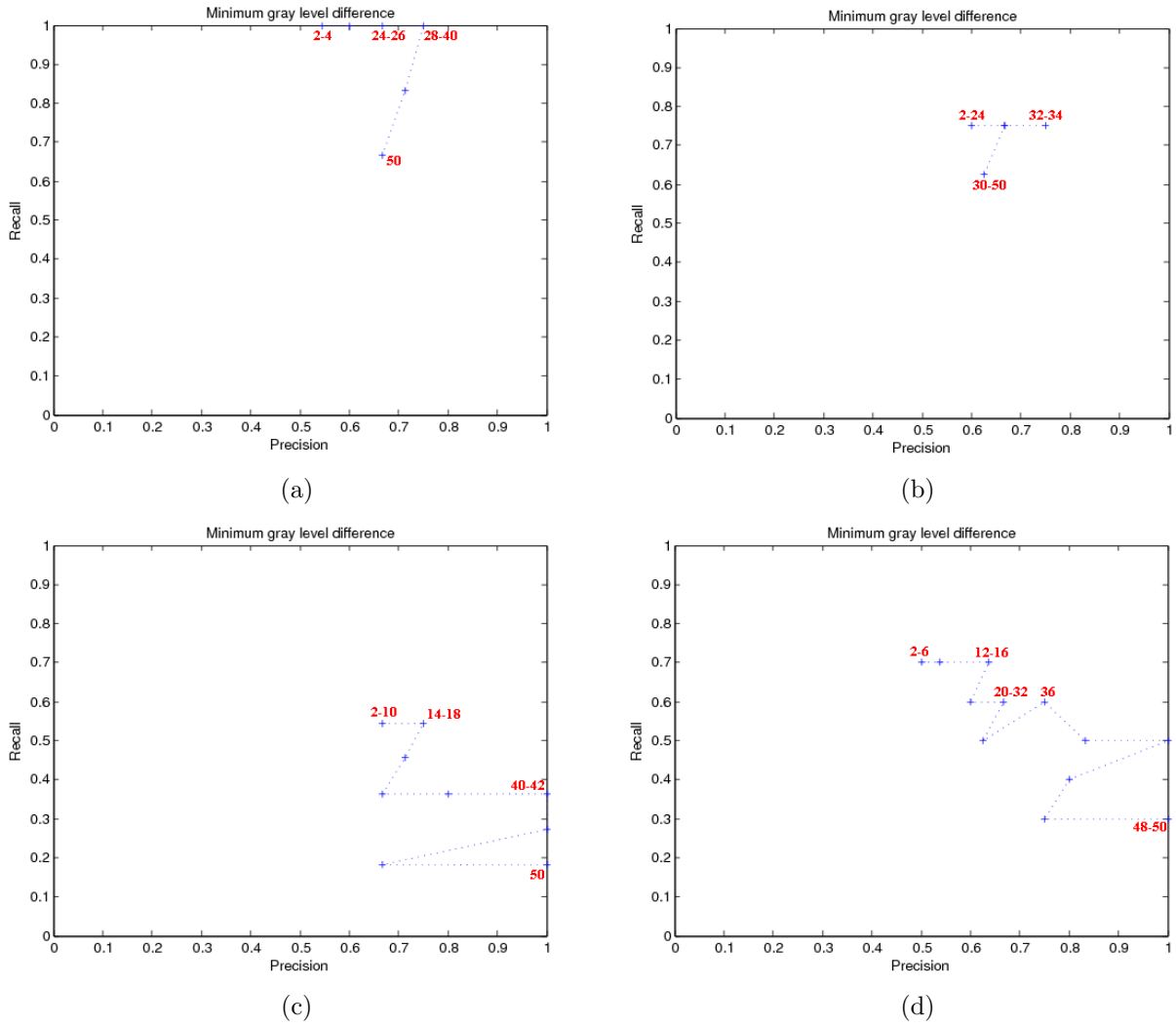
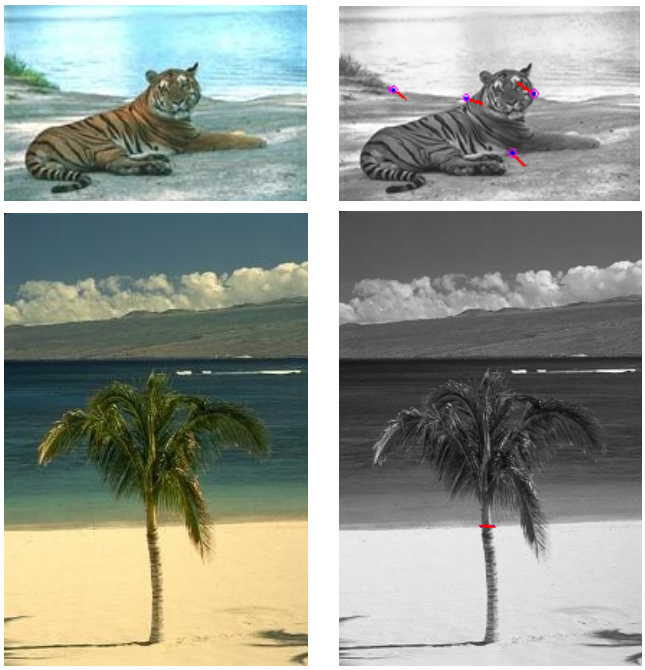


Figure 3.38: Precision and Recall as a function of the minimum gray level difference. The dynamic range is $\{2, 4, ..., 48, 50\}$.



(a) (b)

Figure 3.39: (a) Original image. (b) Result of the junction detection on the original image.

have demonstrated that more global approaches for junction detection give quite satisfactory results. However, one of the main goal of this Ph.D. dissertation being to improve segmentation accuracy by exploiting local depth cues, the question formulated above would become licit only with one condition. The condition consists in that the new pixel modeling proposed in section 3.1.3.2 would achieve an improvement in term of segmentation accuracy with respect to the work Calderero and Marques [Cal08], even without detecting T-junctions. To test this condition we have applied the statistical region merging algorithm proposed in section 3.1.3.2 to the entire image domain. As can be observed in the examples in Figure 3.45, even in relatively early stages of the merging process (see Figure 3.45 (b)), the segmentation in correspondence of T-junctions is not accurate. Furthermore, when we attempt to get a segmentation with a reduced number of regions, the T-junctions disappear (see Figures 3.39 (c)). Instead, when applying the method described in 3.1.3, T-junctions are correctly detected (see Figure 3.39). These results can be attributed to an insufficient locality of the region merging process. Indeed, the order with which nodes are merged is established taking into account the complete set of image regions. Another key element of the performances of the algorithm proposed in section 3.1.3 is that a small neighborhood of the center of candidate points is omitted in the local segmentation process, avoiding that the points of this small neighborhood, which may have a very similar window around them, can be merged together.

The second question is: what are the practical advantages of the statistical pixel modeling

(a) (b)

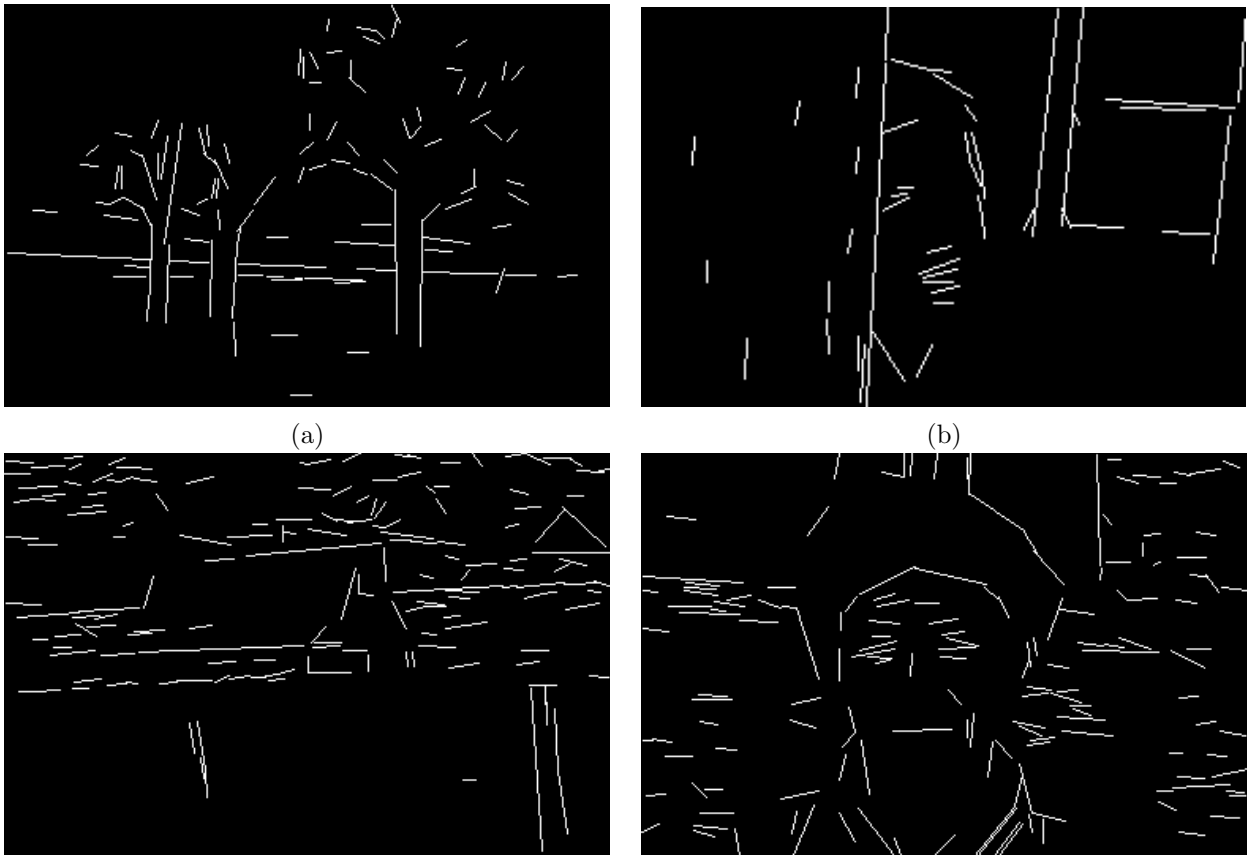
(c) (d)

Figure 3.40: Images used to compare the performances of the line segment-based T-junction detector and the region merging-based T-junction detector.

(a) (b)

(c) (d)

Figure 3.41: Gray level version of the images in Figure 3.40.



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

Figure 3.42: Results of applying LSD to the image in Figure 3.41.

(a) (b)

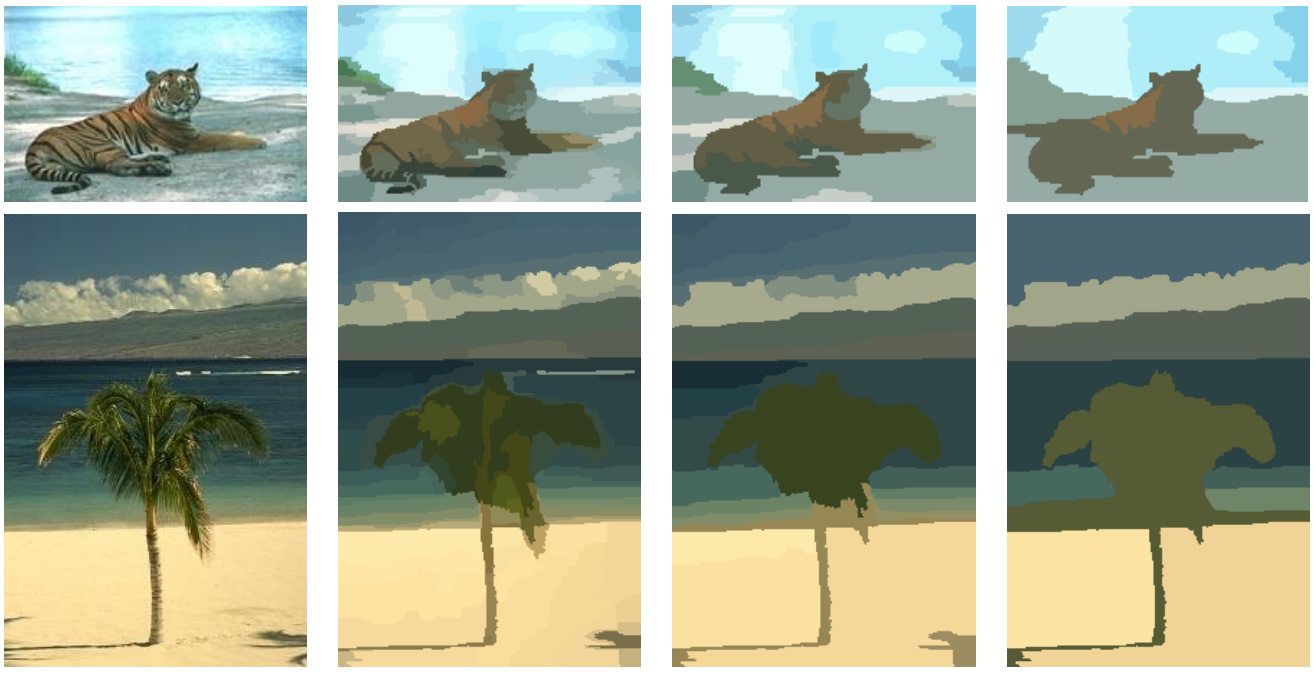
(c) (d)

Figure 3.43: Result of applying the line-segment based junction detector to the images in Figure 3.41. Only the detected T-junctions have been visualized.

(a) (b)

(c) (d)

Figure 3.44: Result of applying the region-merging based junction detector to the images in Figure 3.40.



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

Figure 3.45: Examples of image segmentation by the statistical region merging algorithm proposed in section 3.1.3.2. (a) Original image. (b) Segmentation with 80 regions. (c) Segmentation with 25 regions. (d) Segmentation with 10 regions.

with respect to the deterministic pixel modeling in the context of T-junction branch extraction? To illustrate the efficiency of the statistical modeling of pixels, Figure 3.46 shows some examples of comparison between a deterministic local segmentation, where each pixel is modeled by its color value, and a statistical local segmentation, where each pixel is modeled through a probability distribution. As can be observed, the statistical local segmentation 3.46 (g),(h), and (i) gives much better visual results with respect the deterministic local segmentation 3.46 (d),(e), and (f).

3.2 Transparency

In chapter 2, the depth cue of transparency has been introduced and the Transmittance Anchoring Principle (TAP), which summarizes all the geometrical and photometrical constraints to be held for perceiving transparency, has been stated and discussed (see section 2.1.2.3). In this section, we propose an algorithmic translation of the TAP. Recall that the detection of transparency involves not only the detection of X-junctions, but also a photometrical characterization which requires to check the polarity constraint. The detection of X-junctions is performed by the line segment detector proposed in section 3.1.2, since the region merging approach has been currently developed only for T-junctions. Let A, B, C, and D be the four regions delimited by

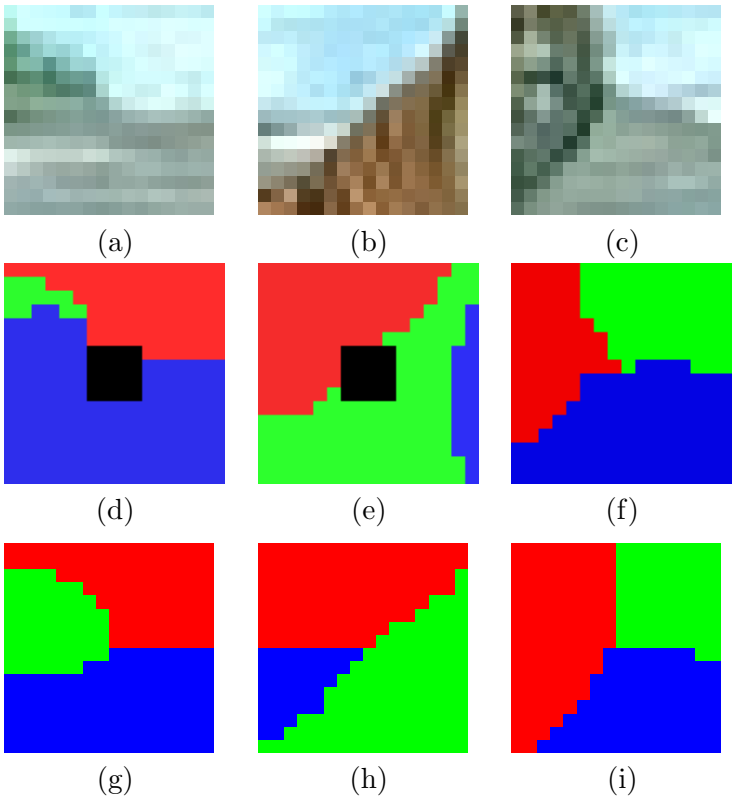


Figure 3.46: Example of comparison: (a), (b), and (c) Neighborhood of a candidate point to be segmented. (d),(e), and (f) Segmentation of the images respectively in (a), (b), and (c) by a pixel modeling of order zero. (g), (h), and (i) Segmentation of the images respectively in (a), (b), and (c) by the statistical region modeling proposed in 3.1.3.2. Points in 3.46 (d),(e) are discarded before the branch propagation in Ω since the segmented neighborhood does not fit the validation conditions described in 3.1.3.4.

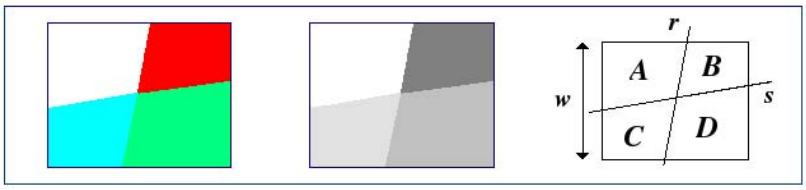


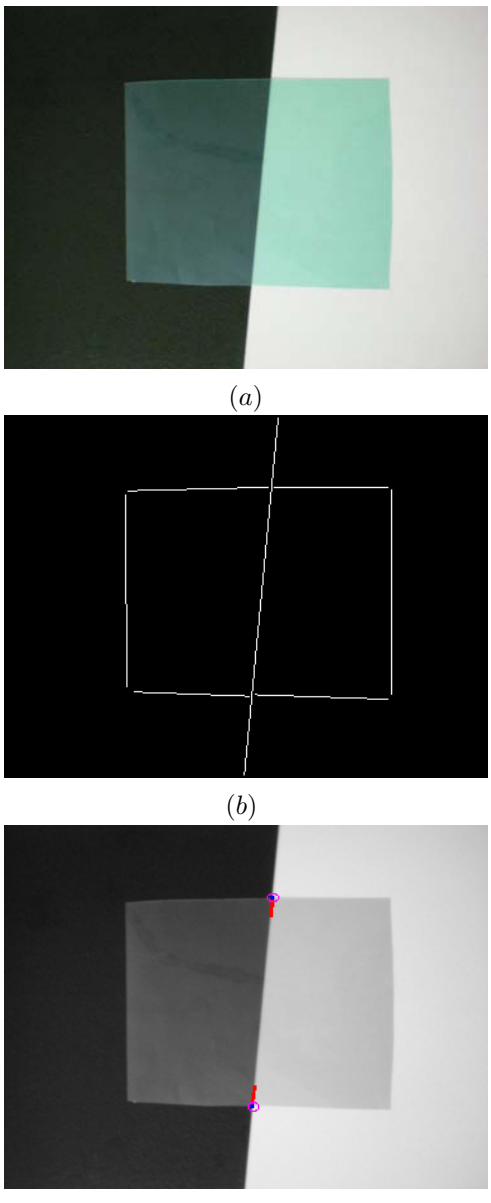
Figure 3.47: The polarity constraint tells us that s is the contour of the transparent object, since the polarity of the contrast between pairs of adjacent regions delimited by r ((A, B) and (C, D)), does not change when s is crossed

the contours r and s forming the X-junction and a squared window of size w centered at the junction center (see Figure 3.47). The gray level representative of each region, a, b, c and d is obtained as a median value on each region.

If the regions A and B are separated by r and A and C are separated by s, then the polarity constraint is satisfied if the difference $a - c$ has the same sign as the difference $b - d$ or if the difference $a - b$ has the same sign as the difference $c - d$. In the latter case, s is the contour of the transparent object and r is the contour of the underlying object. In the former case the contrary is true. Figure 3.48 shows an example of real image involving transparency. The vectors point to the region closer to the viewpoint. As can be observed, by performing the algorithm proposed above, transparency is detected and a correct depth interpretation is given.

3.3 Visual Completion

From chapter 2 it has emerged that visual completion is one of the processes through which local depth estimations are organized into a global depth ordering. In section 2.1.2.1, it has also emerged that the process of visual completion is performed by the visual system only when the relatability condition between visible contours is held. However, the relatability condition is merely geometric and does not take into account the photometric profile of the contours under consideration. In this work, we formalize the concept of relatability and we also impose a photometric condition for visual completion. Let a_i b_i and c_i be respectively the mean color of regions A_i, B_i and C_i delimited by the contours forming the T-junction and a squared window of size w centered at the junction center (see Figure 3.49), where the index $i \in 1, 2$ refers to the corresponding regions belonging to the two relatable T-junctions. The relatability condition is checked only at candidate relatable pairs of T-junctions such that the mean color of the region forming the top, say a_1, and the regions forming the stem b_1 and c_1 have a mean color similar respectively to a_2, b_2 and c_2 or a_2, c_2 and b_2. For each candidate pair of T-junctions, if the outer angle Φ, formed by the intersection of the lines on which the stems of the T-junctions lie on, is less than $\pi/2$, the pair of T-junctions is considered relatable. Quantitative variations of



(*a*)

(*b*)

(*c*)

Figure 3.48: (*a*) Original image. (*b*) Segments detected by applying the LSD. (*c*) Image where the vectors point to the region closer to the viewpoint.

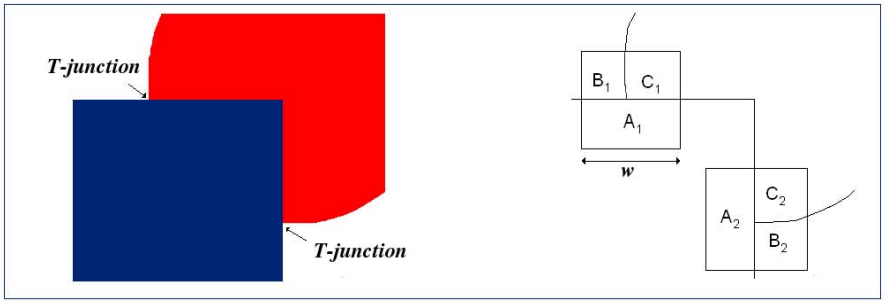


Figure 3.49: Amodal completion: pairs of relatable regions (A_1,A_2),(B_1,B_2), and (C_1,C_2) have similar color.

relatability (see section 2.1.2.1) are used to choose the best relatable T-junctions, when multiple candidate pairs satisfying the relatability conditions are possible. The value of Φ is used as first criterion and, in case of angle parity, the offset between the two stems of the T-junctions is computed (see Figure 3.50 (a_3)).

In the case of camouflage, the occluding object matches the color of only one of the two background objects and therefore pairs of T-junctions that lead to modal completion show up as pairs of L-junctions (see Figure 3.50 (b_3)). We shall call the L-junctions that lead to modal completion *degenerated T-junctions*. Pairs of degenerated T-junctions are detected using the quantitative variations of relatability.

3.4 Convexity

In chapter 2 it has been stressed that in the absence of viewing conditions that lead to partial objects obscuration, the factors that determine which regions are perceived as foreground and which as background, is related to the shape of the regions and not to their contrast polarity, nor any other texture property. With respect to other global shape properties, convexity has proved to have a stronger influence on figural organization. As it has been illustred in section 2.1.3.4, any convex curve (even if not closed) suggests itself as the boundary of a convex body on the foreground.

From a mathematical point of view, the convexity of a curve is related to the sign of its curvature vector. Let $u : \Lambda \subset \mathcal{R}^2 \rightarrow \mathcal{R}$ be a real image, Du the gradient of u and x a point of Λ such that $Du(x) \neq 0$ and, the level line of u that passes through x is a C^2 Jordan arc Γ. Then the curvature vector $\kappa(u)$ at x is defined by

$$\kappa(u)(x) = -curv(u)(x)\frac{Du}{|Du|}, \tag{3.9}$$

where $curv(u)(x)$ is the curvature of the level line Γ of u passing through x. The curvature

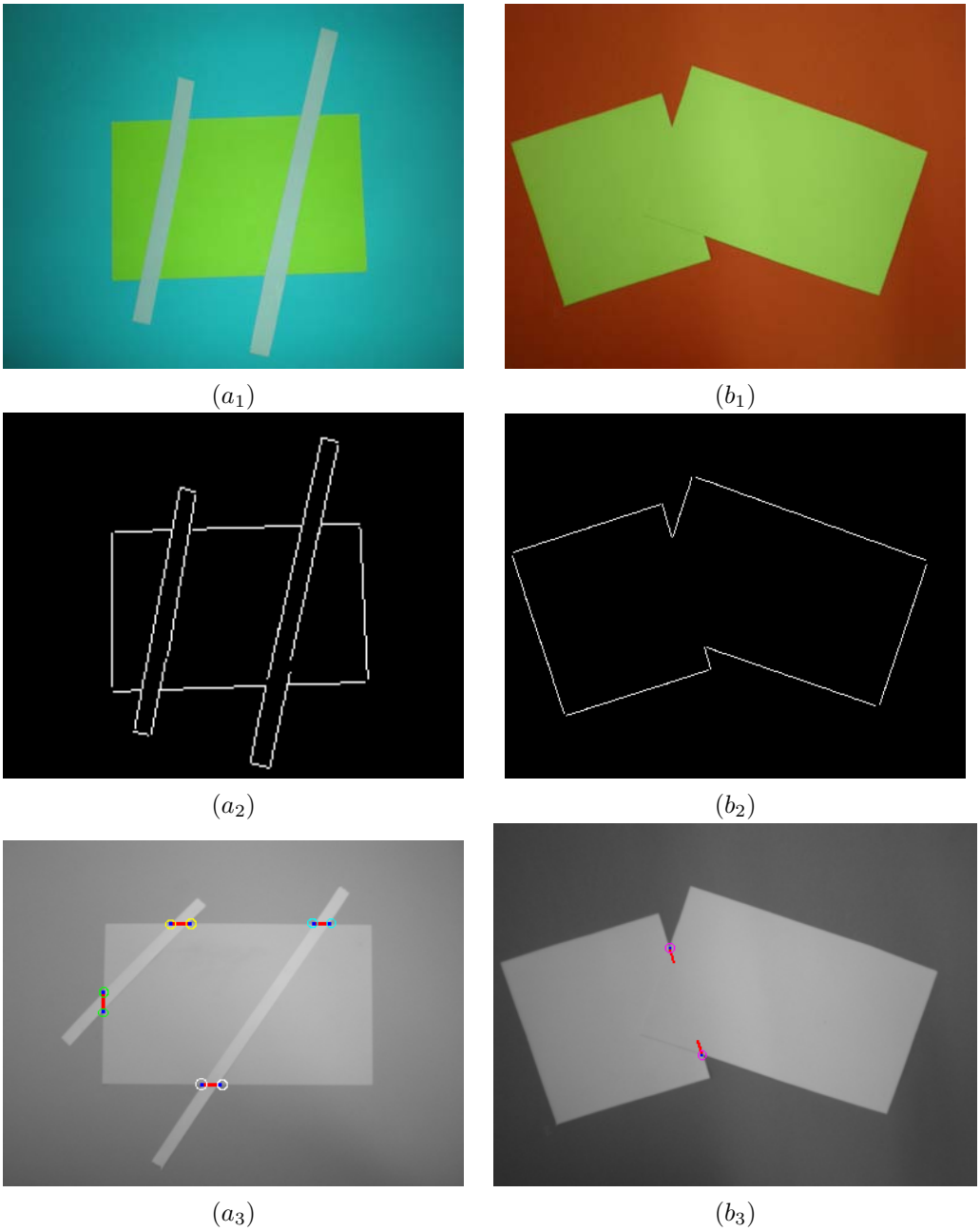


Figure 3.50: (a_1), (b_1) Original images. (a_2), (b_2) Segments detected by LSD. (a_3), (b_3) Local Depth Information is represented through a vector that points to the region closer to the viewpoint. In (a_3) pairs of relatable T-junctions have been marked with circle of the same color.

vector $\kappa(u)(x)$ is normal to Γ at x as the gradient vector $\frac{Du}{|Du|}$ and points towards the center of the osculating circle. Therefore, it points always to the convex part of a curve The formula defining $curv(u)(x)$ is

$$curv(u)(x) = \frac{1}{|Du|^3} D^2u(Du^\perp, Du^\perp)(x) = \frac{u_y^2 u_{xx} - 2u_x u_y u_{xy} + u_x^2 u_{yy}}{(u_x^2 + u_y^2)^{3/2}}(x), \tag{3.10}$$

To discretize the equation 3.10 we have used the finite difference scheme proposed in [Gui04]. Let us consider the discrete image model U and let $\Delta x = \Delta y$ be the pixel width on the discrete grid. A finite difference scheme for a differential operator T consists in writing T as a linear combination of the values of u on a fixed 3×3 stencil (see Figure 3.51). The problem of the discretization consists in determining the coefficients of the linear combination that ensure consistency with the differential operator. A finite difference scheme is said consistent if, when the length of the grid mesh Δx tends to zero, the discrete difference scheme tends to the differential operator. Indeed, numerically, Δx is equal to 1 but in the continuous image model it is considered as an infinitesimal length with respect to the image scale. In order to obtain the coefficient of the linear combination which ensure consistency we procede as follows. The differential operator $curv(u)$ can be expressed as:

$$curv(u) = \frac{u_{\xi\xi}}{|Du|}, \tag{3.11}$$

where ξ is the direction orthogonal to the gradient and $u_{\xi\xi}$ is the second derivative of u in the direction orthogonal to the gradient. By defining θ the angle between the horizontal direction and the gradient, ξ and $u_{\xi\xi}$ can be expressed in terms of θ:

$$\xi = (-sin(\theta), cos(\theta)) = \left(\frac{-u_y}{\sqrt{u_x^2 + u_y^2}}, \frac{u_x}{\sqrt{u_x^2 + u_y^2}}\right), \tag{3.12}$$

$$u_{\xi\xi} = sin^2(\theta)u_{xx} - 2sin(\theta)cos(\theta)u_{xy} + cos^2(\theta)u_{yy}. \tag{3.13}$$

The coefficients of the linear combination of the values of u on the fixed stencil would depend on θ. By symmetry, being the direction of ξ defined modulo π, the coefficients of points symmetrical with respect to the central point of the stencil must be equal (see Figure 3.51). By using the order 2 Taylor formula, we can remove the second derivatives from the equation 3.13 by expressing them as linear combination of the values of u on the fixed stencil. Therefore $u_{\xi\xi}$ can be written as linear combination of the values of u on the fixed stencil with coefficients λ_i:

$$\begin{aligned}(u_{\xi\xi})_{(i,j)} = \frac{1}{\Delta x^2}(&-4\lambda_0 u_{i,j} + \lambda_1(u_{i,j+1} + u_{i,j-1}) + \lambda_2(u_{i+1,j} + u_{i-1,j}) \\ &+ \lambda_3(u_{i-1,j-1} + u_{i+1,j+1}) + \lambda_4(u_{i-1,j+1} + u_{i+1,j-1}) + o(\Delta x)^2).\end{aligned} \tag{3.14}$$

By substituting for each partial derivative in equation 3.13 its discrete version, obtained by using the order 2 Taylor formula, and by using the equation 3.14 it is possible to find the values

λ_4	λ_1	λ_3
λ_2	$-4\lambda_0$	λ_2
λ_3	λ_1	λ_4

Figure 3.51: A 3×3 stencil.

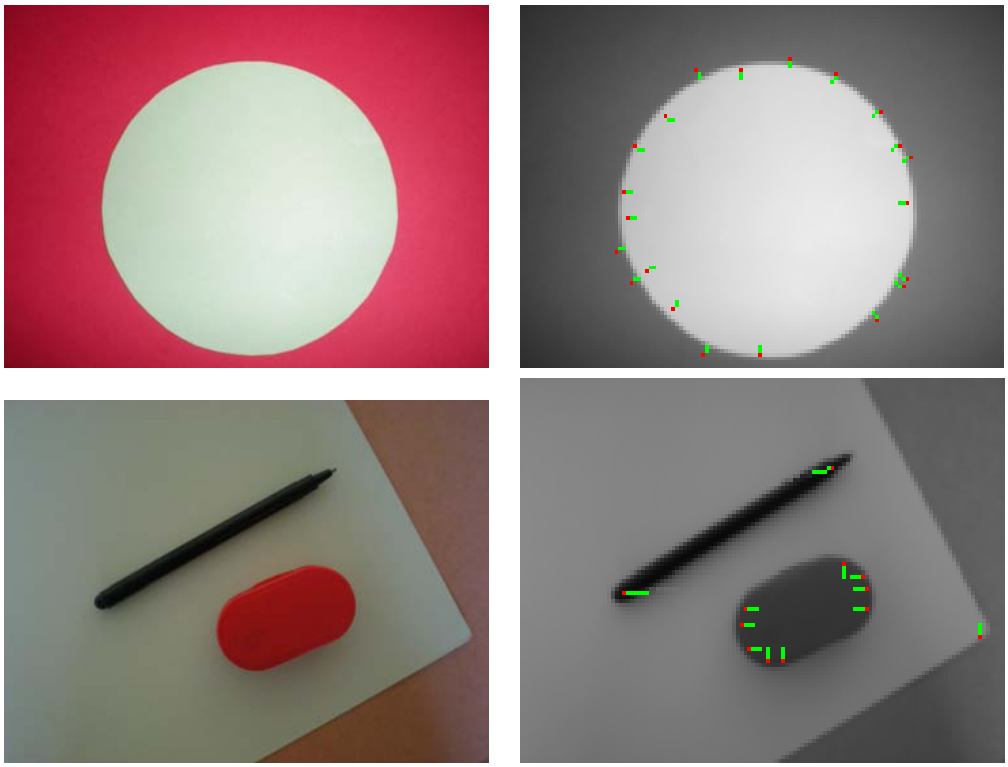
of the coefficients λ_i as function of θ (see [Gui04] for more details). The consistency is ensured by the fact that the discrete version of the gradient and of the partial derivative expressed by the order 2 Taylor formula are consistent. When the gradient $|Du| = 0$, the direction of the gradient is unknown and therefore $u_{\xi\xi}$ is not defined. In this case, $u_{\xi\xi}$ is replaced by half the Laplacian, which is equal to the sum of the two second derivative in orthogonal directions.

In Figure 3.52 are shown some examples of detection of convexity on real images. The convexity of a curve at a given image point is represented through a vector that has the same direction and orientation as the curvature vector at that point and therefore points to the regions closer to the viewpoint. For readability issues, for each cluster of curvature vectors, only the vector having the biggest magnitude has been drawn. As can be observed, the depth interpretation is correct.

3.5 Chapter summary

This chapter has focused on monocular depth cue detection. Two different approaches for detecting occlusion have been proposed: a line segment based approach and a region merging based approach. A comparative evaluation of the performance of the two methods has demonstrated that the region merging approach gives in general better results specially in images where the color information plays a crucial role in perceiving object contours. Algorithmic translations of the TAP for detecting transparency and of the relatability concept for detecting the inducers of visual completion have been presented. A mathematical translation of the cue of convexity has also been proposed.

In the following chapter, we address the problem of integrating depth information arisen from the above mentioned monocular depth cues to infer a global consistent depth ordering.



(*a*) (*b*)

Figure 3.52: (*a*) Original image. (*b*) The convexity at a given image point is represented through a vector having the same direction and orientation of the curvature vector at that point.

Chapter 4

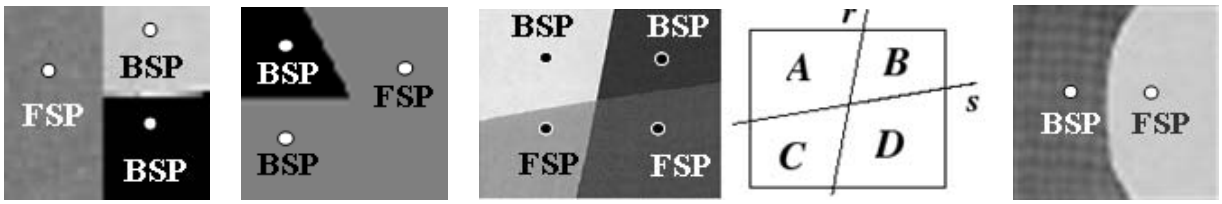
Monocular Depth Cue Integration

This chapter presents two different frameworks for monocular depth cue integration: a diffusion-based and a region-merging based framework. Section 4.1 is devoted to the diffusion-based framework. The encoding of the initial depth values arisen from monocular depth cues is illustrated, the nonlinear filter underlying the diffusion process is analyzed and experimental results are discussed. Section 4.2 details the region-merging based framework. The construction of a hierarchical region-based representation of the image, avoiding regions in occlusion to merge, is described and a graph formalization of the depth relationships between the regions of the final partition is proposed. The operations on the graph allowing to infer a global depth ordering are illustrated and experimental results are discussed. Finally, section 4.3 concludes the chapter with a comparative discussion on the performances of the two proposed approaches.

4.1 Diffusion-based framework

From chapter 2 it has emerged that monocular depth cues such as occlusion, camouflage, transparency and visual completion are crucial to identify the shape and depth relationships of depicted objects (see section 2.1.2). It has also been shown that, in absence of the viewing conditions that lead to partial object obscuration, configural cues are useful for distinguishing the foreground objects. A review of the most important theoretical perspectives in vision emerged during the last century has been done, evidencing the deep impact that the Gestalt School has had on more recent monocular depth perception theories. Nevertheless, it has also been remarked that the merely qualitative and descriptive character of Gestalt theory constitutes the major limitation for an effective exploitation in computer-based applications.

In this section, we attempt a mathematical and computational translation of Gestalt laws and principles governing the monocular perception of depth, from the detection of sparse local depth cues to their integration into a consistent, global depth perception. The detection of local depth



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

Figure 4.1: FSPs and BSPs arising from: (a) T-junctions, (b) L-junctions, (c) Transparency, (d) Convexity.

cues relies on the monocular depth cue detectors proposed in chapter 3. The depth integration process is based on the use of a nonlinear filter, which iteratively extends initial depth values arisen from monocular depth cues to the entire image domain until the stability is attained, allowing to recover the relative depths of depicted objects. This kind of strategy assumes that the set of the initial depth values to be propagated is available. However, monocular depth cues indicate solely the presence of a depth gradient between two or more surfaces, without providing any information about the absolute depth values of the piece of surfaces involved. The way initial depth values are derived from the detected monocular depth cues is detailed in the next section.

4.1.1 Computing Initial Depth Values

In the following, we shall call *source points* the set of image points for which the initial depth values can be inferred from the detected monocular depth cues. We shall call foreground source points (FSPs) the source points marking the regions that are closer to the viewpoint and background source points (BSPs) the source points marking the regions that are more distant to the viewpoint. Since the junction centers may not be accurately localized and the junction branches may not be precisely recovered, marking as source points all pixels in a neighborhood of the junction center would not be a good starting point for an accurate recovery of occlusion boundaries, which is one of the main goal of this Ph.D. dissertation. Instead, our strategy is to use as source points only one or a few pixels for each piece of surfaces involved and to remit the task of accurately computing occlusion boundaries in a neighborhood of a junction to the diffusion process itself. Hence, source points are positioned at a distance d from the junction center. This distance has to be of at least four pixels to be able to jump over blurred transition zones between different objects due to the image acquisition process.

The number of source points for each depth cue depends on the number of surfaces involved. In the case of occlusion and camouflage (see Figure 4.1 (a) and (b)), there are a single FSP and two BSPs. In the case of transparency (see Figure 4.1(c)), there are a two FSPs and two BSPs. In the case of convexity (see Figure 4.1 (d)), there are a single FSP and a single BSP. Source points arising from a T-junction at point P are computed as follows (see Figure 4.2(a)). Let s

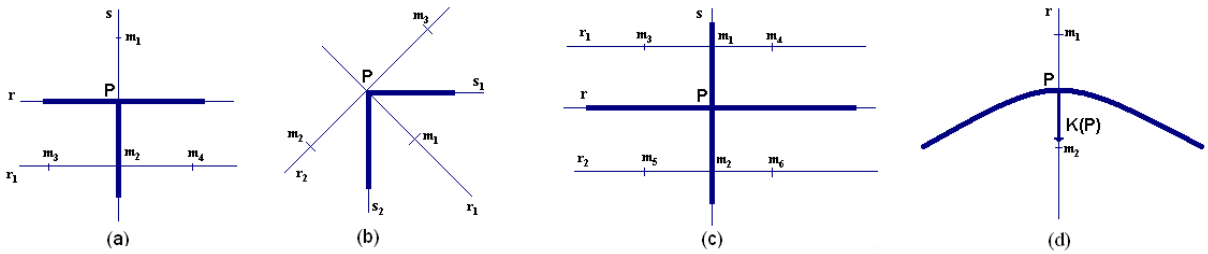


Figure 4.2: Computing BSPs and FSPs from: (a) Occlusion (T-junctions), (b) Camouflage (L-junctions), (c) Transparency (X-junctions), (d) Convexity (curved contour).

be the line containing the segment that forms the stem of the T-junction. Let m_1 and m_2 be the points belonging to s and having distance d from P. If m_2 is the point lying on the stem and r_1 the line perpendicular to s and passing through m_2, then m_1 is the FSP and the points m_3 and m_4 belonging to r_1 and having distance d from m_2 are the BSP. Source points arising from an L-junction at point P are computed as follows (see Figure 4.2(b)). Let r_1 be the bisecting of the angle formed by the segments that lie on the lines s_1 and s_2. Let m_1 be the point lying on r_1 and having distance d from P. Let r_2 be the line perpendicular to r_1 and passing through P. Let m_2 and m_3 be the points belonging to r_2 and having distance d from P. If s_1 is the stem, then m_1 and m_3 are the BSPs and m_2 is the FSP. Source points arising from transparency at point P are computed as follows (see Figure 4.2(c)). Let s be the line containing the contour of the transparent object, and m_1 and m_2 be the points belonging to s and having distance d from P. Let $r1$ be the line perpendicular to s and passing through m_1, and r_2 the line perpendicular to s and passing through m_2. Let m_3 and m_4 be the points belonging to r_1 and having distance d from m_1, and m_5 and m_6 the points belonging to r_2 and having distance d from m_2. If the gray level difference between m_4 and m_6 is larger than the gray level difference between m_3 and m_5, then m_3 and m_5 are FSPs whereas m_4 and m_6 are the BSPs. Source points arising from convexity at point P of a curve are computed in the following way (see Figure 4.2(d)). Let r be the line passing through P and having the direction of the gradient at P. Let m_1 and m_2 be the points belonging to r and having distance d from P. If m_1 is the point lying on the half-line having origin in P and oriented as the curvature vector at x, then m_1 is the BSP and m_1 the FSP.

Figure 4.3 (a) is an example of images where the local depth gradient arisen from convexity and occlusion is represented respectively by red and blue vectors that point to the region closer to the viewpoint. Convexity and occlusion are an example of conflict between local depth cues (see Figure 4.3 (a)). Indeed, T-junctions are a particular case of superposition of corners and in correspondence of each corner the curvature of the level line is high and the perception of convexity is strong. Following Gestalt principles, the depth cues which give the more global interpretation *masks* the other. In this case, occlusion inhibits convexity (see Figure 4.3 (b)). In

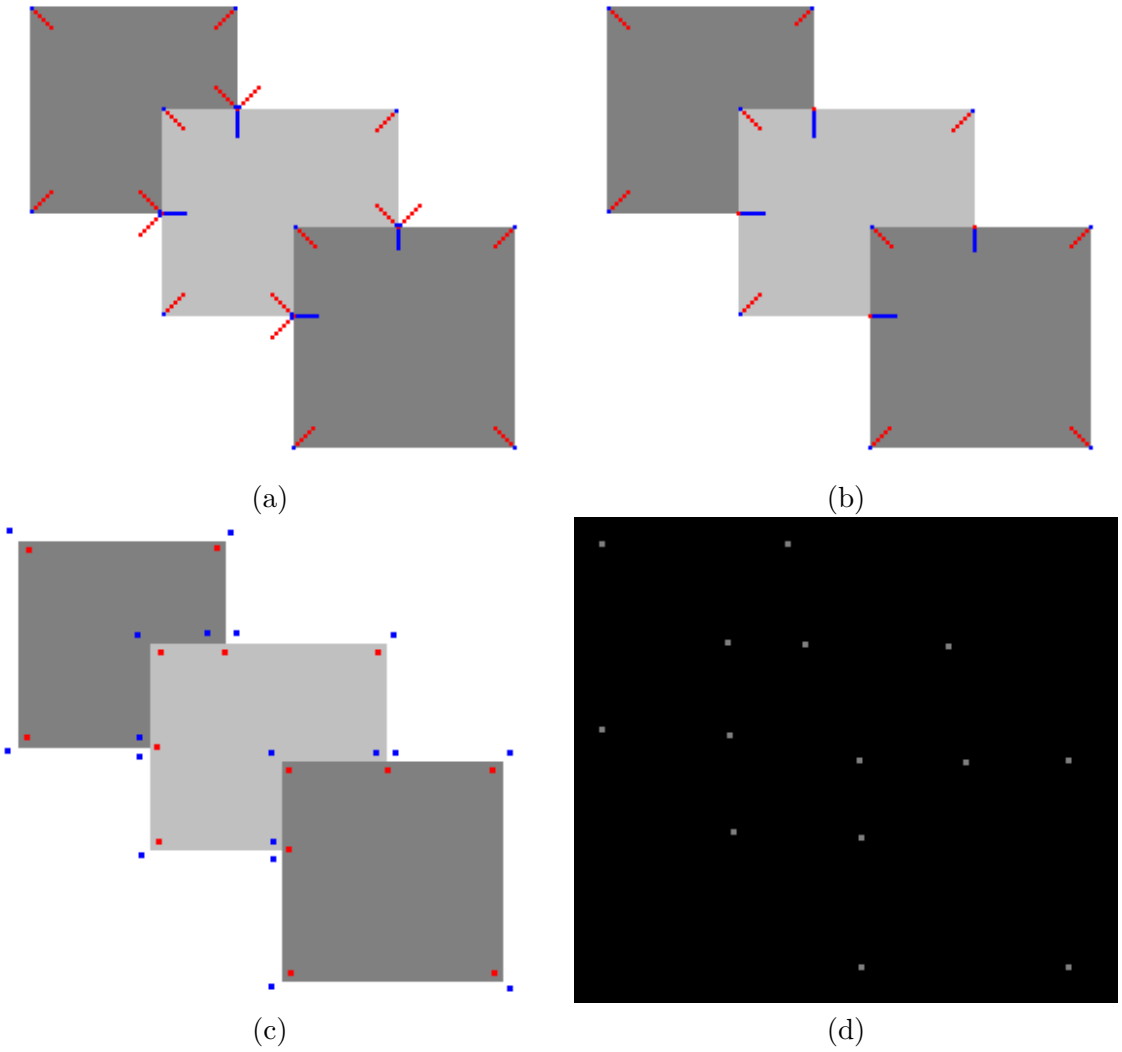


Figure 4.3: (a) Gray level image, where depth relationships arisen from convexity and occlusion are represented respectively by red and blue vectors, that point to the foreground region. The depth interpretations arisen from convexity and occlusion are always in conflict. (b) Occlusion inhibits convexity. (c) Gray level image, where BSPs and FSPs are marked in blue and red respectively. (d) Depth image: points corresponding to FSPs are initialized with a positive value (marked in white), whereas the rest of the image is initialized with value zero.

general, all local and non-local depth relationships are encoded by acting in an additive fashion under non-conflictive conditions (collaboration) and in a exclusive fashion under conflictive conditions (masking).

Once the position of source points is computed (see Figure 4.3 (c)), the depth image z is initialized by assigning a positive value to FSPs, and zero to BSPs. The rest of the image is initialized with value zero (see Figure 4.3 (d)).

4.1.2 Depth Diffusion

The diffusion process is based on the use of a neighborhood filter. In the context of image denoising, a neighborhood filter is any filter which replaces the actual value of the gray level (color) at a given pixel x by an average of the values of pixels in a neighborhood of x, defined taking into account the gray level (color) values. In the simplest and more extreme case, the neighborhood of a given point x of the image $u : \Lambda \subset \mathcal{R}^2 \rightarrow \mathcal{R}$ is defined as: $G(x,h) = \{y \in \Lambda | u(x) - h < u(y) < u(x) + h\}$, where h is a threshold. Therefore, pixels belonging to the whole image can be used for the estimation at pixel x. This is the case of the non-local means [Bua05], whereas earlier version of the neighborhood filter such as the Yaroslavsky filter [Yar85], the SUSAN filter [Smi97] and the bilateral filter [Tom98], define the neighborhood of a given pixel x by taking into account simultaneously spatial proximity and gray level (color). In the Yaroslavsky neighborhood filter (YNF), the filtered value at a point $x \in \Lambda$ is given by:

$$YNF_{h,\rho}u(x) = \frac{1}{C(x)} \int_{B_\rho(x)} u(y) e^{\frac{-|u(x)-u(y)|^2}{h^2}} dy, \tag{4.1}$$

where $B_\rho(x)$ is a ball of center x and radius ρ, h is the filtering parameter, which controls the decay of the exponential function, and

$$C(x) = \int_{B_\rho(x)} e^{\frac{-|u(x)-u(y)|^2}{h^2}} dy \tag{4.2}$$

is the normalization factor. Inside an homogeneous region, the gray level values slightly fluctuate because of the noise. In this case, the YNF filter computes an arithmetic average of the full neighborhood. Instead, at a contrasted edge separating two regions, if the gray level difference between both regions is larger than h, the neighborhood filter computes the average of pixels belonging to the same region as the reference pixel. Thus, the main advantage of this approach is that it allows to remove slight fluctuations due to noise without blurring the edges.

In [Bua06], Buades et al. have demonstrated that the YNF is equivalent to the Perona-Malik equation [Per90] when the size of the spatial neighborhood tends to zero. Indeed, they demonstrated that the YNF acts as an evolution PDE with two terms that makes the signal evolves proportionally to its second derivative. The first term is proportional to the second derivative of u in the direction of the gradient $\eta = Du(x)/|Du(x)|$, which is tangent to the level

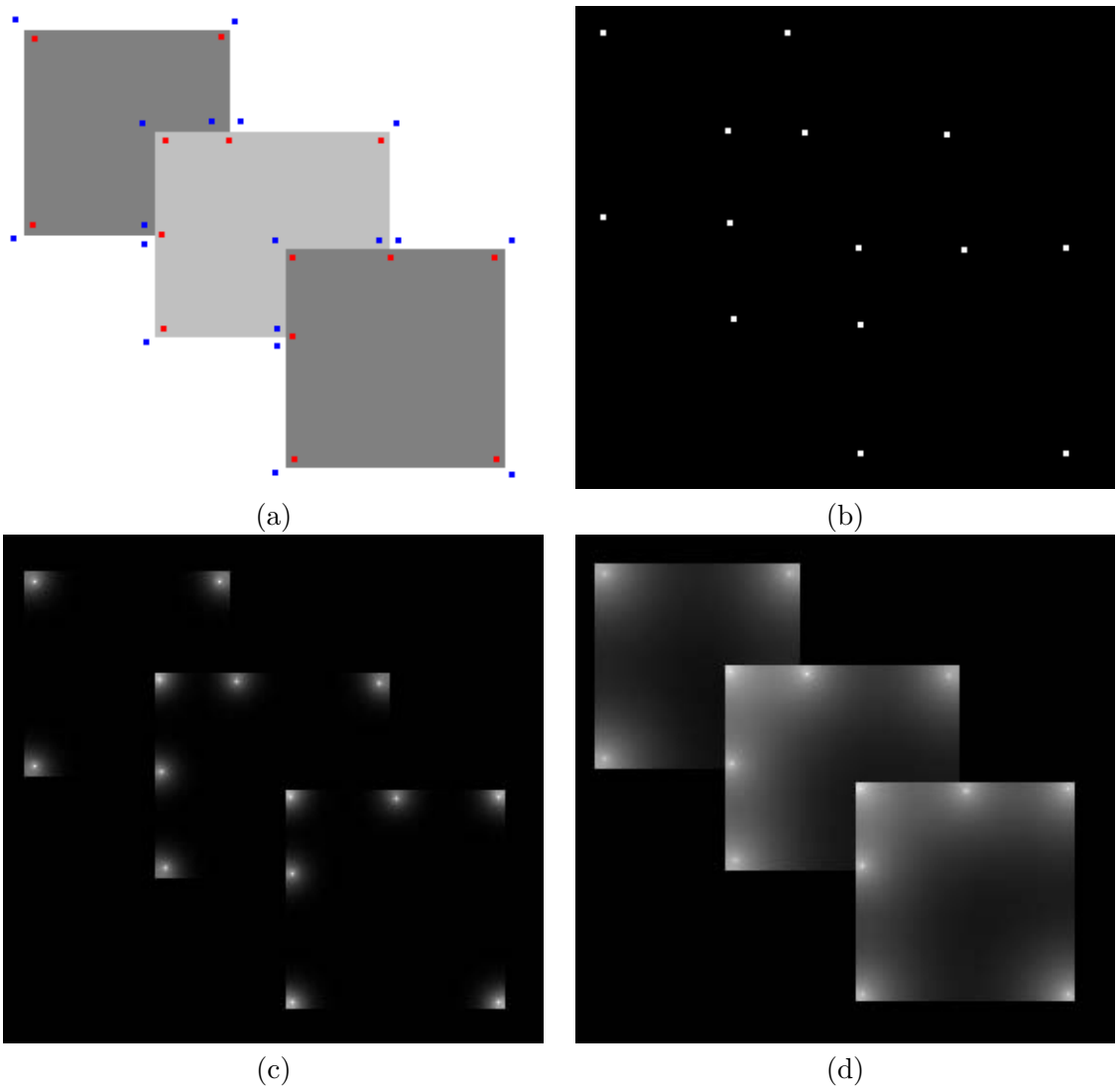


Figure 4.4: Example of depth diffusion by using equation 4.3. (a) Gray level image, where BSPs and FSPs are marked in blue and red respectively. (b) Depth image, where points corresponding to FSPs are initialized with a positive value (marked in white) and the rest of the image with value zero. (c) and (d) Depth images after an increasing number of iterations of the DDF.

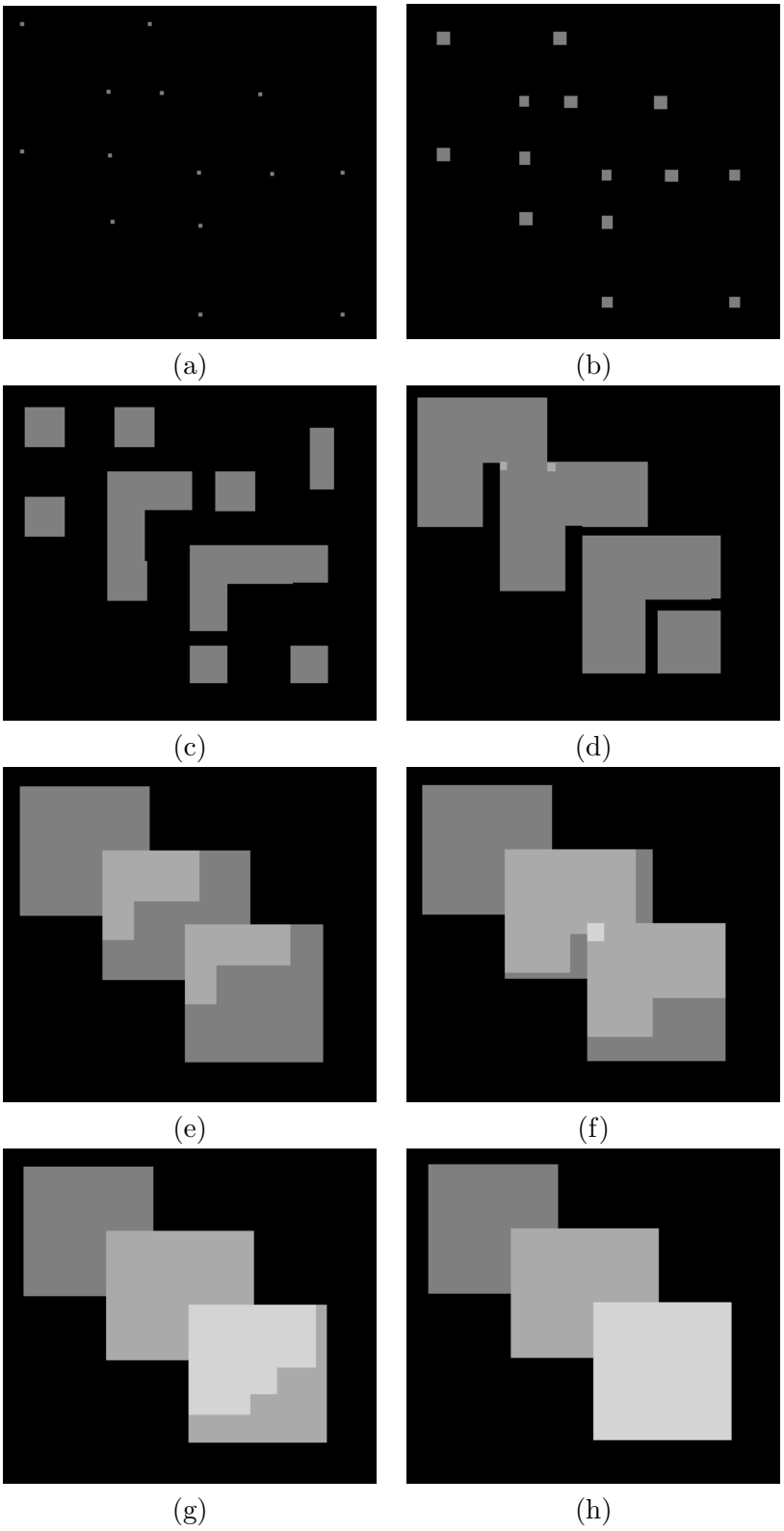


Figure 4.5: Example of depth diffusion by using equation 4.5. (a) Depth image, where FSPs have been initialized with a positive value (marked in gray) and the rest of the image with value zero. (b), (c) , (d), (e), (f), and (g) depth images corresponding to an increasing number of iterations. After each iteration, the depth difference between corresponding FSPs and BSPs is forced to be at least equal to the initial depth difference Δ, by adding Δ to FSPs when the difference between corresponding FSPs and BSPs is less than Δ.(h) Final depth image.

line passing through x. The second term is proportional to the second derivative of u in the direction perpendicular to the gradient $\xi = Du(x)^{\perp}/|Du(x)|$, which is orthogonal to the level line passing through x.

In the context of depth diffusion, we use a neighborhood filter to iteratively diffuse initial depth values $z(x)$ arisen from monocular depth cues to pixels which are similar in gray level (color). The idea of using the gradient of the original image rather than that of the depth image to define the neighborhood is not new since it has already been used by Yin et al.[Yin04] to improve the depth maps produced by stereo algorithms. By smoothing the depth map using the gray level (color) gradient of the original image instead of that of the depth map itself, the edge information is incorporated into the depth map, assuring that discontinuities in depth are consistent with gray level (color) discontinuities. Our work differs from that of Yin et al.[Yin04] in that the depth map is directly obtained by the diffusion process itself by applying it to some sparse and localized source points. FSPs are initialized with value 1, while the rest of the image is initialized with value zero. The depth diffusion filter (DDF) we propose is as follows:

$$DDF_{h,\rho}z(x) = \frac{1}{C(x)} \int_{B_\rho(x)} z(y) e^{\frac{-|u(x)-u(y)|^2}{h^2}} dy, \tag{4.3}$$

where $B_\rho(x)$ is a ball of center x and radius ρ, h is the filtering parameter which controls the decay of the exponential function, and

$$C(x) = \int_{B_\rho(x)} e^{\frac{-|u(x)-u(y)|^2}{h^2}} dy \tag{4.4}$$

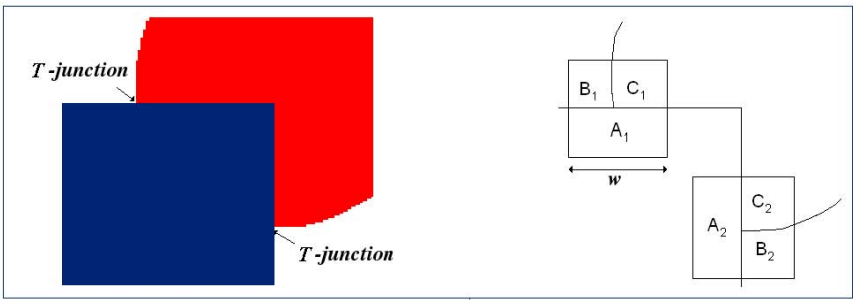
is the normalization factor. We set $\rho = 1$ and $h = 10$.

4.1.3 Internal boundary conditions

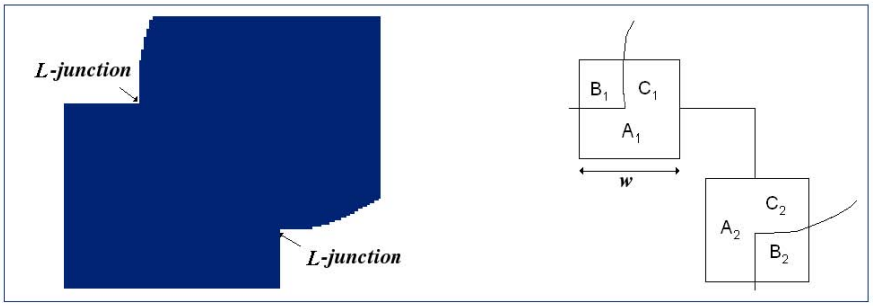
Equation 4.3 is applied iteratively until the stability is attained. After each iteration, the values of FSPs and BSPs are modified in order to hold at least the initial depth gradient. In practice, when the difference between the values of a FSP and the corresponding BSP became less than 1, the value 1 is added to the value of the FSP. This constraint corresponds to Neumann internal boundary conditions which are understood as a prespecified jump on the $c\frac{Dz}{Dn}$ as the boundary is crossed, where c is a positive constant and n is the normal to the boundary. This allows one to handle simple sorting when objects are located in multiple layers. Figure 4.4 is an example of the diffusion through the DDF. As can be sensed looking at Figures 4.4 (c) and (d), by taking an average of the depth values in a neighborhood, a very large number of iterations would be needed to attain the stability.

To make the diffusion process faster, the following equation is used

$$DDF_{h,\rho}z(x) = \sup_{B_\rho(x)} z(y) e^{\frac{-|u(x)-u(y)|^2}{h^2}} dy, \tag{4.5}$$

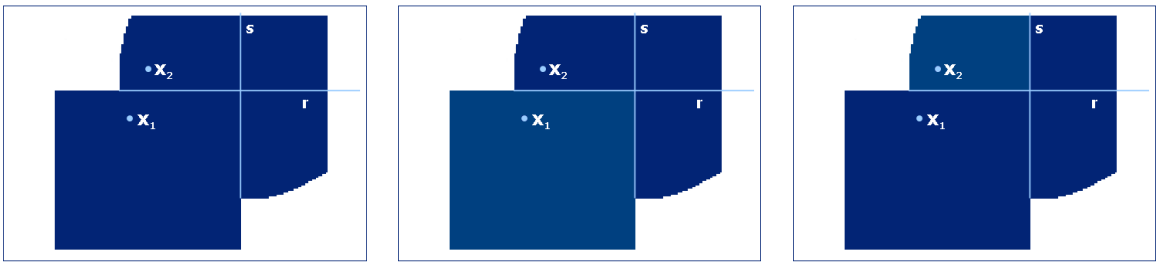


(a)



(b)

Figure 4.6: (a) Amodal completion: pairs of relatable regions (A_1,A_2), (B_1,B_2), and (C_1,C_2) have similar color. (b) Modal completion: pairs of relatable regions (A_1,A_2), (B_1,B_2), and (C_1,C_2) have similar color and the FSP (A_i) has similar color as one of the two BSPs (C_i).



(a) (b) (c)

Figure 4.7: (a) The points x_1 and x_2 belong to different objects, however their neighborhood has the same color. We change the way the neighborhood is defined so that: (b) The neighborhood of x_1 has to intersect the light blue region. (c) The neighborhood of x_2 has to intersect the light blue region.

while equation 4.3 is used only in the last iterations (see Figure 4.5).

In the case of occlusion and transparency there is also a depth order between the two regions separated respectively by the stem of the T and by the contour of the surface visible through the transparent object. However, occlusion and transparency cues do not carry any information about the partial order between the objects respectively in partial occlusion or in transparency and the background. This depth order can be inferred by other cues, such as convexity or visual completion. When information about this partial order is present, the depth gradient between one of the BSPs and the FSPs increases. This is the reason for which we force source points to hold "at least" the initial depth gradient. To deal with amodal completion, after each iteration, pairs of source points marking relatable regions (see Figure 4.6 (a)) are forced to maintain the same depth. In the case of modal completion, one of the two BSP has a gray level similar to the one of the FSP (see Figure 4.6 (b)). For this reason we modify the way the neighborhood is defined (see Figure 4.7). Let r and s be the lines the modal contours lie on. The neighborhood N_ρ is defined as follows: $N_\rho = \{y \mid y \in B_\rho(x), y \in \alpha_r(x), y \in \beta_s(x)\}$, where $\alpha_r(x)$ is the half image plane including x with origin the line r and $\beta_s(x)$ is the half image plane including x with origin the line s.

4.1.4 Experimental Results

We tested our model on a set of real images (taken by a digital camera) involving occlusion, transparency, convexity, visual completion (both amodal and modal) and self-occlusion. In Figures 4.8 and 4.9 we show for each experiment four images: the original image (see Figures 4.8, 4.9, and 4.10 (a_i)); the image showing the segments found by applying LSD on the original image (see Figures 4.8, 4.9, and 4.10 (b_i)); the image where the initial depth gradient at depth cue points is represented through vectors pointing to the region closer to the viewpoint (red vectors arise from T-junctions, green vectors arise from local convexity and each of them represents the point having the biggest curvature value of the connected components obtained by thresholding the curvature) (see Figures 4.8, 4.9, and 4.10 (c_i)); the depth image obtained by performing the proposed method (see 4.8, 4.9, and 4.10 (d_i)). The depth map is rendered through gray level values (high values indicate regions that are close to the camera). In the example on the first column of Figure 4.8, local convexity induces to see the disk over the table. In the second column (see Figure 4.8), there is an example involving convexity and occlusion: it shows that the proposed method is able to handle simple sorting in presence of multiple depth layers. In the example on the third column (see Figure 4.8), occluding contours have different depth relationships at different points along its continuum. However, the proposed method performs well also in this ambiguous situation. In the first and the second columns of Figure 4.9 there are examples of amodal and modal completion respectively: in the former case, the detection of pairs of relatable T-junctions leads to see the green piece of paper partially occluded by the white strips

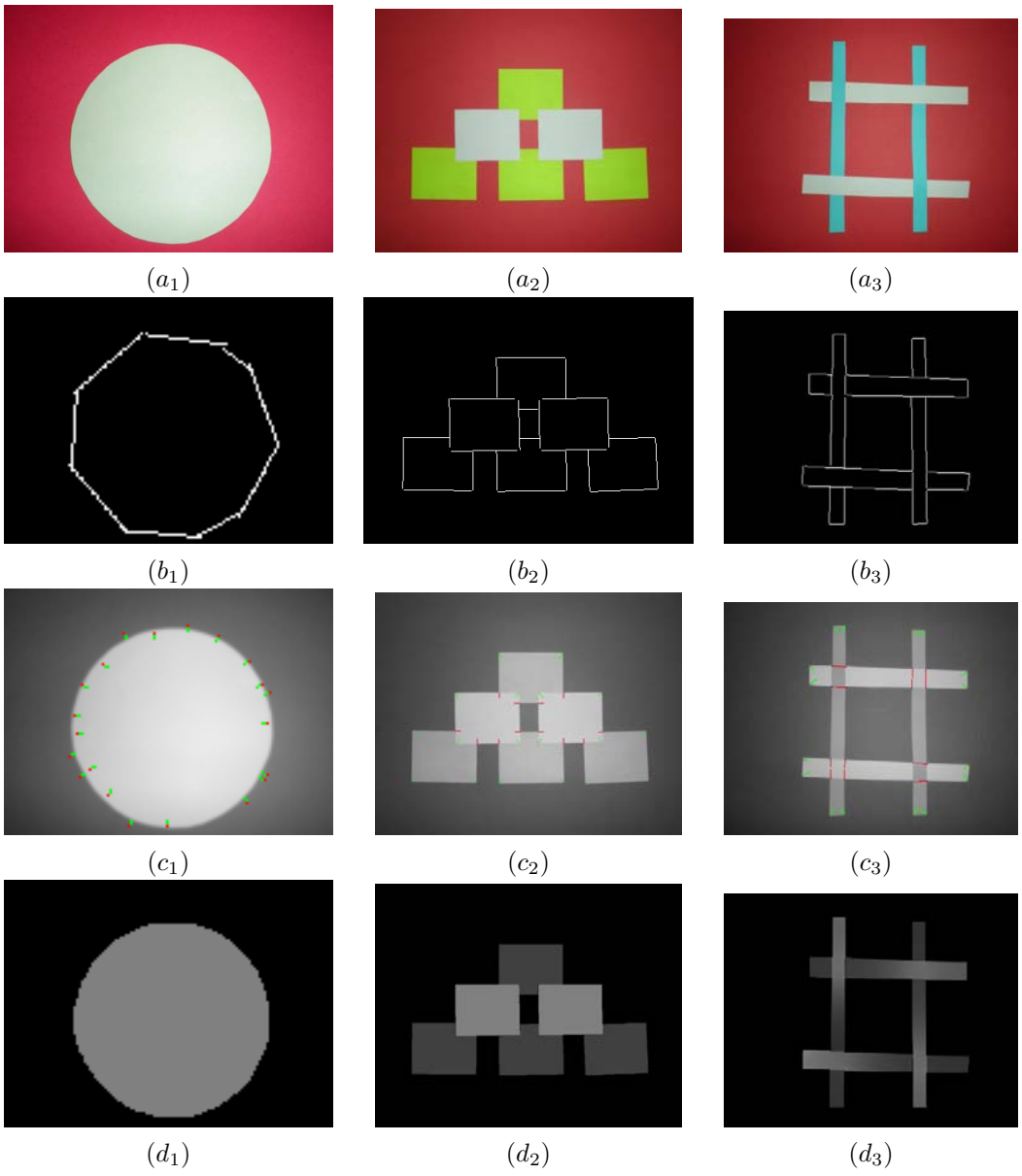


Figure 4.8: (a_i) Original image. (b_i) Segments detected by LSD. (c_i) Local depth cues are represented through vectors that point to the region closer to the viewpoint. (d_i) Depth image.

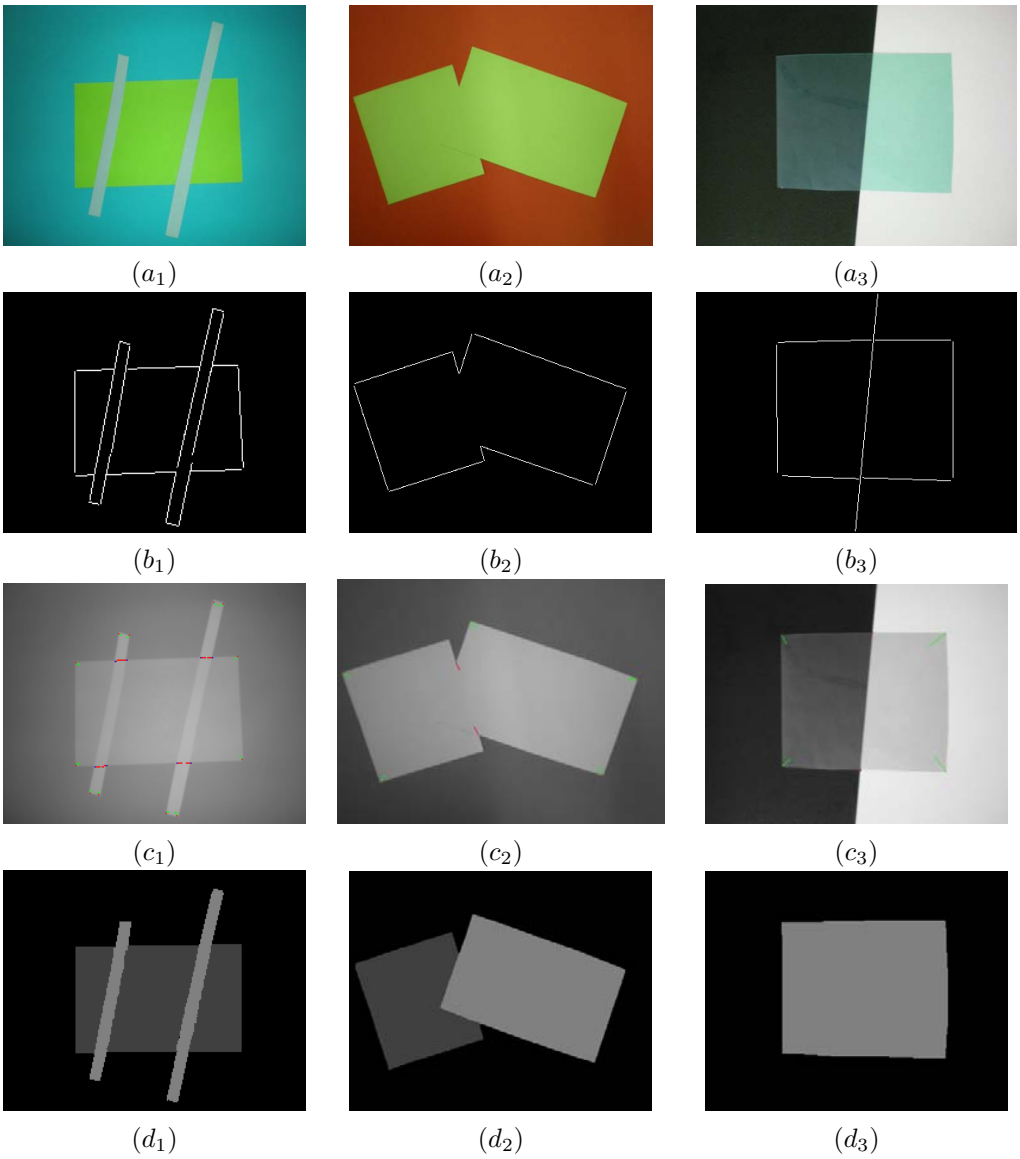


Figure 4.9: (a_i) Original image. (b_i) Segments detected by LSD. (c_i) Local depth cues are represented through vectors that point to the region closer to the viewpoint. (d_i) Depth image.

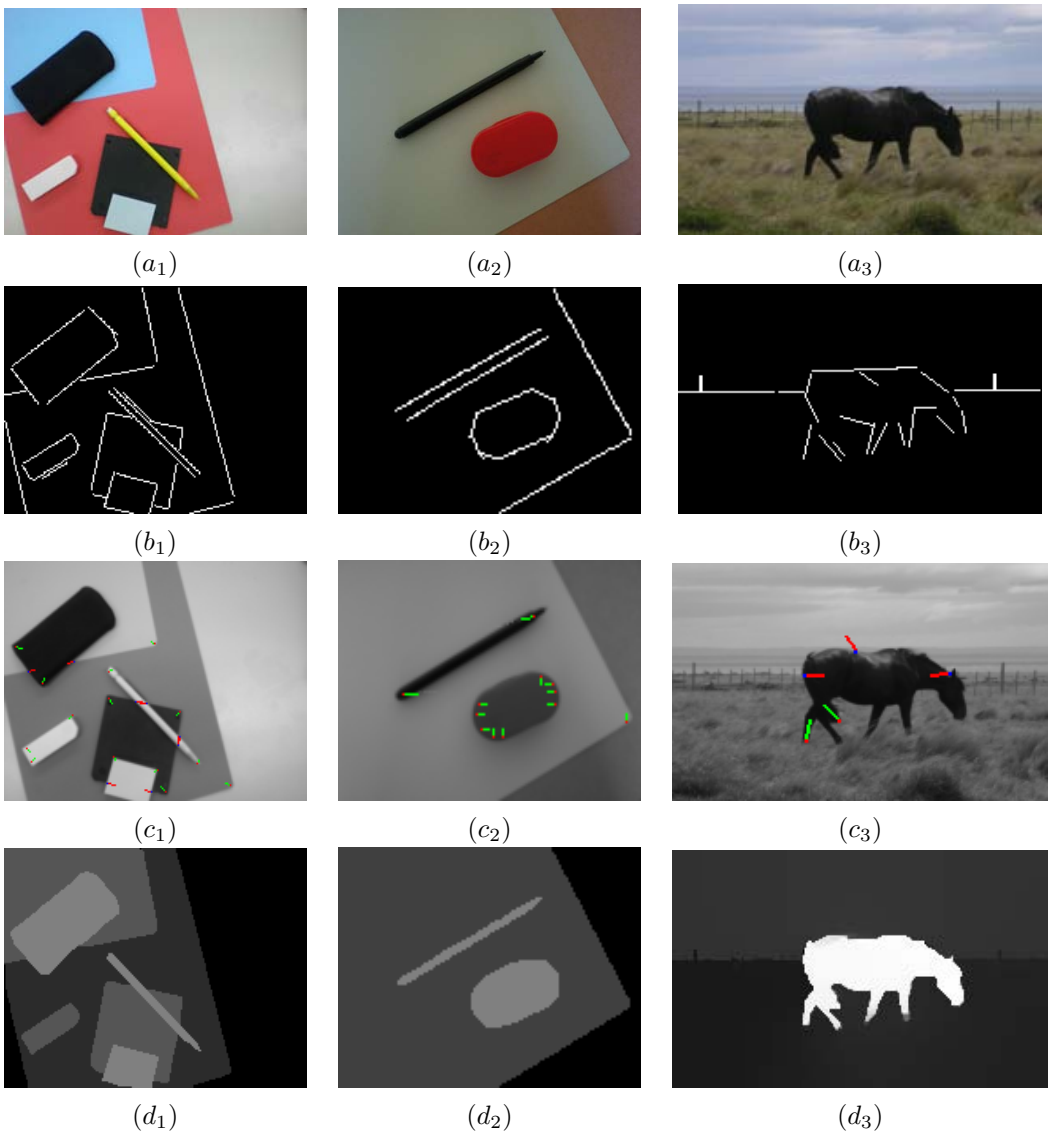


Figure 4.10: (a_i) Original image. (b_i) Segments detected by LSD. (c_i) Local depth cues are represented through vectors that point to the region closer to the viewpoint. (d_i) Depth image.

as a meaningful unit; in the latter case, the detection of a pairs of degenerated T-junctions leads to see the rectangle in front of the square. In the example on the third column, the transparency phenomenon is correctly interpreted. On the first and the second columns of Figure 4.10 there are examples of indoor scenes, for which a proper solution is obtained. Instead, on the third column there is an example of outdoor scene involving a conflict. The T-junction detected on the back of the horse is due to a reflectance discontinuity and its local depth interpretation is incorrect. However, on the depth map, the shape of the horse appear clearly on the foreground since the diffusion process allowed to overcome the local inconsistency.

4.2 Region-merging based framework

From chapter 2 it has emerged that one of the means by which depth is organized into perceptual units is the CDAP (see section 2.1.2.2). This principle expresses a constraint on the relationship between signal contrast and perceived depth. When more than one signal contrast is present, the CDAP provides a strong constraint to determine a global, consistent depth interpretation from local image data. By local image data we refer to a color or to a contrast discontinuity to which a local depth gradient has been assigned. As it has been shown in section 2.1.2.2, any single local signal discontinuity has a variety of possible depth interpretations. However, this ambiguity may be drastically reduced in presence of depth cues. For example, consider the image in Figure 4.11 (a): it contains a certain number of T-junctions (see Figure 4.11 (b)). In correspondence of each T-junction, the region delimited by the roof appears to be in front to the regions delimited by the stem. By extending the branch of each T-junction following the depth discontinuity, which corresponds to the color discontinuity, the image can be partitioned in a limited number of regions (see Figure 4.11 (c)). Due to the presence of T-junctions, the region A appears to be in front of the regions B, D and C; the region B appears to be in front of the regions F, E and G; region D appears to be in front of the regions E and G; finally, region C appears to be in front of the region D and G. By applying the CDAP to these depth assignments, a global depth interpretation can be inferred (see Figure 4.11 (d)). This simple demonstration confirms that T-junctions, and in turn occlusion, can be used to successfully assign depth in images.

However, it needs to be taken into account that the local interpretation of T-junctions is not always consistent with the global depth intepretation, as it happened in the example analyzed above. Indeed, T-junctions do not arise solely from occlusion (see Figure 4.12 (a)) but they may also be the result of a reflectance discontinuity (see Figure 4.12 (b)). In addition, in an automatic framework, false positive detections of T-juntions may also be the cause of a misleading local depth assignment. As a consequence, an algorithm that uses only the cue of occlusion to infer a global depth ordering must envisage a mechanism to solve possible conflicting intepretations. Another issue that needs to be addressed is that depth information provided by

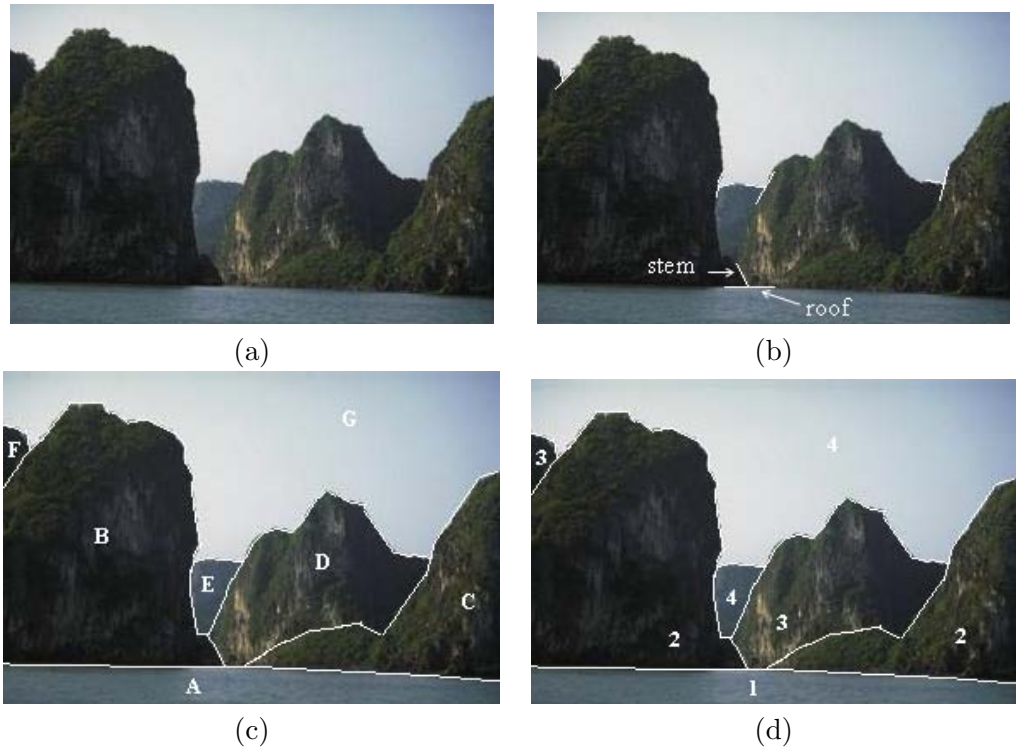


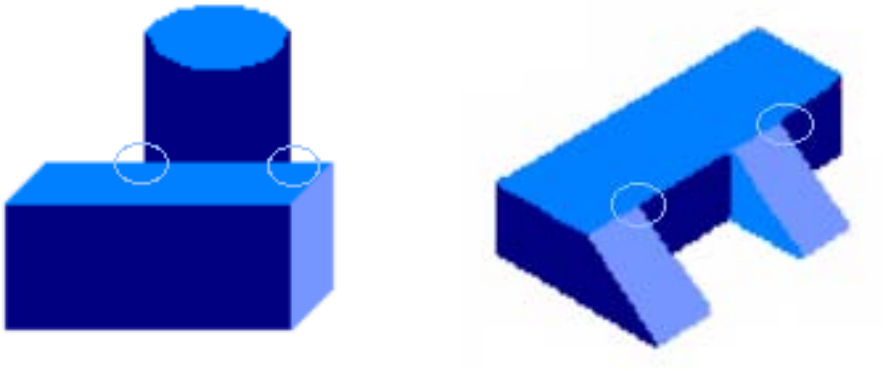
Figure 4.11: (a) An image of a natural scene. (b) T-junction branches are marked in white.(c) Result of extending the T-junction branches following the color discontinuities.(d) Smaller numbers correspond to regions closer to the viewpoint.

T-junctions is limited to a small neighborhood of the T-junction centers. Therefore, to reason globally about depth relationships, the T-junction branches need to be propagated following the depth discontinuities, as it has been done in Figure 4.11 (c). Inspired by the CDAP and taking into account the above mentioned issues, we have developed an automatic algorithm that uses only the cue of occlusion to infer global, consistent depth ordering.

The proposed strategy involves three main steps. First, occlusion relations signaled by T-junctions are detected; second, the image is segmented by using a BPT-based statistical region-merging algorithm which preserves the previously detected T-junctions; third, the depth relations between the regions of the final partition are encoded through a Directed Graph (DG). This formalization allows to easily detect and solve possible conflicting interpretations leading to a global depth ordering. In the following, the last two steps are detailed.

4.2.1 Segmenting the image preserving T-junctions

This section details how to obtain a BPT-based segmentation of the image, which preserves the previously detected T-junctions. Recall that a BPT is a structured representation of a set of hierarchical partitions in which the finest level of detail is given by the initial partition. The set



(a) (b)

Figure 4.12: (a) T-junctions corresponding to depth discontinuities. (b) T-junctions corresponding to reflectance discontinuities.

of the regions of the initial partition may coincide with the set of image pixels or the partition of flat zones, or any other set of regions obtained using any other pre-computed partition. The general region merging algorithm used for creating a BPT requires the specification of the *region model*, *merging order*, and *merging criterion*.

- *Region model*: the region model defines how to represent each region. When two regions are merged, the region model defines how to model its union and what are the main characteristics that should be kept in order to continue the merging process. For instance, if objects are assumed to be homogeneous in color, a region model based on the mean color of each region may be used or, if objects are assumed to be generated by the same probability distribution, a region model based on the color histogram of each region may be used.

- *Merging order*: the merging order determines the order in which pairs of neighboring regions (links) are processed. It is a real valued function of each pair of neighboring regions and is usually based on a similarity criterion between the region models. Each time a link is processed its associated nodes, which correspond to the neighboring regions, are merged together. The merging order is closely related to region model. As the region model, the merging order is related to the notion of objects, that is to the notion of homogeneity with respect to a defined property. It can be seen as a measure of the likelihood that two neighboring regions belong to the same object. For instance, if objects are assumed to be homogeneous in color, a similarity measure based on the color difference should be used or, if objects are assumed to be generated by the same probability distribution, a similarity measure based on the color histogram should be used.

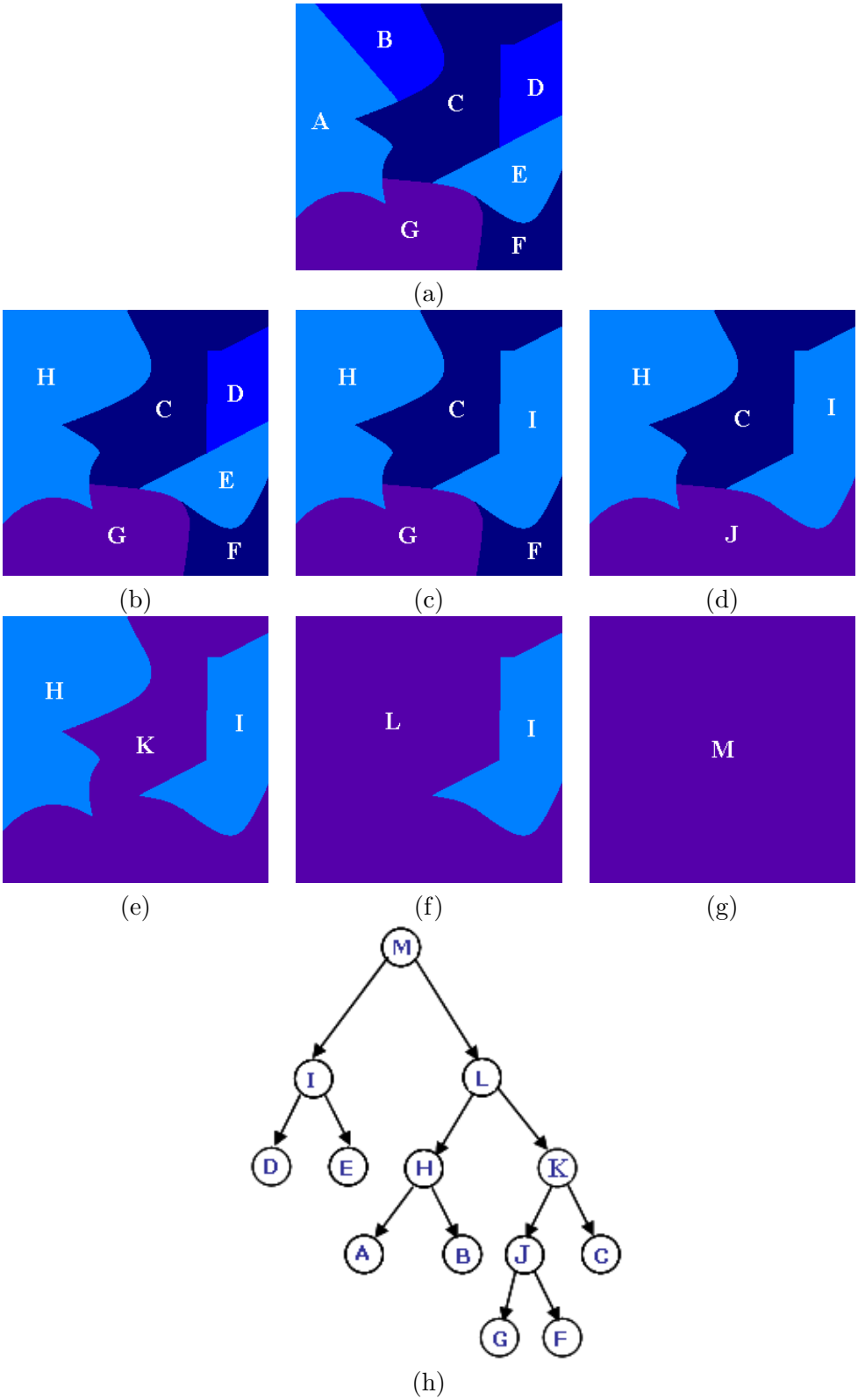


Figure 4.13: Example of BPT creation with a region merging algorithm. (a) Initial partition. (b) Merging step 1. (c) Mering step 2. (d) Merging step 3. (e) Merging step 4. (f) Merging step 5. (g) Merging step 6. (h) BPT representation of the full region merging process.

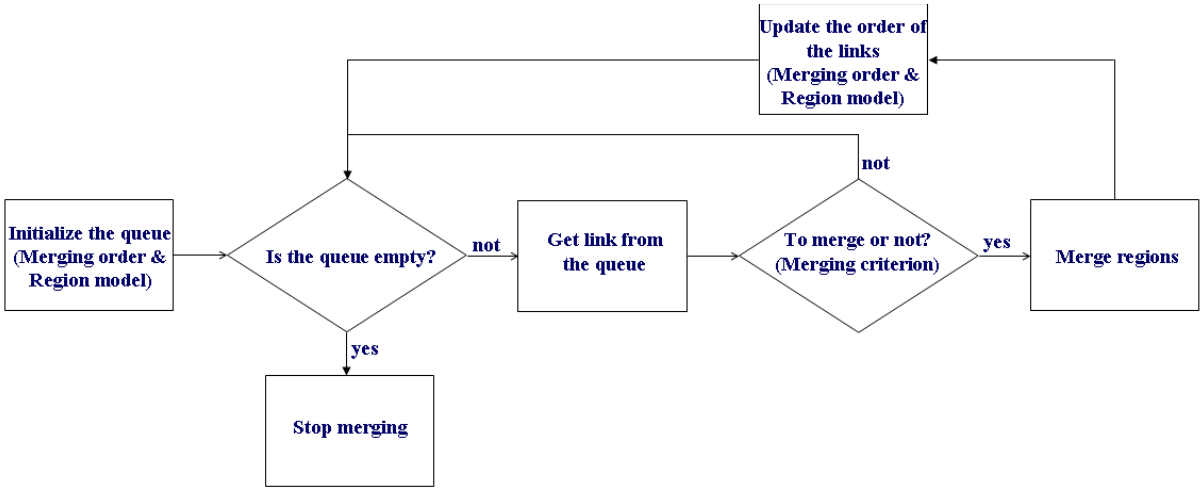


Figure 4.14: Block diagram of the general region merging algorithm.

- *Merging criterion*: each time a link is processed, the merging criterion decides if the merging has actually to be done or not. It is binary valued function (merge or do-notmerge) of each pair of neighboring regions. The merging criterion allows to decide which of the mergings proposed by the merging order should be really done. An example of simple merging criterion is the number of regions. This merging criterion does not modify the order (proposed by the merging order) in which links are processed, but simply acts as a termination criterion. Instead, more complex merging criteria may be used to control more precisely the way regions are merged, acting as a sieve among the set of mergings proposed by the merging order. An example is the maximum size of each region: if the new region resulting from the merging of a pair of neighboring regions proposed by the merging order has a size above the threshold imposed by the merging criterion, the merging is skipped. The merging criterion is usually used in segmentation algorithms to automatically determines the number of regions of the final partition.

Figure 4.13 is an example of BPT created by using the general scheme of the merging process represented in Figure 4.14. In each block of this scheme, the merging notion involved (region model, merging order, and merging criterion) is indicated in parentheses. The initial partition (see Figure 4.13 (a)) is represented through a Region Adjacency Graph (RAG), which encodes neighborhood relations between regions. Then, each region is initialized by computing its model and, for each pair of neighboring regions, the merging order is computed using the model of the associated regions. Depending on the homogeneity value determined by the merging order, the links (representing pairs of neighboring regions) are inserted into a hierarchical queue, used to index and process the links according to its merging order. The link with highest priority

is extracted from the queue and the merging criterion decides whether the neighboring regions represented by the link have to be merged or not. If not, the link is removed from the queue so that the corresponding pair of regions will never be merged and the next link with highest priority is extracted. Instead, if the neighboring regions represented by the link are merged, the RAG structure is updated, the region model of the new region is computed, the neighboring links of the two merged regions are extracted from the queue and the values of their merging order is updated. At this point the iterative process starts again by checking if the queue is empty. The merging process ends when the hierarchical queue is empty. As a result, a BPT is obtained, whose leaves (see Figure 4.13(h)) represent regions that belong to this initial partition (see Figure 4.13 (a)), whose root node represents the entire image support, and whose remaining nodes are associated to regions that represent the union of two children regions.

In this work, we start from the initial partition of all image pixels and we model each pixel statistically by a probability distribution obtained as described in section 3.1.3.2. Hence, our region model relies on the assumption of self-similarity, introduced in section 3.1.3.2, which assumes that the image is a fairly general stationary random process. This assumption is actually more general and more accurate than any other general regularity assumption, which considers objects to be homogeneous in color or texture. As merging order we use a statistical similarity measure between statistical region models proposed by Calderero et al. [Cal08]. The authors proposed two different statistical merging orders: the KL merging order and the Batthacharyya (BHAT) merging order. The KL merging order, that has been already introduced in section 3.1.3.2, is based on measuring the similarity between the probability distributions of the regions and the probability distribution of their merging, weighted by the size of the regions.

KL area-weigthed merging order:

$$KL_{area}(R_i, R_j) = -n_i \cdot D_{KL}(P_i \parallel P_{i\cup j}) - n_j \cdot D_{KL}(P_j \parallel P_{i\cup j}), \tag{4.6}$$

where R_i and R_j are two adjacent regions of size n_i and n_j and statistical distribution P_i and P_j respectively, whose union would generate a new region, $i \cup j$, with empirical distribution $P_{i\cup j}$, and D_{KL} is the Kullback-Leibler divergence [Kul51] between two statistical distributions. The Batthacharyya (BHAT) merging order is based on a size-weighted direct statistical measure of the probability distributions. This criterion leads to merge pairs of adjacent regions with the maximum probability of fusion.

BHAT area-weigthed merging order:

$$BHAT_{area}(R_1, R_2) = \arg \max_{(R_i, R_j)} -min(n_i, n_j) \cdot B(P_i, P_j), \tag{4.7}$$

where

$$B(P_i, P_j) = -log(\sum_x P_i^{1/2}(x) P_j^{1/2}(x)) \tag{4.8}$$

is the Bhattacharyya distance [Bha43]. In both cases, the merging cost depends on the size of the regions. The size term assures that the resulting partitions are size consistent, meaning that the area of the regions tends to increase as the number of regions into the partition decreases. However, this dependence favors the fusion of small regions, delaying the fusion of larger regions. Indeed, small regions cause less significant errors since the error contribution of the union of two small regions is small compared to the contribution resulting from the merging with a large region. As the fusion of large regions is delayed, area weighted merging orders suffer generally from oversegmentation. To provide a trade-off between undersegmentation and oversegmentation, the corresponding area unweighted version of the KL and BHAT merging orders have also been proposed [Cal08]. They are as follows:

KL area-unweigthed merging order:

$$KL_{noarea}(R_i, R_j) = -D_{KL}(P_i \parallel P_{i\cup j}) - D_{KL}(P_j \parallel P_{i\cup j}), \tag{4.9}$$

BHAT area-unweigthed merging order:

$$BHAT_{noarea}(R_1, R_2) = \arg \max_{(R_i,R_j)} \cdot B(P_i, P_j). \tag{4.10}$$

From our practical experience, the merging order which allows to define more precisely self-similar regions is the BHAT area-unweighted criterion.

The advantage of the proposed region model, is illustrated in Figure 4.15. The image to be segmented (see Figure 4.15 (a)) has been generated by smoothing the original bicolor image, having dynamic range in $[0, 1]$, with a Gaussian impulse response of size 25×25 and standard deviation 3, and by adding white noise of mean zero and variance 0.001. The partitions in Figure 4.15 (b_i), (c_i), (d_i) and (e_i) have been obtained by using respectively the area weighted KL merging order, the area unweighted KL merging order, the area weighted BHAT merging order, and the area unweighted BHAT merging order. The partitions on the second row (see Figures 4.15 (b_1), (c_1), (d_1), and (e_1)) have been obtained by modeling each pixel deterministically, by its color value, while the partitions on the third and fourth row by modeling each pixel through a probability distribution defined as in section 3.1.3.2. Experimentally, we have found that, to segment the whole image, for a similarity window of 3×3 and a search window of size 15×15, good values for the filtering parameter h and the number of bins for the luminance component are respectively 50 and 50. As merging criterion, the number of regions has been used, two for the partition on the second row and three for the partition on the third row. As can be observed (see Figure 4.15 (b_2), (c_2), (d_2), (e_2)), the proposed pixel modeling drastically improves the contour definition. The reason is that pixels on the transition zone between two homogeneous regions are grouped together even in presence of noise. Figures 4.15 (b_3), (c_3), (d_3), (e_3) shown the result of segmenting the image in three regions. As can be observed, the third region corresponds to the transition zone between the green and blue regions.

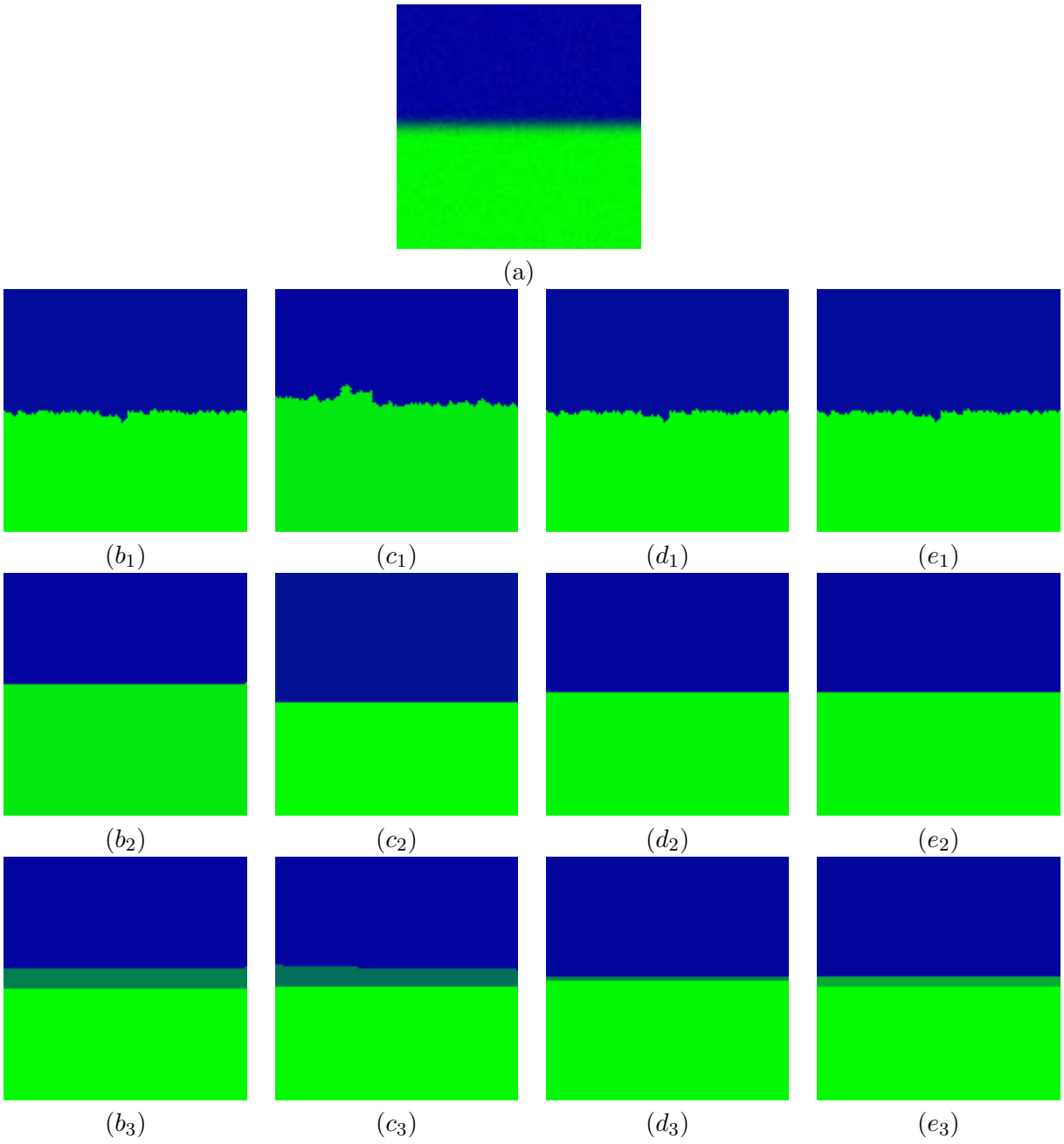
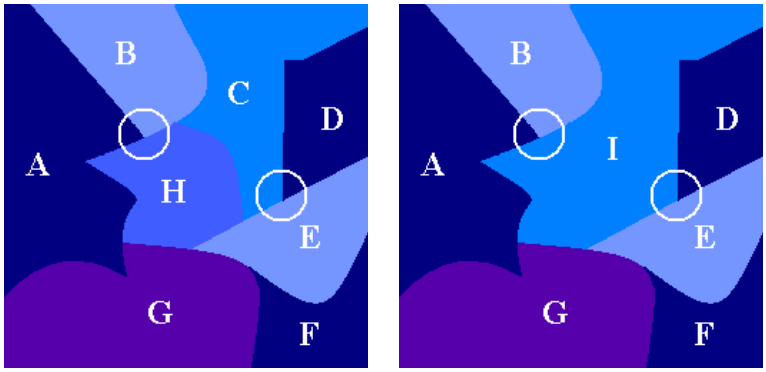


Figure 4.15: (a) Original image to be segmented. Images segmented by using the: (b_i) Area weighted KL merging criterion. (c_i) Area unweighted KL merging criterion. (d_i) Area weighted BHAT merging criterion. (e_i) Area unweighted BHAT merging criterion. The partitions on the second row have been obtained by modeling each pixel deterministically by its color value, whereas the partitions on the last two rows have been obtained by modeling each pixel statistically by its pdf. The difference between the set of partitions with index 2 and the set of partitions with index 3 is that they have been obtained using a number of regions equal to 2 and 3 respectively as merging criterion.



(a) (b)

Figure 4.16: (a) Example illustrating the concept of incompatibility in the merging process. (a) T-junctions are marked by white circles. The regions A, B, and C are incompatible with each other as well as the regions D, E, and C.(b) The regions C and H have been merged to form the region I. The region I is incompatible with A, B, D, and E.

In order to preserve T-junctions, we introduce the concept of *incompatibility*. Two regions are said *incompatible* if they are involved in an occlusion relation and therefore are supposed to belong to different levels of depth. When two regions are incompatible, they cannot be merged. Hence, the concept of incompatibility is used as merging criterion: if the pair of neighboring regions proposed by the merging order are incompatible, the proposed merging is skipped. Incompatibility is an inheritable property. In Figure 4.16, the regions A, B and H are incompatible with each other as well as the regions C, D, E. When the regions C and H are merged to form the region I, all incompatible relations in which C and H are involved, are inherited by I. As a consequence, the region I becomes incompatible with the regions A, B, D, and E. The region-merging process terminates when all regions became incompatible, and therefore no more mergings are allowed. In Figure 4.17 the results of applying the region merging algorithm described in this section without preserving T-junctions and using the number of regions as merging criterion are shown. Both, the KL and the BHAT merging orders, in their weighted and unweighted versions have been used. As can be observed in all four cases (see Figure 4.17 (c), (d), (e), (f)), the 10 regions of the final partition do not correspond to the 10 most perceptually meaningful regions. Instead, when incompatibility is used as merging criterion, T-junctions are preserved and the result of the segmentation is correct (see Figure 4.17 (b)). This improvement is attributable to the fact that the process of grouping by statistical-similarity through the merging order is somehow corroborated by the process of separating by depth-dissimilarity through the merging criterion, being both processes treated under an unified region-merging framework. Indeed, the merging order proposes a merging based on a statistical similarity between the region models and the merging criterion validates or not the proposed merging depending on if the pair of neighboring regions involved belong to the same level of depth or not, that is if they are or not compatible.

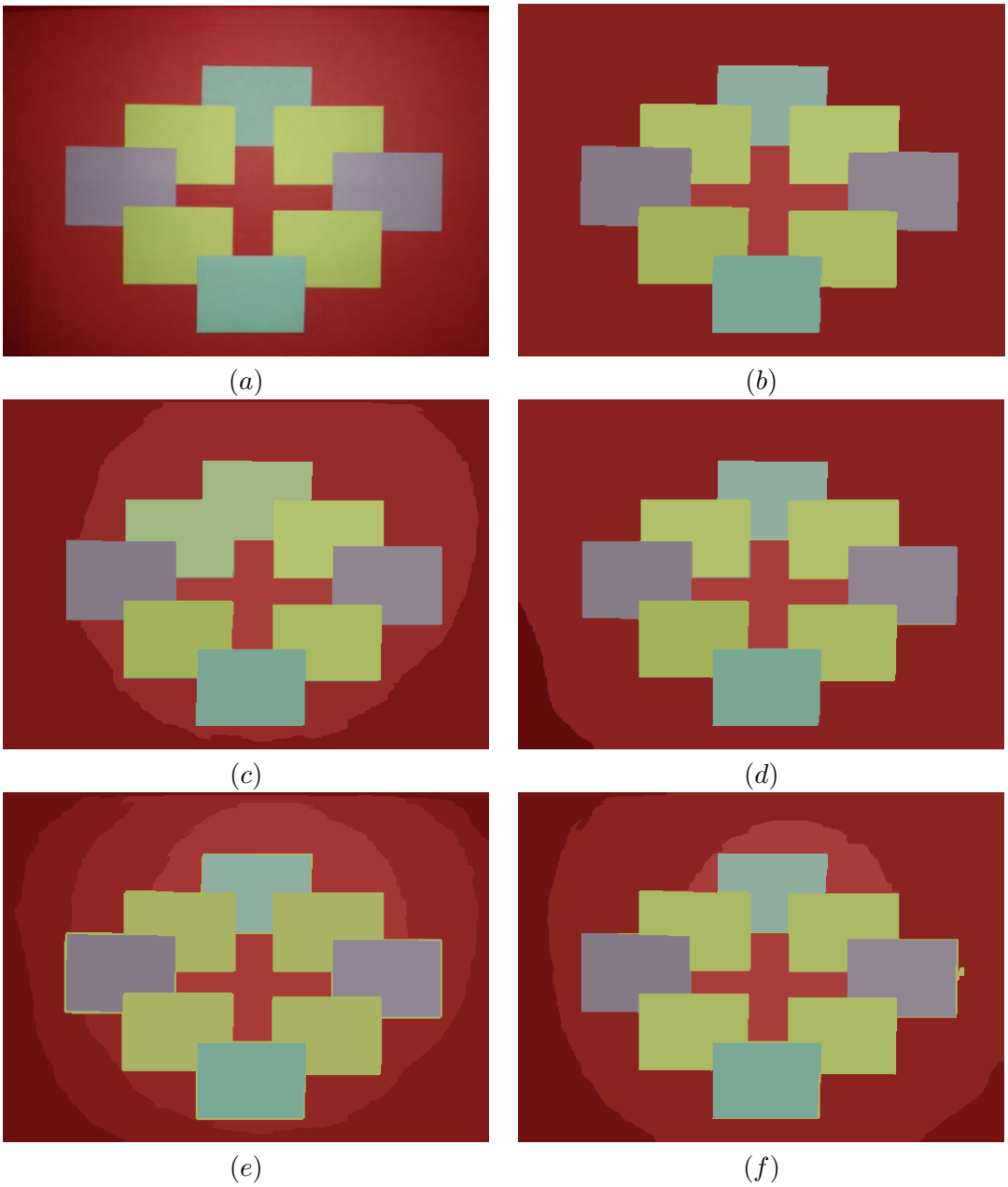


Figure 4.17: (a) Original image to be segmented. (b) Segmentation obtained preserving T-junctions, using the BHAT unweighted merging order. Segmentation results obtained by using the statistical region merging algorithm without preserving T-junctions, by using the: (c) Area weighted BHAT merging order. (d) Area unweighted BHAT merging order. (e) Area weighted KL merging order. (f) Area unweighted KL merging order.

The following section presents a method to infer a global, *consistent* depth ordering between regions of the final partition.

4.2.2 Graph formalization and reasoning

As stated in the previous section, the regions of the final partition are incompatible. For each triple of incompatible regions arisen from an occlusion relation, a depth assessment based on the depth interpretation of T-junctions can be done: locally, the region delimited by the roof of the T appears to be in front of the ones delimited by the stem. However, it needs to be taken into account that the depth interpretation of pairs of T-junctions that share an edge may give rise to an inconsistency. In Figure 4.18 (a) there is an example of first order *inconsistency* (involving a pair of T-junctions). Region C is in front of region A for one T-junction, while the converse is true for the other. Higher order inconsistencies involve more than two T-junctions. We formalize the problem of finding a global, consistent depth interpretation through a Directed Graph (DG). A DG is specified by $DG = (V, E_W, W)$, where V is a set of nodes, E_W is a set of edges and W is the matrix of weights attached to the edges. In our formalization, each node represents a region of the final partition and each directed edge represents the relative depth relation between two regions. Edges are specified as ordered pairs: an edge $e = (X, Y) \in E$ is considered to be directed from X to Y meaning that the region X is in front of the region Y. The weight attached to each edge corresponds to the number of occurrences the depth relation represented by the edge has been inferred from different occlusion relationships. For instance, in Figure 4.18, the weight of the edge $e = (C, A)$ is 2, whereas the weight of the edge $e = (A, C)$ is 1. With this formalization, local constraints are allowed to propagate along the graph and the search for inconsistent pairs of T-junctions is reduced to the search of cycles on the DG (dashed thick red arrows in Figure 4.18(b)). The search for directed cycles is performed by a Depth-First Search (DFS) algorithm [Cor01]. This algorithm may be computationally expensive when the number of nodes involved is high. However, being usually the number of regions of the final partition small, the corresponding computational load is moderate. Inconsistencies are solved by suppressing the edge(s) on the cycle with lowest cost. Since the depth relation associated to the edge with the lowest cost is considered unreliable, the other edge (dashed thin blue arrow in Figure 4.18(b)) associated with the T-junction from which the unreliable depth relation arises is also removed. As a result, a DAG is obtained (see Figure 4.18 (c)). Each DAG gives rise to a *partial order* $\leqslant$ on the set of it nodes V. A relation $\leqslant$ is a partial order on a set V if it satisfies the properties of

- Reflexivity:$X \leqslant X \, \forall \, X \in V$
- Antisymmetry: $(X \leqslant Y, Y \leqslant X) \longrightarrow X = Y$
- Transitivity:$(X \leqslant Y, Y \leqslant Z) \longrightarrow X \leqslant Z$

Many different DAGs may give rise to the same *reachability relation*. The reachability relation of a DAG is the set of all ordered pairs (X, Y) of nodes in V for which there exist nodes $v_0 = X, v_1, \ldots, v_d = Y$ such that $e = (v_i - 1, v_i) \in E \forall 1 \leqslant i \leqslant d$. The reachability relation of a DAG is also called *transitive closure* and corresponds, among all DAGs which give rise to the same reachability relation, to the one with the maximum number of edges. Instead, the DAG, among all the DAGs which give rise to the same reachability relation, with the minimum number of edges is called *transitive reduction*. The transitive reduction of a finite DAG is obtained by removing redundant edges while maintaining identical reachability properties. An edge $e = (X, Y)$ is said *redundant* if there exists a path from X to Y that does not contain the edge. For example (see Figure 4.18 (c)), the edge $e = (G, A)$ is redundant since it is possible to go from the node G to the node A passing through the node H. The graphical rendering of a transitive reduction is called Hasse diagram. Each element of the DAG is drawn on the Hasse diagram as a node and line segments are drawn between these nodes according to the following two rules:

- If $X \leqslant Y$, then the node corresponding to X appears lower in the Hasse diagram than the point corresponding to Y.
- The line segment between the points corresponding to any two nodes X and Y of the set V, is included in the Hasse diagram as a line segment that goes upward from x to y if and only if, $X \leqslant Y$ and there is no Z such that $X \leqslant Z \leqslant Y$ (the edge $e = (X, Y)$ is not redundant).

Any Hasse diagram uniquely determines a partial order, and any finite partial order has a unique transitive reduction.

In our formalization, the Hasse diagram corresponding to the transitive reduction of the DAG is exactly the depth map (see Figure 4.18 (d)). Since there is no depth order between the regions forming the stem of a T-junction, they appear on the Hasse diagram as as leaves (A and B), without any information about their respective depth, unless of course, an order between them can be inferred by other T-junctions.

4.2.3 Experimental results

We tested our algorithm on a set of real images. For each experiment we show four images: the original image (see Figure 4.19(a)); a gray level version of the original image, where the T-junctions detected by using the region-merging algorithm detailed in section 3.1.3.2 are represented through a vector pointing to the region closer to the viewpoint (see Figure 4.19(b)); the segmented image (see Figure 4.19(c)); the map of relative depths, which is rendered as a gray level image (high values indicate regions closer to the viewpoint) (see Figure 4.19(d)). As

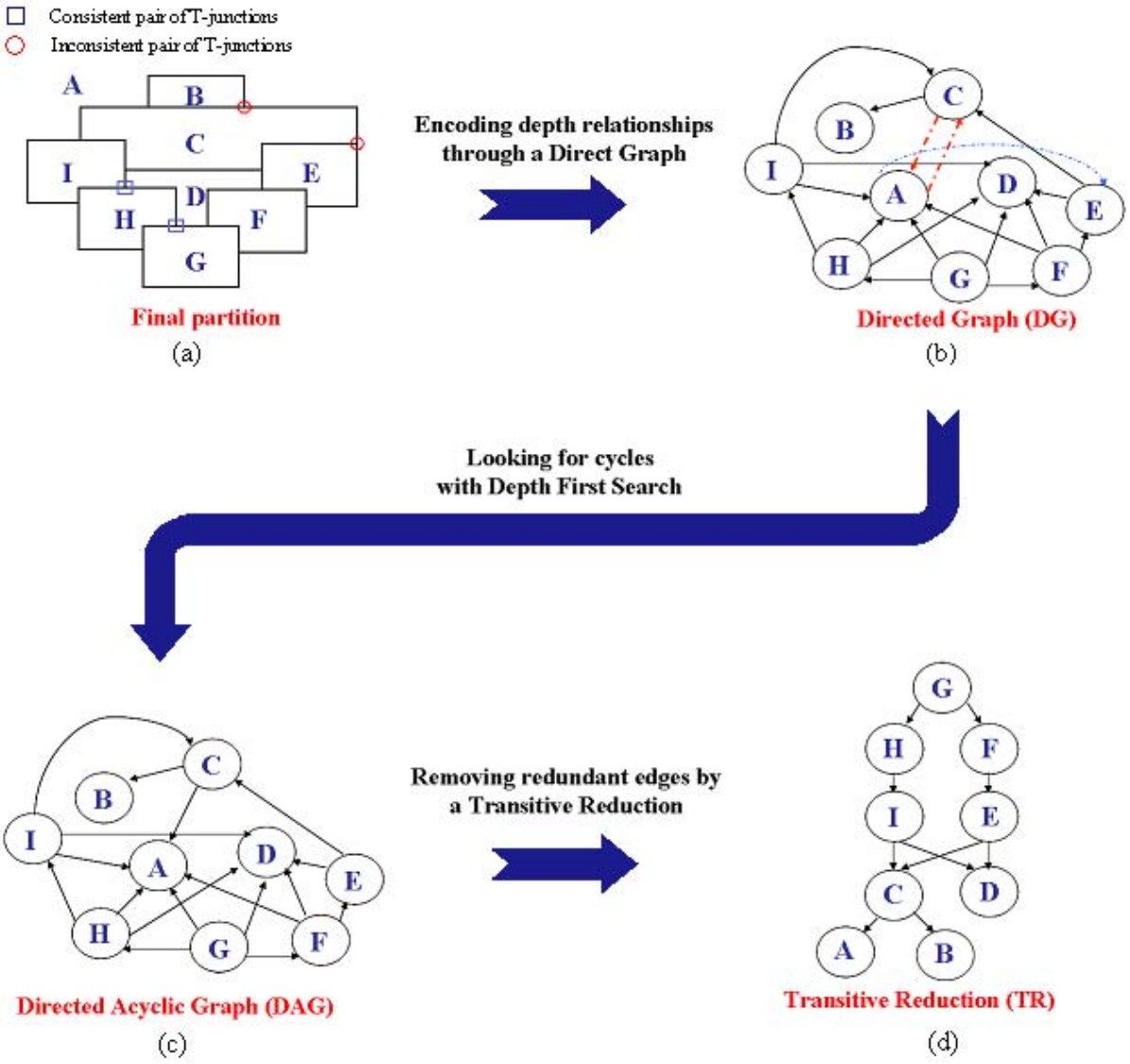


Figure 4.18: (a) Partitioned image. (b) Associated DG. (c) Associated DAG. (d) Hasse diagram resulting from the transitive reduction of the DAG.

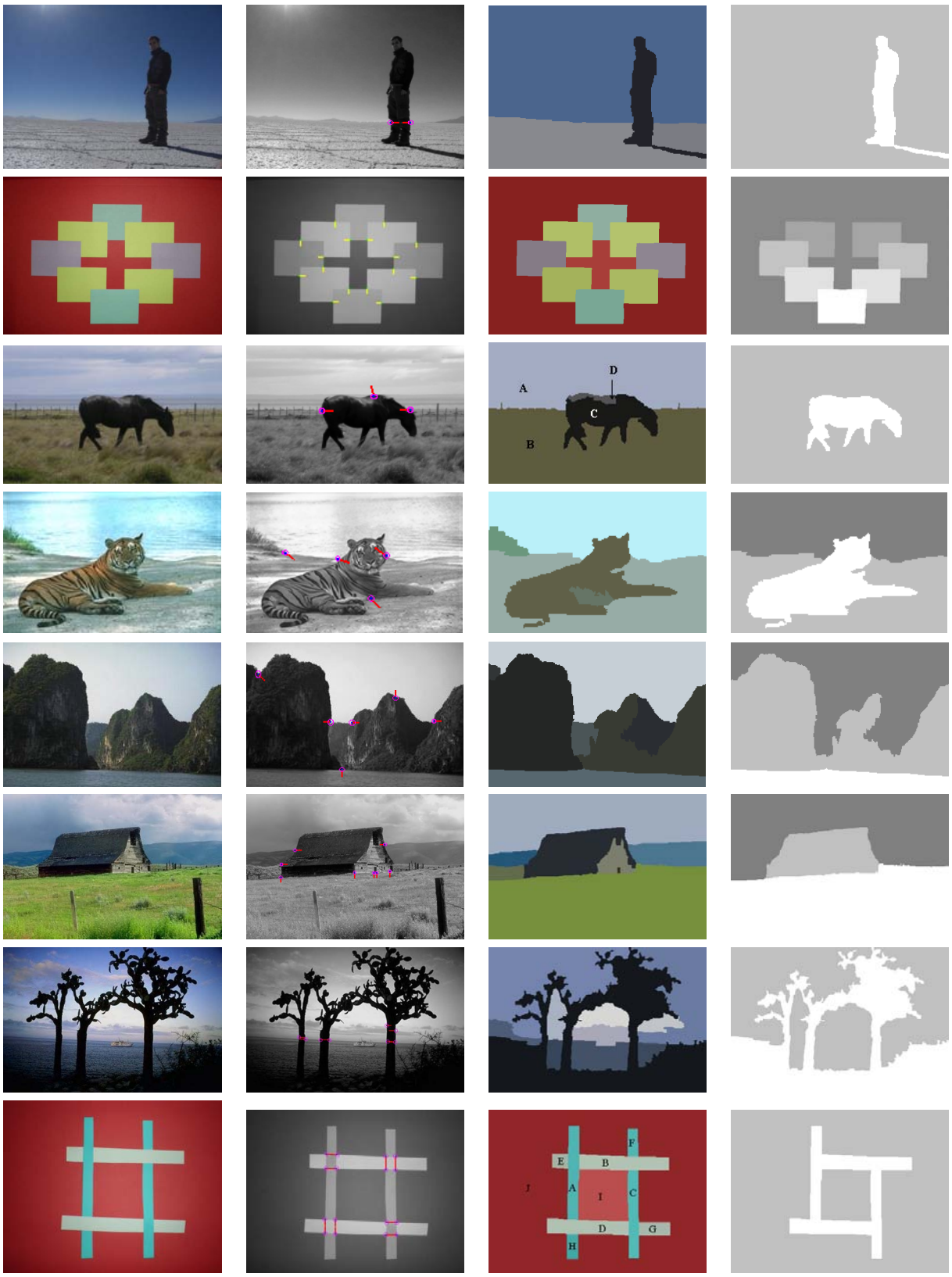


Figure 4.19: Examples of segmentation with depth: (a) Original image. (b) T-junction detection. (c) Segmentation. (d) Depth ordering.

can be observed in the example on the first row and on the second rows, the last level of depth includes two regions, corresponding to leave nodes of the Hasse diagram. In the example on the third row, there is a case of conflict between the regions A and C: while region A is interpreted as foreground and region C as background for one T-junctions, the contrary is true for two T-junctions. The solution of this conflict leads to a correct depth interpretation. A similar case is shown in the example on the fourth row, which involves more depth levels. The following two examples are more complex scenes, for which a correct depth interpretation is obtained. The example on the last two rows illustrates the limitations of the proposed method. In the first example, the regions corresponding to the sky visible through the tree branches have been merged with the region corresponding to the tree branches because there is simply no occlusion relationship allowing to separate them. In general, this happens when the complementary of the foreground region is not a single regions, that is when the foreground regions is not simply connected. In the second example, a case of self-occlusion is involved. The nodes corresponding to the regions A, B, C, and D form a cycle on the DG, characterized by the fact that all edges connecting the nodes of the cycle have the same weight. In this case, all regions corresponding to the nodes of the cycle are considered to belong to the same level of depth and, therefore, they appear as a single region on the map of relative depths.

4.3 Diffusion based framework versus region-merging based framework

In this section, we compare qualitatively the performances of the two proposed approaches for monocular depth cue integration by illustrating their suitability as well as their limitations on a set of real images (see Figure 4.20). For an objective evaluation, the same set of monocular depth cues has been used as input for both approaches.

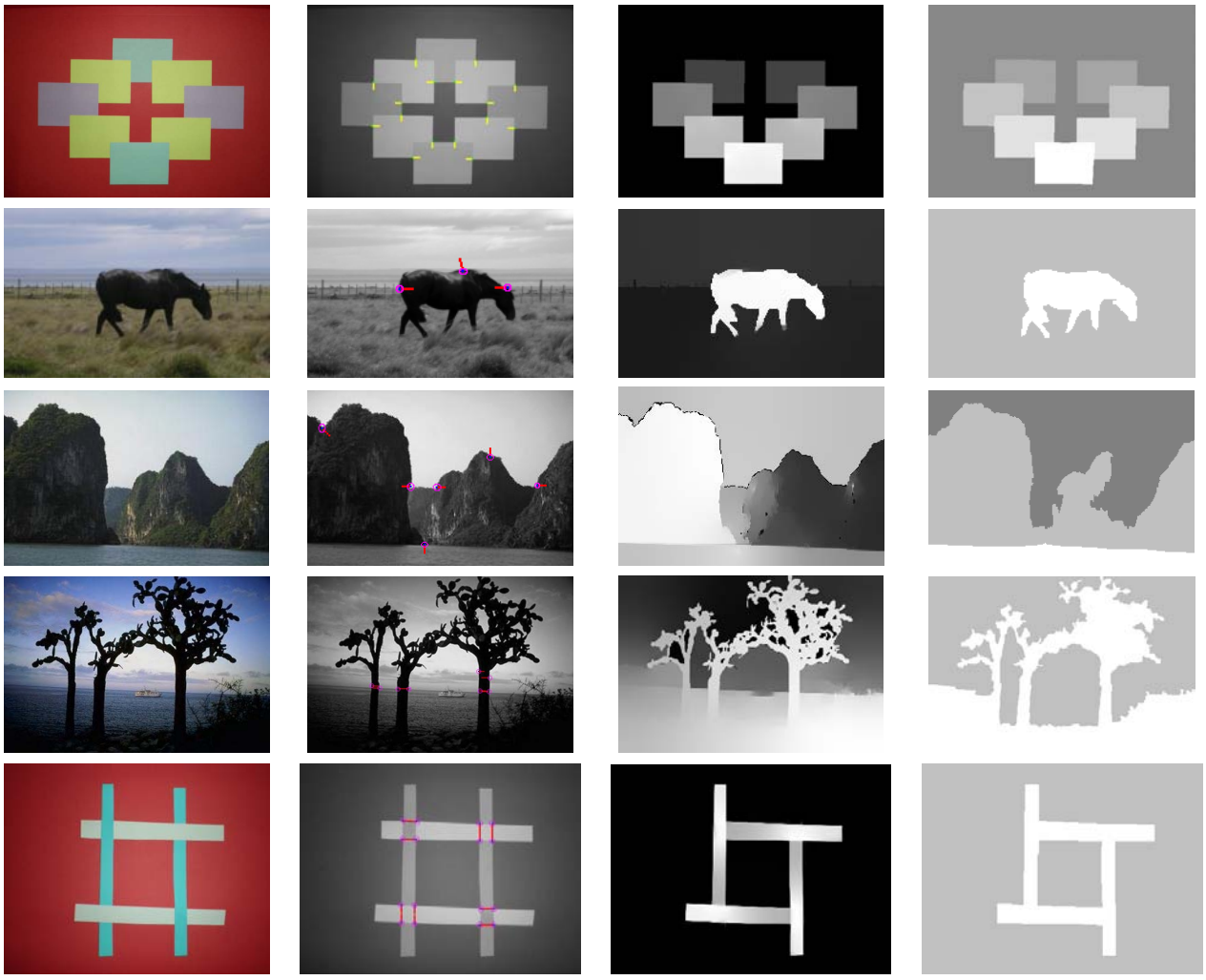
The example on the first row demonstrates that both proposed approaches can deal correctly with images involving multiple depth layers. On the second row, there is an example of conflict between local depth cues. In this case, the region-merging based framework gives a correct depth interpretation, while the diffusion based approach gives only a partially correct results since it leads to see the sky in front of the prairie. The example on the third row involves an image characterized by the presence of regions with high intensity variations beside the presence of a conflict. In this case, while the region-merging based approach gives a quite correct depth interpretation, the diffusion based approach shows its limitation in dealing with this kind of images. Indeed, the diffusion process is stopped in presence of high intensity variations inside a textured region and depth values cannot properly be propagated. For instance, in the map of relative depth, the region corresponding to the water appears to be further away to the viewpoint than the region corresponding to the mountain on the left. This occurs because the

value of the FSP marking the region corresponding to the water is not properly propagated. In the example on the fourth row, an image with a not simply connected foreground region is considered. In this case, while the depth interpretation obtained by performing the diffusion based approach is correct, the region-merging based framework shows a limitation. Indeed, the "holes" of the connected foreground regions disappear during the merging process and the foreground region appear in the map of relative depth as a simply connected region. The example on the last row involves a case of self-occlusion. While the diffusion based approach gives a correct interpretation of this phenomenon, the region-merging leads to see the regions in self-occlusion as single foreground region.

Beside the examples analyzed in this section, it should be taken into account that the current implementation of the region-merging based framework presents a limitation in terms of generality since its field of applicability is restricted to images that contain T-junctions. However, there is no theoretical difficulty to the incorporation in this framework of other monocular depth cues, since a mechanism for solving possible conflicting interpretations has already been envisaged. Beside the gain in generability, including other monocular depth cues would allow to obtain a more detailed depth map, and in turn also a more accurate segmentation in terms of number of meaningful regions. One more issue that need to be stressed here is the robustness with respect to errors in the input, given by the local depth relations. Contrary to the diffusion based approach, the region-merging based approach takes a binary decision in presence of conflicting depth interpretations. This characteristic makes it more exposed to risk of incorrect depth interpretations in presence of a very disturbed input.

4.4 Chapter summary

This chapter has proposed two different frameworks for monocular depth cue integration: a diffusion-based framework and a region-merging based framework. The former, described in section 4.1, relies on the use of a nonlinear filter that iteratively extends sparse initial depth values arisen from local depth cues to the entire image domain. We proved experimentally that the diffusion framework can be used successfully to perform the integration of several monocular depth cues. Experimental results involving occlusion, transparency, convexity, visual completion (both amodal and modal) and self-occlusion have shown a correct interpretation of several real images. In particular, contradictory information given by conflicting depth cues were dealt with correctly by the proposed mechanism which permits two regions to invert harmoniously their depths, in full agreement with the phenomenology, and very diverse Gestalt laws were fused harmoniously within this simple diffusion filter. The latter framework for monocular depth cue integration, detailed in section 4.2, is based on the use of a statistical region merging algorithm that incorporates local depth ordering arising from T-junctions, avoiding regions in occlusion



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

Figure 4.20: Results obtained by performing the two proposed approaches. (a) Original image. (b) T-junction detection. (c) Depth ordering obtained by using the diffusion based framework. (d) Depth ordering obtained by using the region-merging based framework.

to merge. The regions of the final partition and their relative depth relations are then encoded through a DG. Using this formalization, possible conflicting interpretations are easily detected as cycles on the DG and solved leading to a DAG. The depthmap is then obtained as the Hasse diagram corresponding to the transitive reduction of the DAG. Experimental results have proved that the proposed framework gives a correct interpretation of a variety of real images by relying solely on the cue of occlusion.

A qualitative analysis of the performances of the proposed methods has evidenced the suitability and limitations of both approaches. On one side, it has been shown that the region-merging based approach is best suited for images including textured regions with high intensity variations, that cannot properly be dealt with by the diffusion based approach. On the other side, it has been shown that the diffusion-based approach is particularly suited for the integration of a variety of monocular depth cues and it can correctly handle cases, such as self-occlusion and not simply connected foreground regions, that cannot properly be treated by the region-merging based approach. Hence, the region-merging based framework is more robust but the diffusion-based framework is more flexible.

In the next chapter, we shown how to exploit the depth ordering information extracted by the proposed methods for image filtering purpose.

Chapter 5

Depth-oriented Image Filtering

This chapter demonstrates the interest of monocular depth ordering estimation for region-based image filtering purpose. In the context of region-based image analysis, the purpose of a filter is often to remove some image regions that do not fulfill a certain simplification criterion. By taking advantage of monocular depth ordering estimation, a novel depth-oriented image filter is proposed in this chapter, which allows to remove image regions following a depth criterion so that, for instance, the foreground regions can be automatically removed or the background region automatically substituted. Section 5.1 illustrates the proposed filtering approach, detailing how the regions to be removed are selected and how they are restituted on the filtered image. The experimental validation is reported in section 5.2.

5.1 Filtering strategy

Classical region-based image filtering techniques, namely connected operators [Hei91, Vin93, Cre95, Sal95, Mey97, Sal98, Sal00, Ouz05], act by taking a decision on which image regions should be removed and which should be preserved on the basis of a given simplification criterion. Typically, the simplification criterion is related to some low-level feature of the region under consideration, such as the gray level or the geometry, or to the result of applying some morphological operator. In all cases, the decision is not related to the structure of the scene depicted in the image. As a consequence, these filters can be useful for removing some image details before performing more complex image processing tasks such as segmentation and editing, but they are not suitable for object-oriented filtering applications.

In chapter 4, the problem of monocular depth ordering estimation has been addressed, leading to the proposal of two different strategies for estimating the map of relative depths in single images. This spatial understanding of single images enables the introduction of an important intermediate layer in image filtering application. By basing the decision on which regions should be

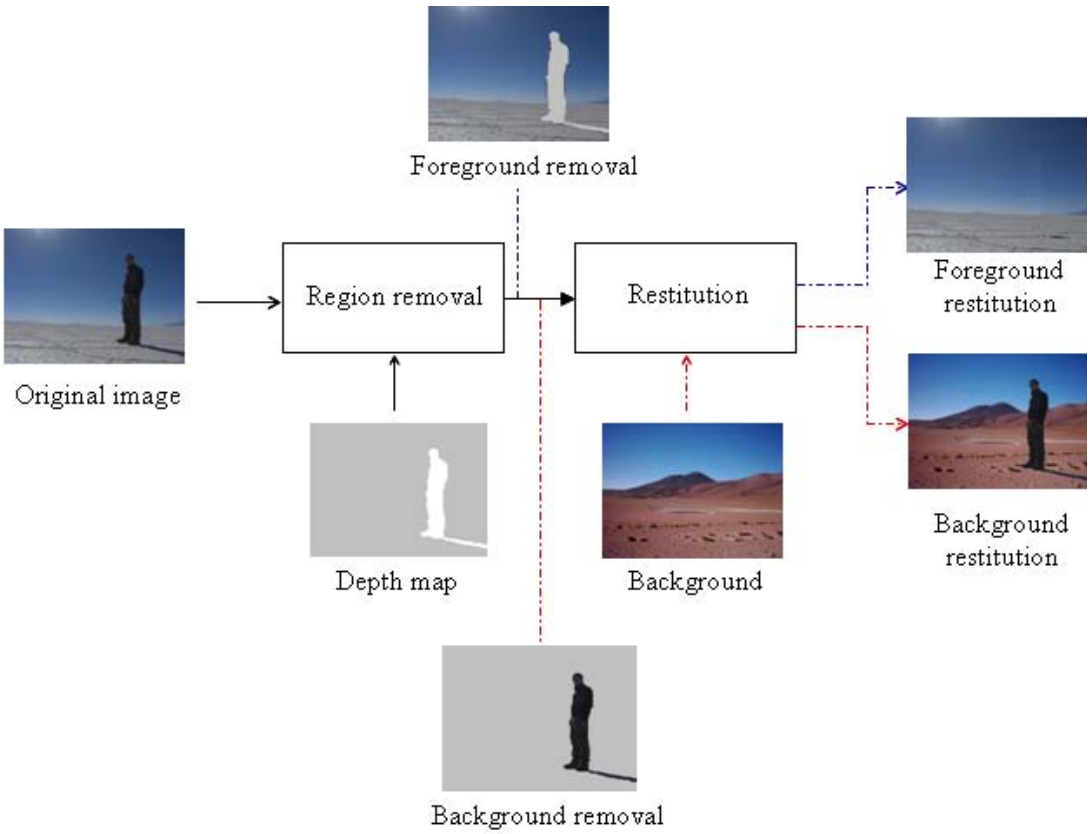


Figure 5.1: Depth-oriented image filtering scheme for automatic foreground object removal and background substitution. Foreground object removal requires only two inputs: the original image and its associated depth map. Background substitution requires an additional input: the new background image.

removed and which should be preserved on the map of relative depths, new filtering applications can be developed. Contrary to classical region-based simplification criteria, the depth-oriented criterion we propose is more related to the notion of objects and, as a consequence, it is suitable for object-oriented filtering applications, such as, for instance, automatic foreground object removal and automatic background substitution.

The scheme used in this chapter for depth-oriented image filtering is illustrated in Figure 5.1. The system takes as inputs the image to be filtered and its corresponding map of relative depths (hereafter called *depth map*). The depth map is a labeled image, where each label corresponds to a level of depth. The decision on which image regions should be removed and which should be preserved is taken by analyzing the depth map associated to the image. For instance, if the goal is to remove foreground regions, then the regions of support of the closest regions to the viewpoint in the depth map are removed from the original image (see Figure 5.1). The depth map can be obtained by using one of the two approaches for monocular depth ordering estimation proposed in chapter 4. However, it should be taken into account that the estimation of depth obtained

by performing the diffusion-based approach does not correspond to a labeled image. Hence, before using it in the context of a depth-oriented filter, the depth image obtained by performing this method has to be segmented. The segmentation can be done, for instance, by using the region-merging algorithm proposed in section 3.1.3.2. The way the removed regions are replaced depends on the kind of application. If the goal is to remove foreground regions from the image, a state-of-the-art image completion technique [Kom07] is used, which restores the removed regions as if they were originally not present in the image, giving a visually plausible result (see Figure 5.2). Instead, if the goal is to substitute the background, a simple toggle-mapping is done, which uses another background image either provided by the user, or already present in an image background database. The idea of using image completion as restitution strategy in the context of region-based image filtering has been originally proposed in [Dim07], where the goal was to eliminate the perceptual presence of regions removed by a connected operator [Sal95]. Contrary to what happens in [Dim07], the image completion restitution strategy is applied to all removed regions following the depth-oriented criterion, without the need of any preprocessing aiming at the selection of the regions to be restituted by image completion. The only operation that needs to be performed before applying the restitution strategy, no matter the kind of application, is a morphological dilation with a squared structuring element of size 3×3 of the region of the support of the removed region. This is done to remove from the preserved regions the transition zone in correspondence of object boundaries.

The next section details the image completion technique used as restitution strategy.

5.1.1 Restitution by image completion

Based only on the observed part of an incomplete or consciously masked image, the goal of image completion is to fill the missing part so that a visually plausible outcome is obtained. In [Kom07], this task is posed in the form of a discrete global optimization problem, whose objective function corresponds to the energy of a discrete Markov Random Field (MRF) [Kin80] and is optimized by using an optimization scheme, called Priority-Belief Propagation. In the following, we introduce the concept of discrete MRF and we explain how, in [Kom07], it is used for modeling the problem of image completion.

5.1.1.1 Modeling the problem of image completion through a discrete MRF

Given a discrete set of labels $\mathcal{L}$ and an undirected graph $\mathcal{G} = (V, E)$, composed of a set of nodes V and a set of undirected edges E connecting nodes, a discrete MRF is obtained by associating to each node $p \in V$ a random variable $\hat{x}_p$, which takes values in $\mathcal{L}$. Each node of the graph represents a variable or a group of variables and each edge of the graph connecting two nodes represents a probabilistic dependency between variables or group of variables. Let N

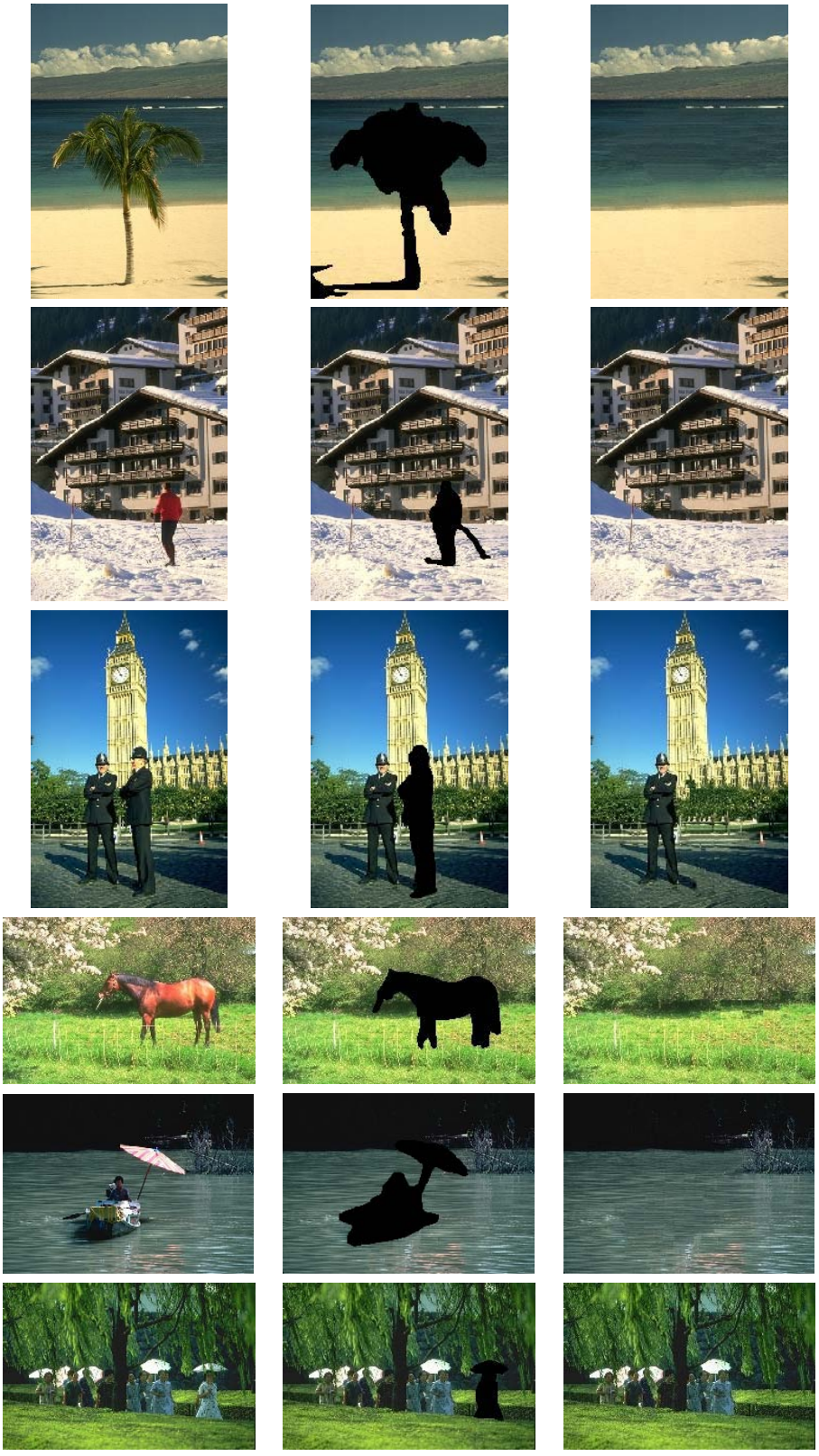


Figure 5.2: Examples of image completion obtained by performing the method in [Kom07]: (a) Original image. (b) Image where the region to be completed is marked in black. (c) Completed image.

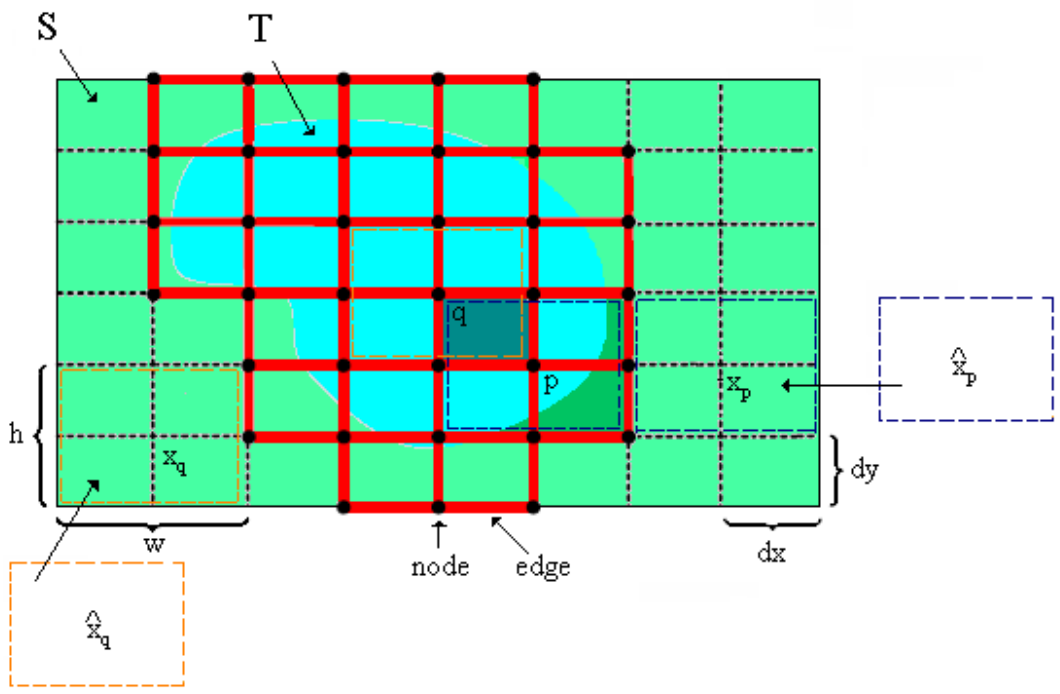


Figure 5.3: Figure illustrating the problem of image completion and how to model it through a MRF. T is the region to be completed (blue region) and S is the source region (green region). The labels of the MRF are the set of $w \times h$ patches of the source region S. The nodes of the MRF are all points of the image grid of size $dx \times dy$, whose $w \times h$ neighborhood intersects the region T (black dots). The edges of the MRF are the segments of the grid connecting pairs of neighboring nodes (red segments). When a patch $\hat{x}_p$ centered at x_p from the source region is assigned to the node p, the single node potential of the node p is given by the SSD on the region of overlap of the patch $\hat{x}_p$, when it is centered at p, with the source region S (dark green region). The pairwise potential of two neighboring nodes p and q is given by the SSD on the region of overlap of the patch $\hat{x}_p$ with the patch $\hat{x}_q$, when they are centered in p and q respectively (dark blue region).

be the number of nodes in the graph and let $\mathcal{X}^N$ be the *sample space*, given by the set of all N-dimensional vectors $x = \{\hat{x}_p | p \in V\}$, having components in $\mathcal{L}$. By assigning to each node $p \in V$ a random variable $\hat{x}_p \in \mathcal{L}$, the MRF assigns to each vector $x \in \mathcal{X}^N$ a probability of mass $p(x)$. The structure of the underlying graph acts as a kind of filter for the allowed distributions, since the probability of mass $p(x)$ has to respect the probabilistic dependencies represented by the graph edges.

To model the image completion problem through a discrete MRF, an undirected graph $\mathcal{G} = (V, E)$ and a set of labels $\mathcal{L}$ have to be defined on the discrete image $U : \Gamma \rightarrow \mathcal{Z}$. Let T be the image region to be completed and S a source region, which is always a subset of $(\Gamma - \Omega)$. To define the graph $\mathcal{G}$, an image grid of size $dx \times dy$ has been considered (see Figure 5.3). The set of nodes V of the graph is given by all points of the image grid whose $w \times h$ neighborhood intersects the target region (black dots in Figure 5.3), while the set of edges E of the graph

represents a 4-neighborhood system on the image grid (red segments in Figure 5.3). The set of labels $\mathcal{L}$ consists of all patches from the source region S (dashed rectangles of size $w \times h$ in Figure 5.3). In this context, to assign a label $\hat{x}_p$ to a node p means to copy the patch $\hat{x}_p$ on the image U centering it at the node p. Therefore, the energy of the MRF should be defined so that only patches that are consistent with the source region S as well as with each other inside T are allowed to be copied in T. To this goal, the proposed energy function includes two terms: a term $V_p(\hat{x}_p)$, called *single node potential*, which measures how well the patch $\hat{x}_p$ agrees with the source region around p (dark green in Figure 5.3), and a term $V_{pq}(\hat{x}_p, \hat{x}_q)$, called *pairwise potential*, which measures how the patches $\hat{x}_p$ and $\hat{x}_q$ agree at the resulting region of overlap (dark blue in Figure 5.3), when they are centered at the neighboring nodes p and q respectively. The measure of agreement is computed as the sum of squared differences (SSD) on the overlapping region. Based on this formulation, the choice of which label $\hat{x} \in \mathcal{L}$ has to be assigned to each node $p \in V$, has to be done so that the following energy is minimized:

$$\mathcal{F}(\hat{x}) = \sum_{V_p \in V} V_p(\hat{x}_p) + \sum_{(p,q) \in E} V_{pq}(\hat{x}_p, \hat{x}_q) \tag{5.1}$$

In the above equation, $V_p(\hat{x}_p)$ represents the single node potential given by

$$V_p(\hat{x}_p) = \sum_{z \in [-\frac{w}{2}\ \frac{w}{2}] \times [\frac{h}{2}\ \frac{h}{2}]} \mathcal{M}_p(p+z)(U(p+z) - U(x_p+z))^2 \tag{5.2}$$

where $\mathcal{M}_p(\cdot)$ is a binary mask, which is equal to one inside the region S, z indicates the displacement on the patch in comparison with respect to the center of the patch, and x_p indicates the center of the patch $\hat{x}_p$ from the source region S. The term $V_{pq}(\hat{x}_p, \hat{x}_q)$ indicates the pairwise potential given by

$$V_{pq}(\hat{x}_p, \hat{x}_p) = \sum_{z \in [-\frac{w}{2}\ \frac{w}{2}] \times [\frac{h}{2}\ \frac{h}{2}]} \mathcal{M}_{pq}(p+z)(U(x_p+z) - U(x_q+z))^2 \tag{5.3}$$

where $\mathcal{M}_{pq}(\cdot)$ is a binary mask, which is equal to one inside the region of overlap between the two patches centered at two neighboring nodes p and q (dark blue in Figure 5.3), z indicates the displacement on the patch in comparison with respect to the center of the patch, and x_p and x_q indicate respectively the center of the patches labeled as $\hat{x}_p$ and $\hat{x}_q$.

The next section details how the above energy is minimized.

5.1.1.2 Optimization by Priority Belief Propagation

In principle, the minimization of the equation 5.1 could be done by using a state-of-the-art algorithm for approximate inference on MRF, called Belief Propagation (BP) [Fre97]. BP is an iterative algorithm that works by propagating local messages between the nodes of a MRF following a given message passing schedule. In the context of image completion, a message

$m_{pq}(\hat{x}_q)$ sent from a node p to its neighbor node q about a label $\hat{x}_q$ indicates how likely the node p thinks that the label $\hat{x}_q$ is adequate for the node q. In order to decide which label $\hat{x}_q$ is the most appropriate for the node q, the node p has to take into account three factors: the pairwise potential $V_{pq}(\hat{x}_p, \hat{x}_q)$, the single node potential $V_p(\hat{x}_p)$, and the opinion of all its neighbors, except q, about how adequate is the label $\hat{x}_p$ for p. Therefore, before sending a message to the node q, the node p must first receive the messages from all its remaining neighbors. Based on this formulation, the message $m_{pq}(x_q)$ corresponds to the following recursive equation:

$$m_{pq}(\hat{x}_q) = \min_{\hat{x}_p \in \mathcal{L}} \{V_{pq}(\hat{x}_p, \hat{x}_q) + V_p(\hat{x}_p) + \sum_{r:r\neq q,(r,p)\in E} m_{rp}(\hat{x}_p)\}, \tag{5.4}$$

where the term $\sum_{r:r\neq q,(r,p)\in E} m_{rp}(\hat{x}_p)$ represents the opinion of all neighbors of p, except q, about how adequate is the label $\hat{x}_p$ for p. At each iteration, each node sends one message per label to all its neighboring nodes, while it also receives messages from these nodes. The messages sent and received at the next iteration are based solely on the messages existing at the current iteration. In this way, all nodes work in a cooperative manner by sending messages to each other about which label they prefer. The process of sending and receiving messages stops when all messages do not change any more.

What makes the BP strategy unfeasible for the image completion problem is the huge number of existing labels. Indeed, being the number of messages sent from a node p to a neighboring node q at each iteration equal to the number of available labels, the computational cost would be intolerable. Furthermore, when the number of available labels is large, the convergence become slower and the probability of falling in a local minimum increases. To overcome these problems, the authors of [Kom07] have proposed a new optimization scheme called Priority-Belief Propagation, which introduces two extensions with respect to standard BP: a *dynamic label pruning* and a *priority-based message passing schedule*. Both extensions rely on the concept of *priority* attached to each node. The priority of a node is related to how confident is the node about the labels that should be assigned to him. The confidence for a node p is related to the number of labels for which the *belief* is above a given predefinite threshold b_{conf}. The belief $b_p(\hat{x}_p)$ of a node p expresses the probability of assigning to p the label $\hat{x}_p$. It is computed as follows:

$$b_p(\hat{x}_p) = -V_p(\hat{x}_p) - \sum_{r:(r,p)\in E} m_{rp}(\hat{x}_p), \tag{5.5}$$

where $V_p(\hat{x}_p)$ is the single node potential and $\sum_{r:(r,p)\in E} m_{rp}(\hat{x}_p)$ represents the opinion of all the neighbors of p about how adequate is the label $\hat{x}_p$ for p. Indeed, if for a given node p the number of labels having a high belief is large, the node is less confident about which label should be assigned to him and, conversely, if the number of labels having a high belief is small, the node is more confident about which label should be assigned to him. As a consequence, to give the higher priority to the more confident nodes, the priority of a node p is defined as follows:

$$priority(p) = \frac{1}{|CS(p)|}, \tag{5.6}$$

where $|CS(p)|$ is the cardinality of the set $CS(p) = \hat{x}_p \in \mathcal{L} : b_p^{rel}(\hat{x}_p) \geqslant b_{conf}$, called *confusion set* and $b_p^{rel}(\hat{x}_p) = b_p(\hat{x}_p) - b_p^{max}, b_p^{max} = max_{\hat{x}_p \in \mathcal{L}} b_p(\hat{x}_p)$, where b_{conf} is a predefinite SSD between patches.

Based on this definition of priority, the dynamic label pruning consists in reducing the number of possible labels for each node by discarting those labels that are unlikely to be assigned to that node. The remaining labels are called *active labels* for that node. The first nodes to send outgoing messages to its neighbors are the nodes whose $w \times h$ neighborhood intersects the source region S. Indeed, only for these nodes the single node potential is not zero and therefore they are more confident about which label they should prefer. Once the node with the highest priority has sent outgoing messages to its neighbors, the priorities of its neighbors nodes are actualized so that, the next node to be visited will be, among all uncommitted nodes, the one with the highest priority. Since the cost of an outgoing message from a node p is proportional to the number of available labels for p, the priority-based message scheduling principle favors the circulation of cheap messages. In addition, when a node p has a high confidence about its labels, its outgoing messages help the neighboring nodes of p to increase their own confidence. Therefore, the message passing schedule based on priority speeds-up the convergence reducing the probability of falling in local minima.

5.1.1.3 Reconstruction of the image

Once the convergence has been attained, each node of the MRF has a patch associated to it. Then, the region T has to be reconstructed by composing the patches. A simple way to reconstruct the image is by visiting the nodes of the MRF in the same order they were visited during the last iteration of the Priority-BP process, by blending them with a weight which is proportional to the final belief of each node. When the patch to be copied intersects the region S, the region of overlap in S is obtained as a mean of the values of the correspondent pixels.

5.1.1.4 Parameter setting

One of the most important parameter to be configured is the size of the patch $w \times h$. In all experiments, we considered squared patches ($w = h$) and we took as size of the grid the half size of the patches, that is: $dx = dy = w/2$. In general, the size of the patches should be sufficiently large to take into account the image structure in S and should increase with the size of the region T to be completed. Experimentally, we found that the minimum value of the patch size needed to take into account the image structure in S is 7×7. In Figure 5.4, we show some results obtained using different patch sizes. As can be observed, smaller patch sizes (see Figure 5.4 (c) and (d)) give a slightly better result in this case since the region to be completed is relatively small compared to the image size. However, also for larger patch sizes the result is

visually plausible (see Figure 5.4 (d) and (e)), even if some block artifacts are present.

Another important issue is how to determine the source region S. Theoretically, S could coincide with the set $\Gamma - T$. However, the larger is the number of nodes, the slower will be the convergence and therefore the probability of falling in a local minimum would increase. On the other side, the source region S should be large enough to include all information needed to complete the target region T. In practice, we computed the region S by dilating the mask with a squared structuring element, whose size has been fixed experimentally to $(w+2) \times (w+2)$.

Experimentally, we found that the above described algorithm is very sensitive to variations of the parameter b_{conf}. Indeed, this parameter is used for the definition of priority, which is the core of the Priority-BP scheme since both, the label pruning and the message passing schedule, are based on the concept of priority. Recall that the priority of a node is inversely proportional to the cardinality of the confusion set, which is determined by the value of b_{conf}, since $CS(p) = \hat{x}_p \in \mathcal{L} : b_p^{rel}(\hat{x}_p) \geqslant b_{conf}$.

Since the number of labels pruned at each iteration depends on the value of the priority, a minimum number of active labels L_{min}, say 10, is always kept. Hence, if the value of the parameter b_{conf} is too high, it could happen that, after a few iterations, all nodes require to active the parameter L_{min}. Then, it will be impossible to distinguish which of these nodes are more confident about their labels and the possibility of visiting the nodes in an incorrect order, will be higher. On the contrary, if the value of b_{conf} is too small, all the label may have a confidence above b_{conf}, so that it does not have any effect. Instead of using a fixed value of the parameter b_{conf}, we proposes to determine it dynamically. In this new scheme, only the L patches with the higher priority are considered, the mean value of their confidence over the L patches is computed for each node and b_{conf} is taken as the largest value of the mean obtained over all the nodes. We take $L = L_{min}$. In this way, the value of b_{conf} is adapted to the dynamic of the process, since after each iteration, its value is determined by the L patches that are considered the more appropriate for the node under consideration.

5.2 Experimental results

In this section, we show some results obtained by applying the proposed depth-oriented filter to a set of real images. In Figure 5.5 and Figure 5.6, we show some examples using respectively image completion and background substitution as restitution strategies. For each example, we show the original image, its map of relative depths, which is rendered as a gray level image (high values indicate regions closer to the viewpoint), and the filtered image. In all examples shown in this section, the depth cues have been detected by using the approach described in section 3.1.3, while the depth cue integration has been performed by using the region-merging based framework described in section 4.2. In the example on the first and on the second row, the last

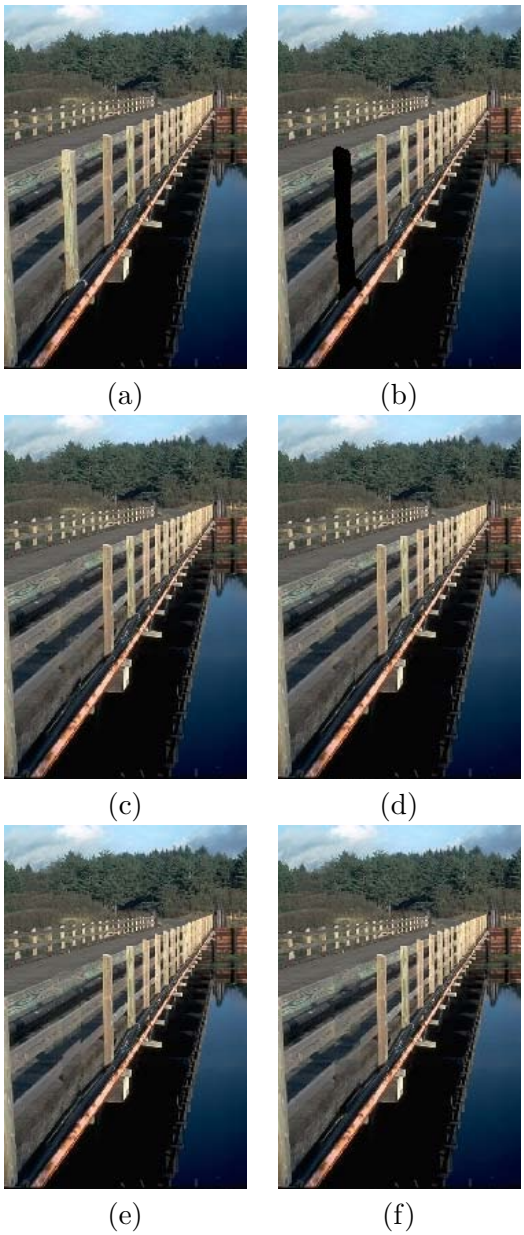


Figure 5.4: Example of results obtained by varying the patch size: (a) Original image. (b) Image where the region to be completed is marked in black. (c) Completed image with patch size 7×7. (d) Completed image with patch size 9×9. (e) Completed image with patch size 15×15. (f) Completed image with patch size 19×19

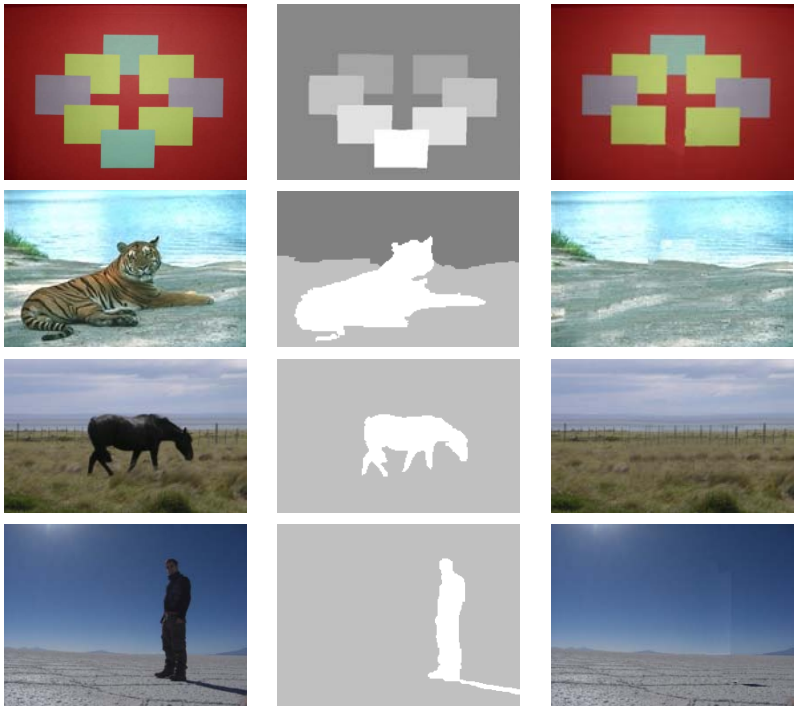


Figure 5.5: Examples of foreground removal: (a) Original image. (b) Map of relative depths. (c) Foreground removal.

level of depth has been removed and substituted by the corresponding regions of the background image. Instead, in the example on the third row, three levels of depth have been detected and the foreground object has been preserved while the regions of support corresponding to the two last levels of depth have been substituted. In the following example involving the house, three levels of depth have been detected. For this image, we first remove and substitute the first level of depth corresponding to the prairie (foreground), then the last level of depth corresponding to the sky (background), and finally both of them so that only the intermediate level is kept.

These examples show that, understanding the spatial layout of images enables the user interface to be aware of the relative depth of the image being edited, so that a given level of depth could automatically be removed and replaced as it were originally not present in the image or by using the corresponding regions of another background image.

5.3 Chapter summary

This chapter has discussed the interest of monocular depth ordering estimation for region-based image filtering. A novel depth-oriented image filter has been proposed, which takes advantage of the estimation of depth ordering in single image to remove image regions following a depth

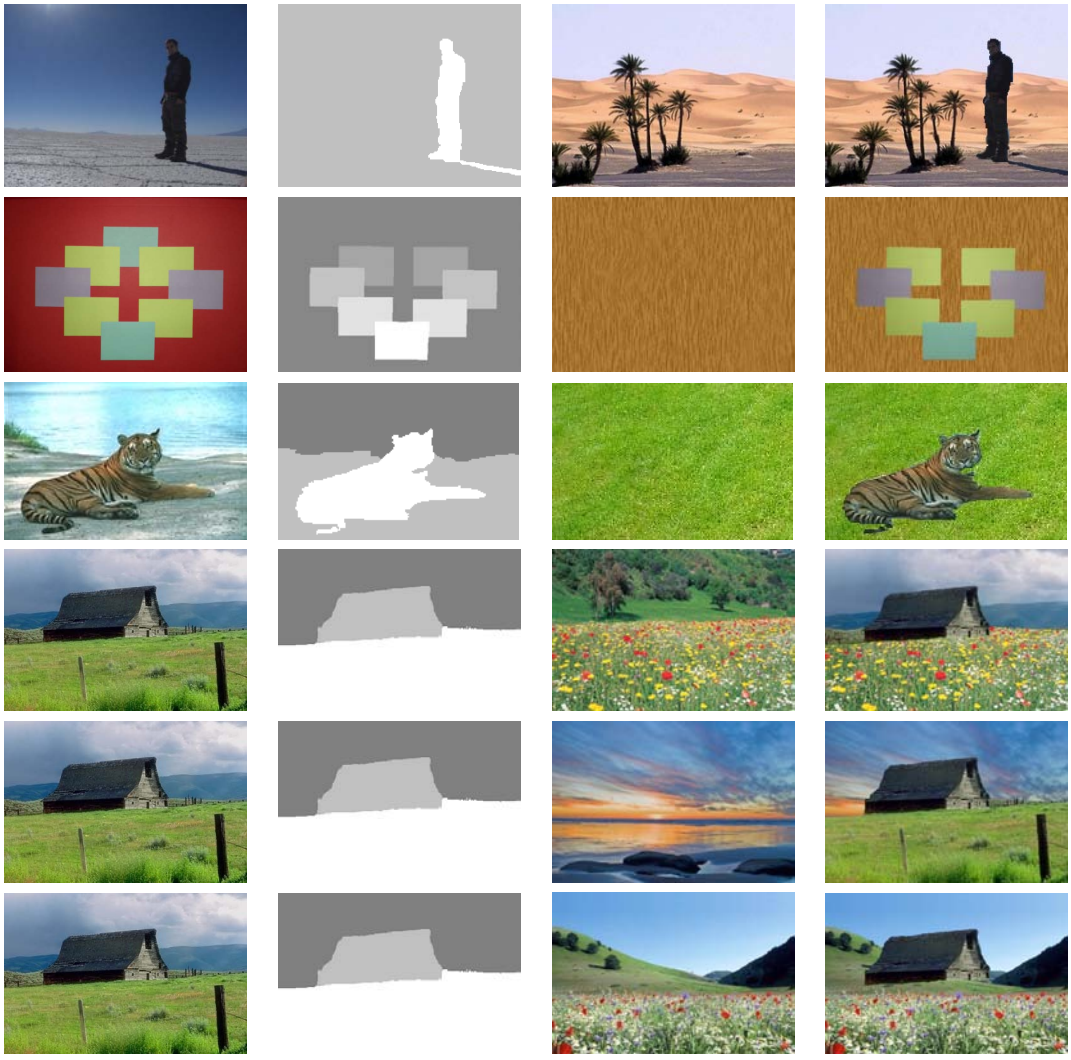


Figure 5.6: Examples of background substitution: (a) Original image. (b) Map of relative depths. (c) Background image. (d) Background substitution.

criterion. The depth-oriented filter is based on a simple toggle mapping operator: it takes a decision on which image regions have to be removed and which have to be preserved on the basis of the relative depth of the regions. The restitution strategy is application dependent. If the goal is to remove foreground regions, an image completion algorithm is used, which allows to replace the removed regions as if they were originally not present in the image, giving a visually plausible result. Instead, if the goal is to change the background, the removed region is replaced by using another background image. Experimental results, have demonstrated the interest of the proposed depth-oriented filtering approach for automatic foreground object removal and background substitution.

Chapter 6

Conclusions and Future Work

This conclusive chapter summarizes the main findings of this Ph.D. dissertation, discusses limitations and proposes possible future lines of research.

6.1 Findings

In chapter 3, a set of monocular depth cue detectors has been proposed including the cues of occlusion, camouflage, transparency, visual completion, and convexity. In particular, two different approaches have been investigated for detecting occlusion cues, namely T-junctions: a line segment-based approach and a region-merging based approach. The line segment-based approach works at a more global scale since it detects T-junctions as intersection of line-segments, which are themselves detected by using global information. Instead, the region-merging based approach works at a more local scale since it acts by segmenting through region-merging a relatively small neighborhood of each candidate point. Comparative results have shown that, in spite of its locality, the region-merging based approach is more robust.

In chapter 4, two different approaches for monocular depth estimation have been proposed and analyzed, namely a diffusion based approach and a region-merging based approach. Both of them act by incorporating depth ordering information provided by a set of monocular depth cues into a grouping process, providing an unified framework for the tasks of image segmentation and depth segregation. In the diffusion-based approach, the unified framework is provided by a nonlinear filter, which iteratively extends the initial depth values arisen from monocular depth cues to the entire image domain. Under this framework, the process of grouping by feature similarity relies on color information, while the process of separating by depth-dissimilarity corresponds to the internal boundary conditions, on the initial depth gradient, of the PDE underlying the nonlinear filter. The mechanism allowing to solve possible conflicting local depth interpretations is incorporated in the diffusion process itself, since a global depth interpretation is obtained

when the convergence is attained. In the region-merging based approach, the unified framework is provided by a region-merging algorithm, which iteratively merges pairs of neighboring regions following a statistical region-merging criterion. Under this framework, the process of grouping by feature similarity relies on a statistical pixel modeling, which exploits the image self-similarity, while the process of separating by depth dissimilarity corresponds to a merging criterion that acts as a sieve on the mergings proposed by the merging criterion. The mechanism allowing to solve possible conflicting local depth interpretations relies on a DG, which encodes the depth relationships between the regions of the final partition and allows to detect and solve possible conflicts as cycles on the graph. The final depth ordering is then obtained as transitive reduction of the DAG. Comparative results have shown that, on one side, the diffusion-based approach is more flexible since it can treat cases such self-occlusion and not simply connected foreground regions that cannot be properly handled by the region-merging based approach. On the other side, the region-merging based approach is more robust since it can deal with images including textured regions with high intensity variations, that cannot be dealt with correctly by the diffusion based approach.

In chapter 5, we have shown that monocular depth ordering estimation can be successfully used in region-based image filtering applications. A new depth-oriented filter has been proposed that allows to remove image regions following a depth criterion, so that image regions corresponding to a prespecified level of depth can be removed and replaced by a visually plausible background. The restitution is done by using either an image completion technique or by using another background image provided by the user or included on a previously constructed image background database.

6.2 Limitations

While the techniques presented in this Ph.D. dissertation are encouraging, there are still numerous issues that cannot be handled by the current state of this research. In particular, the proposed depth cue detectors only act at a given scale and, consequently, their robustness is somewhat limited and they cannot properly handle images involving a large depth range.

The limitations of depth cue detectors also mitigate the performances of both methods proposed for monocular depth cue integration, specially in the case of the region-merging based approach. Indeed, in this case, a binary decision is taken in presence of conflicting depth interpretations. This characteristic makes this approach more exposed to incorrect decisions in presence of a very disturbed input. The region-merging based framework also presents a limitation in terms of generality, since it cannot deal with self-occlusion and not simply connected foreground regions. Instead, the diffusion based framework cannot handle images involving texture with high intensity variations.

The current implementations of the work presented here are also too slow for interactive applications.

6.3 Future Work

To address the above mentioned limitations, we propose some avenues for future research.

The region-merging approach for T-junction detection could be extended to the detection of any kind of junction, and therefore of camouflage, transparency and modal completion, by incorporating a minimization framework to determine the minimum set of wedges that best approximate local data at a given candidate point. To address the problem of scale-variance, the automatic selection of the size of the local neighborhood to be segmented could be envisaged. By selecting its own size of the local neighborhood, the detector could determine the appropriate scale to represent the junction, making a scale invariant junction detector. One simple approach is to determine the scale in a scale space analysis, determining the best representant across multiples scales.

The diffusion based approach could be extended to handle texture with high intensity variations by changing the way the neighborhood is defined. Instead of defining the neighborhood by comparing the values of single pixels, a more complex, rotation invariant comparison involving the neighborhood of the pixels under consideration could be done.

The region-merging based framework could be improved in terms of robustness with respect to false detection by using a probabilistic criterion for solving possible conflicting interpretations. For instance, to each edge on the graph could be assigned a weight that takes into account some properly defined *depth cue likelihood*, which could, for instance, be derived by the contours.

Conclusions et travail future

Ce dernier chapitre résume les principales contributions de cette thèse, discute ces limitations et propose des directions de recherche potentiellement intéressantes pour le future.

Contributions de recherche

Dans le chapitre 3, un ensemble de détecteurs d'indices de profondeur monoculaire a été proposé. L'ensemble inclut les indices d'occultation, de camouflage, de transparence, d'achèvement visuel et de convexité. En particulier, deux approches différentes ont été étudiées pour détecter les indices d'occultations, c'est à dire les jonctions en T: une approche basée sur un détecteur de segments et une approche basée sur le fusionnement de région. L'approche basée sur le détecteur de segments travaille à une échelle plus globale. Les jonctions en T sont détectées comme intersection de segments, qui sont eux même détectés en utilisant une information globale. Par contre, l'approche qui se base sur le fusionnement de région travaille à une échelle plus locale, dans le voisinage de chaque point candidat. Les résultats de comparaison ont démontrés que, malgré sa localité, l'approche basée sur le fusionnement des régions est plus robuste.

Dans le chapitre 4, deux approches différentes pour l'estimation de profondeur monoculaire ont été proposées: une approche reposant sur la diffusion et une approche reposant sur le fusionnement de région. Les deux approches agissent en incorporant l'information sur l'ordre de profondeur donné par un ensemble d'indices locaux de profondeur monoculaire dans un processus de groupement, en donnant un cadre unificateur pour la segmentation des images et la ségrégation en profondeur. Dans l'approche reposant sur la diffusion, le cadre unificateur est donné par un filtre non linéaire, qui étend itérativement les valeurs initiales de profondeur obtenues à partir des indices de profondeur monoculaire, au domaine de l'image. Dans ce cadre, le processus de groupement par similarité se base sur l'information de couleur, et le processus de séparation par dissemblance est lié aux conditions de contour, sur le gradient initial de profondeur, de l'EDP correspondant au filtre non-linéaire. Le mécanisme qui permet de résoudre les interprétations locales conflictuelles est incorporé dans le processus même de diffusion, puisque on obtient une interprétation globale quand le filtre converge. Dans l'approche reposant sur le fusionnement de

région, le cadre unificateur est donné par un algorithme qui fusionne itérativement les pairs de régions voisines en suivant un critère statistique pour définir l'ordre de fusion. Dans ce cadre, le processus de groupement par similarité se base sur un model statistique du pixel, qui exploite l'autosimilarité des images. Le processus de séparation par dissemblance correspond à un critère de fusionnement qui agit comme un tamis sur les fusionnements proposés par le critère statistique de fusionnement. Le mécanisme qui permet de résoudre des interprétations locales de profondeur conflictuelles se base sur un Graphe Orienté (GO), qui encode les relations de profondeur entre les régions de la partition finale et permet de détecter les conflits comme cycles dans le graphe. L'ordre final de profondeur est obtenu comme réduction transitive du GO. Les résultats de comparaison ont démontré que, d'un côté, l'approche reposant sur la diffusion est plus flexible: en fait, elle peut traiter des cas comme l'auto-occultation et des cas où les régions du premier plan ne sont pas complètement connexes (ce cas ne peut pas être correctement traité par l'approche reposant sur le fusionnement de région). De l'autre côté, l'approche reposant sur le fusionnement de région est plus robuste puisque elle peut traiter les images avec régions texturées qui ne peuvent pas être correctement traitées par l'approche reposant sur la diffusion.

Dans le chapitre 5, nous avons montré que l'estimation de l'ordre de profondeur peut être utilisée dans des applications de filtrage orienté région. Un nouveau filtre sensible à la profondeur a été proposé. Il permet d'enlever les régions de l'image en suivant un critère de profondeur. Par exemple, les régions correspondant à un certain niveau de profondeur peuvent être éliminées et replacées par un fond visuellement plausible. La restitution est faite en utilisant une technique d'achèvement visuelle, ou en utilisant une autre image de fond qui peut être donnée par l'utilisateur ou par une base de données d'images de fond construite préalablement.

Limitations

Bien que les techniques présentées dans cette thèse soient encourageantes, il y a encore beaucoup de problème qui ne peuvent pas être traités actuellement. En particulier, les détecteurs d'indice de profondeur proposés agissent seulement à une échelle donnée et, en conséquence, leur robustesse est d'une certaine façon limitée. En particulier, les détecteurs ne peuvent pas traiter proprement les images qui comportent une grande dynamique de profondeur.

Les limitations des détecteurs d'indice de profondeur mitigent aussi les performances des deux méthodes proposées pour l'intégration de l'information de profondeur donnée par les indices de profondeur, surtout dans l'approche reposant sur le fusionnement des régions. En fait, dans ce cas, on doit prendre une décision binaire en présence d'interprétations conflictuelles. Cette caractéristique rend cette approche plus propice à prendre une décision incorrecte en présence d'information très bruitée. L'approche reposant sur le fusionnement de région présente aussi une limitation en termes de généralité, car elle ne peut pas traiter l'auto-occultation et les régions

du premier plan qui ne sont pas simplement connexes. Par contre, l'approche reposant sur la diffusion ne peut pas traiter les images texturées avec de fortes variations d'intensité.

L'implémentation actuelle du travail présenté ici est trop lente pour des applications interactives.

Travail futur

Pour adresser les limitations mentionnées ci-dessus, nous proposons quelques directions pour une future recherche.

L'approche reposant sur le fusionnement de région pour la détection de jonctions en T poudrait être étendue pour traiter tous les types de jonction. Ainsi le même algorithme pourrait détecter le camouflage, la transparence et l'achèvement modal. Il faudra incorporer un cadre de minimisation pour déterminer le plus petit ensemble de régions qui permet d'approximer les données locales dans un voisinage d'un point candidat. Pour traiter le problème de l'invariance vis à vis de l'échelle, la sélection automatique de la taille du voisinage local qui doit être segmenté poudrait être envisagée. En sélectionnant la taille du voisinage local, le détecteur pourrait déterminer l'échelle la plus appropriée pour représenter la jonction. Une approche simple consiste à déterminer la meilleure représentation entre plusieurs échelles dans un cadre d'analyse espace-échelle.

L'approche reposant sur la diffusion poudrait être étendue pour traiter les textures avec de fortes variations d'intensité en changeant la forme avec laquelle le voisinage est défini. En lieu de comparer les voisinages en calculant la différence pixel à pixel, une comparaison plus complexe, invariante vis-à-vis des rotations et engageant le voisinage complet des pixels considérés, pourrait être fait.

L'approche qui fait appel au fusionnement de région pourrait être améliorée en termes de robustesse vis à vis des fausses détections en utilisant un critère probabiliste pour résoudre les interprétations conflictuelles. Par exemple, à chaque arc du graphe on pourrait associer un poids correspondant à la vraisemblance de l'indice de profondeur, qui pourrait, par exemple, être obtenus à partir des contours.

Bibliography

[And00] S. Ando, "Image Field Categorization and edge/corner detection from gradient covariance.", *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 22, nº 2, pags. 179 – 190, 2000.

[Apo06] N. Apostoloff, and A. Fitzgibbon, "Automatic video segmentation using spatiotemporal T-junctions", *Proc. of British Machine Video Conference*, pags. 1–10, 2006.

[Avr03] M. Avriel, *Nonlinear Programming: Analysis and Methods.*, Dover Publishing, 2003.

[Bau00] A. Baumberg, "Reliable Feature Matching Across Widely Separated Views", *Proc. of Computer Vision and Pattern Recognition*, pags. 774–781, 2000.

[Ber04] R. Bergevin, and A. Bubel, "Detection and Characterization of junctions", *Computer Vision and Image Understanding*, Vol. 93, nº 3, pags. 288–309, 2004.

[Bey89] D. Beymer, *Junctions: Their Detection and Use for Grouping Images*, Master Thesis, Massachusetts Institute of Technology, Departament of Electrical Engineering, 1989.

[Bha43] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions.", *Bulletin of the Calcutta Mathematical Society*, Vol. 35, pags. 99 –109, 1943.

[Bie87] I. Biederman, "Recognition by components: a theory of human image understanding.", *Psychological review*, Vol. 2, nº 94, pags. 115–147, 1987.

[Big94] J. Bigun, "A Structure Feature for Some Image Processing Appications Based on Spiral Functions", *Proc. of European Conference on Computer Vision*, pags. 383–394, 1994.

[Bua05] A. Buades, B. Coll, and J.M. Morel, "A review of image denoising algorithms, with a new one", *Multiscale modeling and Simulation, Society for Industrial and Applied Mathematics (SIAM)*, Vol. 4, nº 2, pags. 490–530, 2005.

[Bua06] A. Buades, B. Coll, and J.M. Morel, "The staircasing effect in neighborhood filters and its solution", *IEEE Transactions on Image Processing*, Vol. 15, nº 6, pags. 1499–1505, 2006.

[Bua08] A. Buades, B. Coll, and J.M. Morel, "Nonlocal image and movie denoising", *International Journal of Computer Vision*, Vol. 76, nº 2, pags. 123 – 139, 2008.

[Bul88] H.H. Bulthoff, and H.A. Mallot, "Integration of depth modules: stereo and shading.", *Journal of Optical society of America*, Vol. 63, pags. 1749–1758, 1988.

[Bur86] J.B. Burns, A.R. Hanson, and E.M. Riseman, "Extracting Straight Lines", *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 8, nº 4, pags. 425 – 455, 1986.

[Bur90] J. Burnett, and U. Buy, "Solving integer programming systems using the iminos prototype", Tech. rep., Department of Computer Science, University of Massachusetts, 1990.

[Bur05] J. Burge, M. A. Peterson, and S. E. Palmer, "Ordinal configural cues combine with metric disparity in depth perception.", *Journal of Vision*, Vol. 5(6):5, pags. 534542, 2005.

[Cal08] F. Calderero, and F. Marques, "General region merging approaches based on information theory statistical measures.", *Proc. of International Conference on Image Processing*, San Diego (CA), October 2008.

[Cas96a] V. Caselles, B. Coll, and J.M. Morel, "A Kanizsa programme", *Progress in Nonlinear Differential Equations and their Applications*, Vol. 25, pags. 35–55, 1996.

[Cas96b] V. Caselles, B. Coll, and J.M. Morel, "Topographic maps and local contrast changes in natural images", *Vision Research*, Vol. 25, pags. 317–367, 1996.

[Caz03] M.A. Cazorla, and F. Ercolano, "Two Bayesian Methods for Junctions Classification", *IEEE Trans. on Image Processing*, Vol. 12, nº 3, pags. 317–327, 2003.

[Cha53] A. Chapanis, and R. McCleary, "Interposition as a cue for the perception of relative distance", *Journal of General Psychology*, Vol. 48, pags. 113–132, 1953.

[Cha08] T. Chan, S. Esedoglu, and K. Ni, "Histogram based segmentation using the wasserstein distance", *Scale Space and Variational Methods in Computer Vision (LNCS)*, Vol. 4485, pags. 697–708, 2008.

[Chu97] F. Chung, *Spectral Graph Theory.*, CBMS Regional Conference Serie in Mathematics, 1997.

[Cla90] J. J. Clark, and A. L. Yuille, "Data Fusion for Sensory Information Processing Systems.", *Kluwer Academic Publisher, Norwell, MA*, 1990.

[Clo71] M. Clowes, "On seeing things.", *Artificial Intelligence*, Vol. 2(1), pags. 79–116, 1971.

[Cor01] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, *Introduction to Algorithms*, Chap. Depth-first-search, pags. 540–549, MIT Press and McGraw-Hill, 2001.

[Cre95] J. Crespo, J. Serra, and R.W. Schafer, "Theoretical aspects of morphological filters by reconstruction", *IEEE Transactions on Image Processing*, Vol. 47(2), pags. 201–225, 1995.

[Cut95] J.E. Cutting, and P.M. Vishton, *Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In W.Epstein amd S.Rogers, editors, Perception of Space and Motion.*, Vol. 257, pags. 69–117, Academic Press, San Diego, 1995.

[Del06] E. Delage, H. Lee, and Y. Ng, "A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image", *International Conference on Computer Vision and Pattern Recognition, CVPR*, pags. 1–8, 2006.

[Der94] R. Deriche, and T. Blaszka, "Recovering and Characterizing Image Features Using an Efficient Mode Based Approach.", *International Conference on Computer Vision and Pattern Recognition, CVPR*, pags. 530–535, 1994.

[Des00] A. Desolneux, L. Moisan, and J.M. Morel, "Meaningful alignments", *International Journal of Computer Vision (IJCV)*, Vol. 40, nº 1, pags. 7–23, 2000.

[Dim07] M. Dimiccoli, and P. Salembier, "Perceptual filtering with connected operators and image inpainting.", *International Symphosium on Mathematical Morphology, (ISMM)*, Rio de Janeiro (Brazil), October 2007.

[Dim08] M. Dimiccoli, J.M. Morel, and P. Salembier, "Monocular depth by nonlinear diffusion.", *Indian Conference on Computer Vision, Graphics and Image Processing, (ICVGIP)*, Bhubaneswar (India), December 2008.

[Dim09a] M. Dimiccoli, and P. Salembier, "Exploiting t-junctions for depth segregation in single images.", *In proc. of International Conference in Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei (Taiwan), April 2009.

[Dim09b] M. Dimiccoli, and P. Salembier, "Hierarchical region-based representation for segmentation and filtering with depth in single images.", *Accepted for presentation at International Conference on Image Processing (ICIP)*, Cairo (Egypt), November 2009.

[Efr99] A. Efros, and T. Leung, "Texture synthesis by non-parametric sampling", *In Proc. of International Conference on Computer Vision (ICCV)*, October 1999.

[Ese03] S. Esedoglu, and R. March, "Segmentation with depth but without detecting junctions", *J.Math.Imaging and Vision*, Vol. 18, pags. 7–15, 2003.

[Fal72] G. Falk, "Interpretation of imperfect line data as a three-dimensional scene.", *Artificial Intelligence*, Vol. 3, pags. 101144, 1972.

[Fav03] P. Favaro, A. Duci, Y. Ma, and S. Soatto, "On exploiting Occlusions in Multiple-view Geometry", *Proc. of International Conference on Computer Vision*, pags. 479–486, 2003.

[Fle04] R.W. Fleming, and B.L. Anderson, *The Visual Neurosciences. Cambridge,MA: MIT Press*, pags. 1284–1299, Chalupa, L. and Werner, J.S. Eds., 2004.

[For86] W. Forstner, "A feature based corresponding algorithm for image matching.", *International Archives of the Photogrammetry and Remote Sensing*, Vol. 26, pags. 150–166, 1986.

[For87] W. Forstner, and E. Gulch, "A Fast Operator for Detection and Precise Localization of Distinct Points, Corners, and Centres of Circular Features.", *In Proc. of Intercommission Conference on Fast Processing of Photogrammetric Data (ISPRS)*, pags. 281–305, 1987.

[For94] W. Forstner, "A framework for Low Level Features Extraction", *Proc. of European Conference on Computer Vision*, pags. 383–394, 1994.

[Fow] C.C. Fowlkes, D. Martin, and J. Malik, "The berkeley segmentation dataset and benchmark.", *www.cs.berkeley.edu/projects/vision/grouping/segbench/*.

[Fre91] W. Freeman, and E. Adelson, "The design and use of steerable filters", *IEEE Trans.on Pattern Analysis and Machine Intelligence*, Vol. 13, nº 9, pags. 891 –906, 1991.

[Fre97] B.J. Frey, and D.J.C. MacKay, "A revolution: Belief propagation in graphs with cycles.", *Advances in Neural Information Processing Systems 10*, Vol. 10, pags. 479485, 1997.

[Fre00] W. Freeman, "Learning Low-Level Vision", *International Journal of Computer Vision*, Vol. 40, nº 1, pags. 25–47, 2000.

[Gao07] R.-X. Gao, T.F. Wu, S. C. Zhu, and N. Sang, "Bayesian inference for layer representation with mixed markov random field.", *Energy Minimization methods in Computer Vision and Pattern Recognition*, Vol. 4679, pags. 213–224, 2007.

[Gei96] D. Geiger, and Parida L., "Visual organization for figure-ground separation", *International Conference on Computer Vision and Pattern Recognition, CVPR*, pags. 155–160, 1996.

[Gib50] J.J. Gibson, *The perception of the visual world.*, Oxford, England: Houghton Mifflin., 1950.

[Gio08] R. Grompone von Gioi, J. Jakubowicz, J.M. Morel, and G. Randall, "Lsd: A fast line segment detector with a false detection control.", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, December 2008, digital Object Identifier: 10.1109/TPAMI.2008.30.

[Gra73] G.R. Grape, "Model based (intermediate-level) computer vision", Tech. Rep. Memo, AIM-201 (STAN-CS-73-366), Stanford Artificial Intelligence, 1973.

[Gui04] F. Guichard, J.M. Morel, and R. Ryan, *Contrast invariant image analysis and PDE's.*, Preprint CMLA. Web Address: www.cmla.ens-cachan.fr/Utilisateurs/morel/JMMBookOct04.pdf, 2004.

[Guz68] A. Guzman, "Computer recognition of three-dimensional objects in a visual scene", Tech. Rep. MAC-TR-59, MIT, 1968.

[Har88] C.G. Harris, and M. Stephens, "A combined corner and edge detection", *Proc. of 4th Alvey Vision Conference*, pags. 147–151, 1988.

[Har93] R. Hartley, and R. Gupta, "Computing matched-epipolar projections.", *International Conference on Computer Vision and Pattern Recognition, CVPR*, pag. 549555, 1993.

[Har99] R. Hartley, "Theory and practice of projective rectification", *International Journal of Computer Vision (IJCV)*, Vol. 35(2), pags. 1–16, 1999.

[Hei91] H. Heijmans, "Theoretical aspects of gray level morphology", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 13(6), pags. 568–592, 1991.

[Hei93] F. Heitger, and R. von der Heydt, "A computational model of neural contour processing: Figure-ground segregation and illusory contours", *In Proc. of International Conference on Computer Vision (ICCV)*, pags. 32–40, 1993.

[Hei95] F. Heitger, "Feature detection using suppression and enhancement", Tech. rep., Communication Technology Labboratory, Swiss Federal Institute of Technology ETH, ETH-Zurich, Switzerland, 1995.

[Hel25] H.L.F. von Helmholtz, *Treatise on Physiological Optics.*, James P.C.Southall, 1925.

[Hit50] W.H. Hittelson, "Size as a cue to distance: static localization.", *American Journal of Psychology*, Vol. 64, pags. 54–67, 1950.

[Hoi07] D. Hoiem, A.N. Stein, A.A. Efros, and M. Hebert, "Recovering Occlusion Boundaries from a Single Image", *In Proc. of International Conference on Computer Vision (ICCV)*, pags. 1–8, 2007.

[Hue71] M.F. Hueckel, "An Operator which Locates Edges in Digitized Pictures", *Journal of the ACM*, Vol. 18, pags. 113–125, 1971.

[Huf71] D.A. Huffman, "Impossible objects as nonsense sentences.", *Machine Intelligence*, Vol. 8, pags. 475–492, 1971.

[KÖ3] U. Köthe, "Edge and junction detection with an improved structure tensor.", *25th Symposium of the German Association for Pattern Recognition (DAGM), Magdeburg, Lecture Notes on Computer Science (LNCS)*, pags. 25–32, 2003.

[Kan76] G. Kanizsa, and W. Gerbino, *Vision and artifact*, Chap. "Convexity and symmetry in Figure-Ground organization.", pag. 2532, In M. Henle (Ed.),New York: Springer, 1976.

[Kan79] G. Kanizsa, *Organization in Vision: Essays on Gestalt Perception*, Praeger, New York, 1979.

[Kan96] G. Kanizsa, *La Grammatica del Vedere*, Diderot, 1996.

[Kar04] S. Karlsson, *Monocular Depth from Occluding Edges*, PhD Thesis, Lund: Centre for Mathematical Sciences, 2004.

[Kau74] L. Kaufman, *Sight and mind: an introduction to visual perception.*, Oxford University Press, 1974.

[Kel77] D.H. Kelly, "Visual contrast sensitivity.", *Optica Acta*, Vol. 24, pags. 107–129, 1977.

[Kel91] P.J. Kellman, and T.F. Shipley, "Visual interpolation in object perception", *Current Directions in Psychological Science*, Vol. 1, nº 6, pags. 193–199, 1991.

[Kel98] P.J. Kellman, and M. Arterberry, *Thr crodle of knowledge.*, Cambridge, MA: MIT Press, 1998.

[Kel01] P.J. Kellman, and T.F. Shipley, *From Fragments to Objects: Segmentation and Grouping in Vision*, North Holland, 2001.

[Kin80] R. Kindermann, and J.L. Snell, *Markov Random Field and their applications.*, American Mathematical Society, Providence, R.I., 1980.

[Kit82] L. Kitchen, and A. Rosenfeld, "Gray level corner detection", *Pattern Recognition Letters*, Vol. 1, nº 2, pags. 95 –102, 1982.

[Koe88] J.J. Koenderink, and W. Richards, "Two-dimensional curvature operators.", *Journal of Optical society of America*, Vol. A5, pags. 1136–1141, 1988.

[Koe96] J.J. Koenderink, A.J.V. Doorn, and A.M.L. Kappers, "Pictorial surface attitude and local depth comparisons.", *Perception and Psychophysics*, Vol. 58(2), pags. 163–173, 1996.

[Koe98] J.J. Koenderink, "Pictorial relief.", *Phylosophical Transactions of the Royal Society of London*, Vol. 356, pags. 1071–1086, 1998.

[Kof35] K. Koffka, *Principles of Gestalt psychology.*, Oxford, England: Harcourt, Brace, 1935.

[Kof58] K. Koffka, *Figure-Ground perception in: Readings in perception. Translated from German by M. Wertheimer.*, Princeton, NJ: Van Nostrand, 1958.

[Kog02] N. Kogo, C. Strecha, R. Fransen, G. Caenen, J. Wagemans, and L.J. Van Gool, "Reconstruction of subjective surfaces from occlusion cues", *Biologically Motivated Computer Vision: second workshop of BMVC*, pags. 311–312, 2002.

[Kom07] N. Komodakis, and G. Tziritas, "Image completion using efficient belief propagation via priority scheduling and dynamic pruning", *IEEE Trans. on IP*, Vol. 16, nº 11, pags. 45–78, 2007.

[Kul51] S. Kullback, and R.A Leibler, "On Information and Sufficiency.", *The Annals of Mathematical Statistics*, Vol. 22, nº 1, pags. 79–86, 1951.

[Lag04] R. Laganiere, and R. Elias, "The detection of junction features in images.", *In Proc. of International Conference in Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 3, pags. 573–576, 2004.

[Lan60] A.H. Land, and A.G. Doig, "An Automatic Method for Solving Discrete Programming Problems.", *Econometrica.*, , nº 2, pags. 497520, 1960.

[Lev02] E. Levinia, *Statistical issues in texture analysis.*, PhD Thesis, Berkley, CA, 2002.

[Li89] D. Li, G.D. Sullivan, and K.D. Baker, "Edge detection at junctions.", *In Proc. of Alvey Vision Conference*, pags. 121–125, 1989.

[Lin94] T. Lindeberg, "Junction detection with automatic delection of detection scales and localization scales.", *In Proc. of International Conference on Image Processing (ICIP)*, Vol. 1, pags. 924–928, 1994.

[Liv87] M.S. Livingstone, and D.H. Hubel, "Psychophysical Evidence for Separated Channels for the Perception of Form, color, Movement and Depth.", *The Journal of Neuroscience*, Vol. 7(11), pags. 3416–3468, 1987.

[Low04] D.G. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision (IJCV)*, Vol. 60, nº 2, pags. 91–110, 2004.

[Mad94] S. Madarasmi, and D. Ting-Chuen Pong Kersten, "Illusory contour detection using MRF models", *World Congress on Computational Intelligence*, pags. 4343 – 4348, 1994.

[Mai08] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik, "Using Contours to Detect and Localize Junctions in Natural Images", *International Conference on Computer Vision and Pattern Recognition, CVPR*, pags. 1–8, 2008.

[Mal87] J. Malik, "Interpreting line drawing of curved objects", *International Journal of Computer Vision (IJCV)*, Vol. 1(1), pags. 73–103, 1987.

[Mar82] D. Marr, *Vision*, W. H. Freeman and Company, New York, 1982.

[Mar86] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color and texture cues.", *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 26, pags. 530–549, 1986.

[Mar07] P. Maragos, and G. Evangelopoulos, "Leveling cartoons, texture energy markers, and image decomposition", *Proc. 8th International Symp. on Mathematical Morphology*, Rio de Janeiro (Brazil), October 2007.

[McD04] J. McDermott, "Psychophysics with junctions in real images", *Perception*, Vol. 33, nº 9, pags. 1101 1127, 2004.

[Met74] F. Metelli, "The perception of transparency", *Scientific American*, Vol. 230, pags. 354–366, 1974.

[Met75] W. Metzger, *Gesetze des sehens.*, Waldemar Kramer, 1975.

[Mey97] F. Meyer, "The levelings", *Proc. of International Symposium on Mathematical Morphology (ISMM).*, Vol. 2, pags. 211–214, 1997.

[Mik02] K. Mikolajczyk, and C. Schmid, "An affine invariant interest point detector", *In Proc. of European Conference on Computer Vision ECCV*, pags. 128–142, 2002.

[Mik04] K. Mikolajczyk, and C. Schmid, "Scale and affine invariant interest point detector", *International Journal of Computer Vision (IJCV)*, Vol. 60, nº 1, pags. 63–86, 2004.

[Mon71] N. Montanari, "On the optimal detection of curves in noisy pictures", *Communications of the ACM archive*, Vol. 14, nº 5, pags. 335 – 345, 1971.

[Mor04] P. Mordohai, and G. Medioni, "Junction Inference and Classification for Figure Completion using Tensor Voting", *Computer Vision and Pattern Recognition Workshop, CVPRW*, Vol. 4, pags. 56–64, 2004.

[Mum89] D. Mumford, and J. Shah, "Optimal Approximations of Piecewise Smooth Functions and Associated Variational Problems", *Journal on Communications in Pure and Applied Mathematics*, Vol. 42, pags. 577–685, 1989.

[Nak92] C. Nakayama, and S. Shimojo, "Experiencing and perceiving visual surfaces.", *Science*, Vol. 257, pags. 1357–1363, 1992.

[Nar64] E.A. Naradaya, "On estimating regression.", *Theory of Probability and its Applications.*, Vol. 10, pags. 186–190, 1964.

[Ni09] K. Ni, X. Bresson, T. Chan, and S. Esedoglu, "Local histogram based segmentation using the wasserstein distance", *International Journal of Computer Vision (IJCV)*, Vol. 84(1), pags. 97–111, 2009.

[Nit90] M. Nitzberg, and D. Mumford, "The 2.1-D Sketch", *In Proc. of International Conference on Computer Vision (ICCV)*, pags. 138–144, 1990.

[O'S94] R.P. O'Shea, S.G. Blackburn, and H. Ono, "Contrast as depth cue", *Vision research*, Vol. 34, nº 12, pags. 1595–1604, 1994.

[Ouz05] G.K. Ouzounis, and M.H.F. Wilkinson, "Second order attribute filters using max-trees", *Proc. of International Symposium on Mathematical Morphology, (ISMM)*, pags. 65–74, 2005.

[Par98] L. Parida, D. Geiger, and R. Hummel, "Junctions: Detection, Classification, and Reconstruction", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 20, nº 8, pags. 687–698, 1998.

[Pen85] A.P. Pentland, "A new sense for depth of field", *In Proc. of International Conference on Computer Vision (ICCV)*, pags. 839–846, 1985.

[Per90] P. Perona, and J. Malik, "Scale-space and edge detection using anisotropic diffusion.", *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 12, nº 7, pags. 629–639, 1990.

[Pet93] M. A. Peterson, and B. S. Gibson, "Shape recognition inputs to Figure-Ground organization in three-dimensional displays", *Cognitive Psychology*, Vol. 25, pags. 383–429, 1993.

[Pro99] M. Proesmans, and L.V. Gool, "Grouping based on coupled diffusion maps. contour and grouping in computer vision", *Shape, Contour and Grouping in Computer Vision. LNCS*, pags. 196–216, 1999.

[Rao02] R.P.N. Rao, Olshausen B.A., and Lewicki M.S., *Probabilistic models of the brain: perception and neural function.*, The MIT Press, 2002.

[Ren06] X. Ren, C.C. Fowlkes, and J. Malik, "Figure/ground assignment in natural images", *In Proc. of European Conference on Computer Vision ECCV*, pags. 614–627, 2006.

[Rob65] L. Roberts, "Machine perception of three-dimensional solids.", *Optical and Electro-Optical Information Processing, OEOIP*, pag. 159197, 1965.

[Rom91] B.M. Romeny, L.M. Florack, J.J. Koenderink, and M.A. Viergever, "Invariant third order properties of isophotes: T-junction detection", *Proc. of 7th Scand. Conf. on Image Analysis*, pags. 346–353, 1991.

[Ros92] L. Rosenthaler, F. Heitger, O. Kubler, and R. von der Heydt, "Detection of general edges and keypoints.", *In Proc. of European Conference on Computer Vision ECCV*, pags. 78–86, 1992.

[Rot09] D. Rother, and G. Sapiro, "3d reconstruction from a single image", *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence. IMA PreprInternational*, 2009.

[Rub21] E. Rubin, *Visuell wahrgenommene figuren.*, Kobenhaven: Glydenalske boghandel, 1921.

[Rub98] Y. Rubner, C. Tomasi, and L.J. Guibas, "A metric for distributions with applications to image databases.", *In Proc. of International Conference on Computer Vision (ICCV)*, pags. 59–66, 1998.

[Ruz01] M.A. Ruzon, and C. Tomasi, "Edge, junction, and corner detection using color distributions.", *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2001.

[Sal95] P. Salembier, and J. Serra, "Flat zone filtering, connected operators and filters by reconstruction", *IEEE Transactions on Image Processing*, Vol. 3(8), pags. 1153–1160, 1995.

[Sal98] P. Salembier, A. Oliveras, and L. Garrido, "Anti-extensive onnected operators for image and sequence processing", *IEEE Transactions on Image Processing*, Vol. 7(4), pags. 555–570, 1998.

[Sal00] P. Salembier, and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation and information retrieval", *IEEE Trans. on IP*, Vol. 7(4), pags. 561–576, 2000.

[Sau99] E. Saund, "Perceptual organization of occluding contours of opaque surfaces", *Computer Vision and Image Understanding*, Vol. 76, nº 1, pags. 70–82, 1999.

[Sau05] E. Saund, "Logic and mrf circuits for labeling occluding and thinline visual contours.", *International Conference on Neural Information Processing Systems, NIPS*, 2005.

[Sax07] A. Saxena, Min Sun, and A.Y. Ng, "Learning 3-d scene structure from a single still image", *In Proc. of International Conference on Computer Vision (ICCV)*, pags. 1–8, 2007.

[Shi00] J. Shi, and J. Malik, "Normalized Cuts and Image Segmentation.", *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 26, pags. 530–549, 2000.

[Sin03] M. Singh, and X. Huang, "Computing Layered Surface Representations: An algorithm for Detecting and Separating Transparent Overlays", *International Conference on Computer Vision and Pattern Recognition, CVPR*, pags. 1–8, 2003.

[Sin08] E.D. Sinzinger, "A model-based approach to junction detection using radial energy.", *Pattern Recognition*, Vol. 41, pags. 494–505, 2008.

[Smi97] S.M. Smith, and M. Brady, "Susan - a new approach to low level image processing", *International Journal of Computer Vision (IJCV)*, Vol. 23, nº 1, pags. 45–78, 1997.

[Smi05] P. Smith, T. Crummond, and R. Cipolla, "Generalizing Swendsen-Wang to Sampling Aribtrary Posterior Probability", *IEEE Trans. Pattern Analalysis Machine Intelligence*, Vol. 27, nº 4, pags. 1239–1253, 2005.

[Ste01] X.Y. Stella, T.S. Lee, and T. Kanade, "A hierarchical markov random field model for figure-ground segregation", *International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition, EMM CVPR*, pags. 110–133, 2001.

[Sug84] K. Sugihara, "An algebric approach to shape-from-image problem.", *Artificial Intelligence*, Vol. 23, pags. 59–95, 1984.

[Thi07] S.R. Thiruvenkadam, F. Chan, T, and B.W. Hong, "Segmentation under occlusion using selective shape prior", *Scale Space and Variational Methods in Computer Vision*, Vol. 4485, pags. 191–202, 2007.

[Tom98] C. Tomasi, and R. Manduchi, "Bilateral filter for gray and color images", *In Proc. of International Conference on Computer Vision (ICCV)*, pags. 988–994, 1998.

[Vas98] G. Vasari, *Life of the Artists*, Oxford University Press, 1998.

[Vin93] L. Vincent, "Morphological gray scale reconstruction inimage analysis: Applications and efficient algorithms", *IEEE Transactions on Image Processing*, Vol. 2(2), pags. 176–201, 1993.

[Vin05] J.F. Rigaud (translator) L. da Vinci, *Treatise on Painting*, Dover Publications, 2005.

[Wal75] D.L. Waltz, *The Psychology of COmputer Vision*, Chap. Understanting Line Drawings of scenes with shadows., P. Winston Ed.,McGraw-Hill, New York, 1975.

[Wat64] G.S. Watson, "Smooth regression analysis.", *Sankhya: The Indian Journal of Statistics.*, Vol. 26, pags. 359–372, 1964.

[Wer23] M. Wertheimer, "Untersuchungen zur Lehre der Gestalt, II", *Psycologische Forschung*, Vol. 4, pags. 301–350, 1923.

[Whe38] C. Wheatstone, "Contributions to the physiology of Vision, Part I: On some remarkable and hitherto unobserved, phenomena of binocular vision.", *Phylosophical Transactions of the Royal Society of London*, Vol. 128, pags. 371–394, 1838.

[Wil90] L.R. Williams, "Perceptual organization of occluding contours", *In Proc. of International Conference on Computer Vision (ICCV)*, pags. 133–137, 1990.

[Wue96] S. Wuerger, R. Shapley, and N. Rubin, "On the visual perceived direction of motion, Hans Wallach: 60 years later", *Vision Research*, Vol. 25, pags. 317–367, 1996.

[Wys82] G. Wyszecki, and W.S. Stiles, *Color Science: Concept and Methods, Quantative Data and Formulae.*, New York: John Wiley and Sons, 1982.

[Yar85] L.P. Yaroslavsky, "Digital picture processing - an introduction", 1985, new York: Springer-Verlag.

[Yin04] J. Yin, and J.R. Cooperstock, "Improving Depth Maps by Nonlinear Diffusion", *In Proc. of 12th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, pags. 1–8, 2004.

[Yon79] A. Yonas, *Attached and cast shadows, in Perception and Pictorial Representation*, C.F.Nodine and D.F.Fisher,Praeger, 1979.