Depth Recovery from Stereo Matching Using Coupled Random Fields

Ramya Narasimha

UNIVERSITÉ DE GRENOBLE

N° attribué par la bibliothèque
par ***THÈSE***

pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : "Mathématiques et Informatique"

préparée au laboratoire **Laboratoire Jean Kuntzmann (LJK)**
dans le cadre de **l'École Doctorale**
"Mathématiques, Sciences et Technologies de l'Information, Informatique"

---

préparée et soutenue publiquement par

Ramya Narasimha

le 14 Septembre 2010

Méthodes d'estimation de la profondeur par mise en correspondance stéréoscopique à l'aide de champs aléatoires couplés

---

sous la direction de: Prof. Radu Horaud et Dr. Elise Arnaud

JURY

Prèsident	**Prof. Bernard Ycart**
Rapporteurs	**Prof. Steven W. Zucker**
	Prof. Beatrice Pesquet-Popescu
Éxaminateurs:	**Dr. Marc Sigelle**
	Prof. Nikos Paragios
	Prof. Florence Forbes

Acknowledgements

First of all, I would like to thank my supervisors, Prof. Radu Horaud and Dr. Elise Arnaud, for their help and support during the period of my thesis. My extended thanks to Prof. Florence Forbes for her invaluable insight and helpful discussions.

I am grateful to the members of the jury, Dr. Marc Sigelle, Prof. Steven Zucker, Prof. Nikos Paragios, Prof. Beatrice Pesquet-Popescu and Prof. Bernard Ycart for accepting to read my manuscript and for providing me with feedback. I would especially like to thank my friends and colleagues, Miles Hansard, and Diana Mateus who accepted to review the earlier drafts of this manuscript. I would like to thank Antoine Letouzey for helping me with the French part of my thesis.

Special thanks go to my friend Lamiae Azizi, for her rare hospitality and friendship, for offering me a great place to stay during my thesis writing and for making sure that I am always properly fed!

I am deeply grateful to Benjamin, Regis, and Simone for their technical assistance during the writing of my thesis. Finally, I woulkd like to thank my friends Vasil, Fabio, Pierre, Gaetan, Amael, Avinash, Kiran, Visesh, Julie, Xavi, Antoine (D), and Jan, who made my stay in Grenoble very enjoyable, both from the social and from the research aspects.

Résumé en français

La Stéréo-vision se réfère à la perception visuelle en trois dimensions (3D) des structures du monde vues par les deux yeux. En d'autres termes, grace à la stéréo-vision on a la capacité de distinguer la profondeur relative des différents objets de la scène. Mais comment cela est-il possible ?

Dans le cas des primates, les deux yeux regardent vers l'avant avec un chevauchement important de leur champs de vision. En raison de ce chevauchement, les deux yeux ont une vue presque identique du monde. Cependant, en raison de leur séparation horizontale, ils voient le monde à partir de points de vue légèrement différents. Par conséquent, chacun reçoit une image légèrement différente de la scéne en trois dimensions. La différence entre les positions des points dans les deux images est appelée **disparité binoculaire**. Cette disparité binoculaire est liée aux distances relatives des objets aux yeux.

Au lieu de directement comprendre le système visuel humain, on peut tenter de l'imiter avec de la stéréo-vision par ordinateur. Cela se fait à l'aide de deux caméras, une pour chaque oeil, et d'un ordinateur (qui se comporte comme le cerveau) qui s'interface avec les caméras et traite les images stéréo pour finalement fournir les informations de profondeur. Bien que le modèle stéréo artificiel ne suive pas exactement le système visuel humain, il s'en inspire fortement. Même après 40 ans de recherche en stéréo-vision par ordinateur, il existe encore de nombreux problèmes non résolus qui doivent être abordés. Dans cette thèse, notre objectif est de répondre à certains de ces problème et de proposer des solutions possibles.

La stéréo-vision par ordinateur est utilisée dans de nombreux autres domaines que la simulation de la vision humaine. C'est un outils utilisé dans le cadre de la navigation de robotique, pour extraire la profondeur des objets d'une scène pour permètre de détecter et d'éviter les obstacles qui s'y trouvent. Un des exemples les plus évident d'un tel système est le (Goldberg et al. [2002]) Mars Rover qui a été équipé d'un système de caméras stéréoscopique et qui a roulé sur le sol de la planète Mars. La Stéréo-vision a également trouvé des applications dans la restitution du relief topographique, l'architecture, l'ingénierie, la fabrication et la géologie. Plus récemment, la vision stéréoscopique a été utilisée pour le suivi de milliers de points sur un visage humain ou d'autres surfaces pour l'animation de

personnages. Il y a aussi une demande pour des systèmes temps-réel de vision stéréo pour détecter et suivre la pose de marqueurs pour des applications chirurgicales. En outre, des recherches sont également en cours (Balakrishnan et al. [2007]) pour créer des systèmes portables de stéréo-vision pour l'aide aux personnes mal-voyantes.

Stéréo-vision par ordinateur

La stéréo-vision par ordinateur, que nous appelerons plus simplement stéréo-vision dans la suite, repose la configuration suivante. L'élement principal est le couple de caméras (gauche et droite) qui sont placées l'une à coté de l'autre avec un décalage latéral. Ces caméras capturent la scène à partir de deux points de vue différents, les images obtenues sont ensuite transmises à un ordinateur. L'idée est maintenant de trouver les endroits dans les images qui correspondent au même point physique dans l'espace. Avec cette information, ainsi que la géométrie des caméras stéréo, il est possible de déterminer les emplacements en trois dimensions de tous les points dans l'image. Les principaux problèmes qui doivent être abordées en stéréo-vision sont les suivants: *l'étalonnage , la mise en correspondance stéréo et la reconstruction 3D.* L'**étalonnage** consiste à déterminer les paramètres intrinsèques et extrinsèques de la caméra. Par paramètres intrinsèques, on entend la distance focale, la taille des pixels, et le point principal (où l'axe optique rencontre le plan image), et par des paramètres extrinsèques, les positions relatives et les orientations de chaque caméra. Le **problème de mise en correspondance** consiste à déterminer les emplacements des points dans les images de chaque caméra qui sont les projections d'un même point physique de l'espace 3D. La recherche de ces points est déterminée par la **géométrie épipolaire** des caméras. La géométrie épipolaire est la géométrie projective intrinsèque entre les deux points de vue. Il dépend des paramètres (éventuellement inconnus) intrinsèques et extrinsèques des deux caméras. *En termes de correspondance stéréo, la géométrie épipolaire limite la recherche du point correspondant dans une image à une droite, appelée la* **droite épipolaire**, *dans l'autre image. Autrement dit, la recherche est réduite d'une couverture totale de l'image à une recherche 1D le long de la droite épipolaire. Cette contrainte imposée par la géométrie épipolaire est appelée la* **contrainte épipolaire**. En général, les droites épipolaires sont inclinées. Par conséquent, la recherche de points correspondant prend du temps car les pixels à comparer reposent sur des droites obliques dans l'espace image. Ce problème peut être résolu en transformant les images gauche et droite de telle sorte que les droites épipolaires des deux images sont alignées et parallèles à l'axe horizontal. C'est ce qu'on appelle le processus de **rectification**. Les paramètres de la caméra obtenus à l'aide de l'étalonnage peuvent être combiner avec les informations de correspondance pour reconstruire la scène 3D. Ceci est fait grâce à la **triangulation**, qui établie une relation entre la profondeur et la disparité. étant donné les points correspondants dans les deux images, on peut calculer leur disparité égale à la différence entre leurs positions. Cette disparité est inversement proportionnelle à la profondeur, pour une calibration de la caméra connue. Par conséquent,

en utilisant la carte de disparité et les paramètres de caméra, nous pouvons réaliser la **reconstruction 3D** de la scène. Toutefois, le problème qui reste à résoudre est toujours: *comment trouver les correspondances stéréo?*

Les méthodes utilisées pour la mise en correspondance stéréo

L'objectif est maintenant de mettre en correspondance les points des images gauche et droite après réctification. On compare la similitude des pixels à des endroits candidats, $\mathbf{x} = (x_l, y_l)$ et $(x_r, y_r) = (x_l + d, y_l)$, en faisant varier les valeurs de d. Mais une simple comparaison des intensités dans les images aux deux positions ne suffit pas car il peut y avoir des régions sans texture, des motifs répétitifs et deux images peuvent avoir différents éclairages qui peuvent conduire à des résultats ambigus. Certaines méthodes de mise en correspondance (Hannah [1974], Marr and Poggio [1979], Pollard et al. [1985], Baker and Binford [1981]) utilisent certaines caractéristiques dans les images, comme les bords, les coins ou les contours, car ils sont plus résistants aux changements d'éclairage et produisent des correspondances avec une plus grande certitude. En raison de la large gamme d'applications dans les domaines du rendu d'image et de la modélisation 3D, la plupart des techniques d'aujourd'hui se concentrent sur la recherche d'un ensemble dense de correspondances stéréo. Le problème consistant à trouver ces correspondances denses a été largement étudié. Certaines méthodes suggérent de comparer les intensités de l'image dans le voisinage autour de chaque point dans les deux images. Ces méthodes font l'hypothèse implicite de lissage local dans le choix du voisinage. La distinction entre ces méthodes réside dans les mesures utilisées pour la comparaison. Ces méthodes sont généralement appelées méthodes *locales*. Les méthodes *globales*, d'autre part, définissent une fonction d'énergie qui implique un coût basé sur les intensités des deux images et un terme de régularisation qui force les disparités à être semblables dans un voisinage. Comme le problème de mise en correspondance est mal determiné, une régularisation explicite est nécessaire pour obtenir une solution physiquement plausible. Cette fonction d'énergie est ensuite minimisée afin d'obtenir la carte de disparité finale. La plupart des méthodes récentes utilisent des techniques basées sur la minimisation d'une énergie utilisant dans le contexte d'une modélisation probabiliste. L'idée principale de ces méthodes consiste à utiliser les *Champs de Markov* (MRF), ainsi que l'inférence bayésienne pour modéliser les disparités. D'une part les champs de Markov définissent les interactions locales entre les disparités et d'autre part l'inférence bayésienne permet à ces interactions d'être incluses dans une distribution a priori. Cette distribution a priori définie l'homogénéité des disparités et le coût dérivant des intensités dans le couple d'images stéréo est présenté comme la vraisemblance. Le théorème de Bayes permet de calculer la distribution a posteriori en utilisant les a priori et la vraisemblance. L'avantage d'utiliser une approche bayésienne est qu'elle offre une approche prometteuse pour ces problèmes mal contraints parce qu'elle traite le problème comme un problème d'inférence afin de trouver l'estimation optimale. Dans de telles approches bayésienne-MRF l'objectif est alors

de maximiser la probabilité a posteriori sur toutes les cartes de disparité possibles. Cette maximisation/minimisation nécessite l'utilisation de techniques d'optimisation tels que, le recuit simulé, champ moyen, la propagation des croyances ou Graph-cuts. Même si ces techniques de modélisation saisissent les interactions locales entre les disparités voisines et permettent d'intégrer les informations stéréo, certains problèmes cruciaux demeurent:

- Certaines zones de la scène qui sont visibles dans une image peuvent être *ocultés* dans l'autre et cela peut conduire à des mises en correspondance incorrectes.
- Le terme de régularisation du modèle pourrait lisser toutes les disparités et conduire à de mauvaises solutions aux limites des objets de la scène.
- Intégrer uniquement les intensités des image stéréo et le terme de lissage condiurait à modéliser des disparités qui peuvent ne pas être compatible avec les propriétés géométriques de la surface.

Afin de prendre en compte ces limitations, certaines informations supplémentaires sont nécessaires dans la modélisation du problème de mise en correspondance. Des information monoculaires tels que, le gradient, les contours ou l'information concernant la couleur d'une image peuvent être utilisées dans le modèle pour fournir de meilleures solutions pour la disparité. En outre, certaines contraintes géométriques supplémentaires doivent être incorporées pour obtenir des solutions valables vis-à-vis de la surface.

Contributions de la thèse

Dans cette thèse, nous nous concentrons sur l'extension des contraintes pour le problème de mise en correspondance stéréo à partir d'indices monoculaires et de contraintes extra-géométrique. A cette fin, nous proposons ce qui suit:

Estimation conjointe de la disparité et des frontières des objets

La première méthode se propose d'estimer conjointement les disparités et les bordures des objets dans un cadre probabilisteunifié. L'idée ici est de s'attaquer au problème de la localisation des discontinuités dans la carte de disparités qui correspondent aux bords des objets dans le monde réel, ainsi que celle de l'estimation des disparités. Ce schéma implique l'incorporation de l'informations venant des gradients d'intensité dans chaque image comme information monoculaire. Tandis que les disparités sont détectés en utilisant les informations stéréo (images gauche et droite), les indices monoculaires aideront à corriger la disparité au niveau des discontinuités et à trouver les limites des objets. Nous modélisons les informations stéréo et monoculaire dans un cadre MRF unifié. Cette partie de nos contributions a été publié dans l'article Narasimha et al. [2008].

Estimation de disparités compatibles avec la surface

La deuxième méthode tient compte des contraintes de surfaciques différentielle dans le modèle de disparité. Ces contraintes proviennent des normales de la surface dans l'espace des disparités. L'idée est de modéliser les disparités de telle manière à ce qu'elles se trouvent sur le plan défini par les normales de la surface. Cette contrainte conduit à des solutions qui sont compatibles avec les propriétés géometrique de la surface. L'idée est d'estimer simultanément la disparité et les normales de la surface, en tenant compte explicitement de leurs influences réciproque. Cela se fait par la modélisation à la fois des disparités et des normales dans un cadre unifié. Ces travaux ont été publiés dans Narasimha et al. [2009] et Narasimha et al. [2010].

Modélisation conjointe probabiliste à partir des champs aléatoires couplés

Le défi majeur des deux méthodes mentionnées ci-dessus est d'intégrer les informations et les contraintes dans un cadre probabiliste unifié, dans lequel la relation entre les disparités et les variables étudiées (les limites des objets ou les normales à la surface) peuvent être établies de façon explicite. A cet égard, nous utilisons l'idée de couplage des champs de Markov, qui permet de modéliser plus d'une variable aléatoire. Ce type de modélisation permet de rendre explicite l'influence d'une variable sur une autre dans le modèle. Une telle méthode probabiliste permet aussi d'utiliser des techniques d'optimisation séparées pour maximiser les distributions a posteriori portant sur chacune des variables ; Elle donne plus de souplesse dans la modélisation et l'optimisation. Une procédure de maximisation alternative est ensuite utilisée pour réaliser une optimisation globale.

Résumé des chapitres

Nous allons maintenant donner une brève description des chapitres ultérieurs. Les chapitres détaillant les principales contributions de la thèse sont indiqués par $\star$. L'organisation de la thèse est la suivante:

Dans le **chapitre 2**, nous proposons un bref aperçu de la littérature du domaine de la mise en correspondance stéréo. Dans ce chapitre, nous fournissons également une introduction aux champs de Markov (MRF) et nous concentrons principalement sur les techniques qui utilisent des modèles MRF pour l'appariement stéréo. Nous fournissons un bref résumé des techniques d'optimisation pour les MRF utilisés pour estimer les disparités. En outre, nous discutons de certaines ambiguités du problème d'appariement stéréo, et les méthodes existantes pour les surmonter. Nous montrons que l'utilisation des informations monoculaires et géométriques sont importants pour la résolution du problème de correspondance stéréo. Cette discussion introduit les motivations des méthodes proposées de la thèse.

Dans le **chapitre 3**, nous décrivons l'idée de champs de Markov couplés (MRF couplés) dans le cadre de nombreuses applications telles que, la restauration d'image, l'estimation des frontières, la segmentation d'images, la segmentation de texture, ainsi que dans l'appariement stéréo. Nous présentons également une stratégie d'optimisation appelée maximisation alternative. Ces modèles et méthodes fournissent la plate-forme sur laquelle les algorithmes proposés dans cette thèse sont construits.

⋆ Dans le **chapitre 4**, nous présentons une méthode pour intégrer les informations monoculaires pour estimer conjointement les bords des objets les disparités dans un cadre probabiliste unifié. Nous utilisons les MRF couplés pour modéliser les disparités stéréo et l'information sur les limites des objets. Nous montrons que ce modèle permet une amélioration mutuelle dans l'estimation des disparité et des frontières. Nous utilisons la maximisation alternative qui fournit une méthode efficace pour estimer les disparités et les frontières. Enfin, les résultats obtenus en utilisant la méthode proposée sont présentés et discutés.

⋆ Dans le **chapitre 5**, nous proposons une méthode pour obtenir une estimation des disparités qui soit compatible avec l'information de surface. Nous le faisons en étendant les contraintes sur le problème d 'appariement stéréo pour intégrer l'information géométrique venant de la surface. Nous présentons l'importance de ces contraintes et montrons que l'ajout de ces contraintes implique l'estimation des normales à la surface. Nous montrons la relation entre les normales à la surface dans l'espace de profondeur et dans l'espace disparité. Nous proposons ensuite un modèle couplé pour estimer simultanément les disparités et les normales à la surface dans l'espace des disparités. La procédure de maximisation alternative utilisée pour estimer les deux variables est discutée et les résultats sont présentés.

Dans le **chapitre 6**, nous résumons les méthodes proposées dans la thèse et mettons en évidence certains des aspects importants qui les concernent. Nous concluons la thèse en fournissant des orientations pour des travaux futurs.

Conclusion

Nous fournissons un bref résumé des deux modèles présentés dans les chapitre 4 et 5 de cette thèse, respectivement. Nous énumérons les caractéristiques communes des deux modèles et, finalement, nous concluons en citant quelques pistes de recherches futures.

Caractéristiques spécifiques de chacune des approches proposées

Caractéristiques de l'estimation de la disparité et des frontières

Nous réalisons une estimation conjointe des disparités et des frontières des objets par l'unification des deux tâche dans un cadre unifié de Markov. Nous définissons un modèle probabiliste commun original qui nous permet d'estimer les disparités à travers un modèle MRF couplés. L'estimation des frontières aide à l'estimation des disparités afin d'améliorer progressivement et conjointement la précision des deux estimations. Les retours de l'estimation des frontières envers l'estimation des disparités se fait par le champ auxiliaire dénommé le champ de déplacement. Ce champ indique les corrections qui doivent être appliquées aux discontinuités dans la carte de disparité dans le but de les aligner avec les limites des objets. Le modèle commun est un MRF lorsque nous considérons les disparités et se réduit en une chaîne de Markov lorsque nous nous concentrons sur le champ de déplacement. Les caractéristiques spécifiques à ce modèle sont les suivantes:

- Le cadre MRF couplés a été présenté, impliquant deux MRFs : un pour les disparités, l'autre pour le champ de déplacement.
- L'influence de l'estimation des frontières a été encodée dans le champ de déplacement pour représenter les directions dans lesquelles des corrections doivent être appliquées aux disparités.
- Le point central de l'idée d'appliquer des corrections sur les disparités au niveau des discontinuités, provient de l'hypothèse que les discontinuités de la disparité se produisent à proximité des frontières réelles et que les discontinuités de profondeur sont en fait les limites des objets.
- Ces corrections ont été intégrées dans le MRF des disparités utilisant la notion de voisinage adaptatif, qui est capable de traiter des systèmes de voisinage non-standard.
- Le champ de déplacement a été réduit à une chaîne de Markov de second ordre qui n'est active qu'aux discontinuités des disparités. Cela nous permet de trouver les vraies positions des frontières des objets sur la base des corrections appliquées au niveau des discontinuités des disparité.
- L'inférence approximative des disparités a été réalisé en utilisant des algorithmes standards tels que BP ou champ moyen et l'inférence exacte du champ de déplacement en utilisant l'algorithme de Viterbi.
- L'algorithme global permet l'extraction simultanée des frontières des objets et les disparités grâce à l'utilisation d'une information monoculaire simple, dans notre cas, le gradient de l'image.

Caractéristiques de l'estimation des disparités et des normales

Le but de ce second algorithme est de récupérer les disparités en conformité avec les propriétés de surface de la scène étudiée. Pour ce faire, nous estimons les disparités ainsi que les normales dans l'espace disparité, en définissant les deux tâches dans un cadre unifié. Nous avons défini un nouveau modèle probabiliste commun à travers deux champs aléatoires pour favoriser à la fois les cohérences intra-champ (entre les disparités voisines, et entre les normales voisines) et inter-champs (entre les disparités et les normales). L'information géométrique est introduite dans les modèles à la fois pour les normales et les disparités puis elle est optimisée en utilisant une procédure de maximisation alternative appropriée. Le cadre général a les caractéristiques suivantes:

- Un modèle de disparités et de normales à la surface, avec les deux variables modélisées comme des champs aléatoires conditionels (CRF) , a été proposé. Ces CRF sont couplés pour intégrer l'influence réciproque de chaque variable.
- Les deux modèles ont été construits sous l'hypothèse que la scène étudiée est composée de surfaces lisses par morceaux.
- Le CRF des disparités a été défini de telle sorte que le terme d'interaction implique des dérivées au premier ordre des disparités, forçant ainsi les disparités proche à reposer sur un même plan. Ces dérivés ont été extraites du modèle de la normale.
- L'optimisation du CRF des disparités a ensuite été réalisée en utilisant des algorithmes de champ moyen.
- Deux modèles ont été présentés pour les normales, une discrète et une continue.
- Le modèle discret des normales requiert une discrétisation de l'espace des normales et a été optimisé en utilisant des BP. Cependant, ce modèle nécessite une discrétisation dense de l'espace des normales et s'est donc avéré inefficace lors de l'optimisation.
- Le modèle continu quant à lui fourni une meilleure alternative pour le modèle CRF des normales. Le modèle permet l'utilisation de l'algorithme ICM pour l'extraction des normales.

Les caractéristiques partagées par les deux approches proposées

Les deux approches présentées dans cette thèse partagent un certain nombre de caractéristiques communes, telles que:

- Nous proposons des modèles qui, dans un cadre probabiliste, ont permis des distributions conditionnelles qui peuvent modeler explicitement les relations entre deux variables.
- Les deux distributions conditionnelles ont améiloré la flexibilité du modèle global dans le sens où elles peuvent être dépendante ou indépendante en fonction de l'information intégrée.

- L'utilisation de la technique de maximisation alternative pour l'optimisation des deux champs conduit à l'amélioration mutuelle des deux variables en question.
- L'utilisation de l'approche multi-grille dans l'optimisation, permet une interaction à longue portée dans un treillis, sans pour autant réduire la résolution de l'image, mais seulement celle des coûts.
- Les approches proposées ont l'avantage supplémentaire de faire une distinction claire entre le modèle probabiliste et la procédure d'optimisation ultérieure. Des techniques d'optimisation séparées et pré-existentes peuvent être utilisées pour déduire les variables associées à chacun des champs aléatoires.

D'autres directions de recherche

Les deux modèles proposés dans cette thèse traitent de deux questions importantes que sont la localisation des discontinuité dans les disparités et l'extraction des surfaces de disparités. La suite naturelle de ces recherches est de combiner ces deux modèles afin que nous puissions en même temps évaluer les disparités au niveau des discontinuités, extraire la surface des disparités en utilisant des contraintes géométriques et obtenir les frontières des objets. Il s'agirait de gérer trois champs aléatoires, chacun d'eux agissant sur les deux autres d'une manière différente. Etant donné que les contraintes géométriques forcent les disparités à être plus lisses et le modèle des frontières introduit des discontinuités, leur combinaison n'est pas triviale.

En ce qui concerne la forme probabiliste des deux techniques proposées, nous nous sommes concentrés sur la définition valide d'un cadre unifié pour modéliser les coopérations et à utiliser le principe du MAP pour l'inférence. Ce modèle peut être approfondi par la refonte de nos approches dans un cadre Expectation-Maximisation (Dempster et al. [1977]). Un cadre de type EM fournirait également une procédure, théoriquement correctement bien fondée, pour l'estimation des paramètres.

Depth Recovery from Stereo Matching Using Coupled Random Fields

Contents

Chapter 1

Stereo Vision

Stereo vision refers to the visual perception of three dimensional (3D) structures of the world when seen by the two eyes. In other words, with stereo vision one has the ability to distinguish the relative depth of different objects in the scene. But how exactly is this achieved?

In the case of primates, they have two eyes facing forward with a large overlap in their fields of view. As a result of this overlap, the two eyes see almost identical views of the world. However because of the horizontal separation between them they see the world from sightly different vantage points. Each eye receives a slightly different picture of the three dimensional scene around us. The dissimilarity is the difference in the positions of the points in two images, which is referred to as **binocular disparity**. This binocular disparity is related to the relative distances of the objects from the eyes.

This fact can be verified through a quick experiment; hold two fingers up, one in front of the other. While fixating on the closer finger, alternately open and close each eye. One notices that the farther the far finger is from one's eyes (while not moving the near finger), the greater is the lateral shift in its position as one opens and closes each eye. Therefore, it can be concluded that, the difference in line-of-sight shift manifests itself as *disparity* between the left and right eye images.

Although this explanation seems obvious, it was not until 170 years ago that this fact was completely understood. While simple facts about binocular disparities were known to people since ancient times, these were considered as obstacles rather than the basis of stereo vision. Leonardo da Vinci had correctly observed that the two eyes received different views of the 3D scene. But he could not explain how one could see single world of the objects given these different views. Vieth and Müller later discovered the idea of *horopter*, which referred to the locus in space within which the object must lie to appear fused. However, they disregarded the binocular disparities as being too small to be noticed by the two eyes. It was Charles Wheatstone (Wheatstone [1838, 1852]), who through the invention of the

stereoscope demonstrated the significance of binocular disparity. He stated the principle of the stereoscope as follows:

"It being established that the mind perceives an object of three dimensions by means of two dissimilar pictures projected by it on the retina, the following question occurs. What would be the visual effect of simultaneously presenting to each eye, instead the object itself, its projection on a plane surface as it appears to that eye? To pursue this enquiry it is necessary that means should be contrived to make the two pictures, which must necessarily occupy different places, fall on similar parts of both eye."

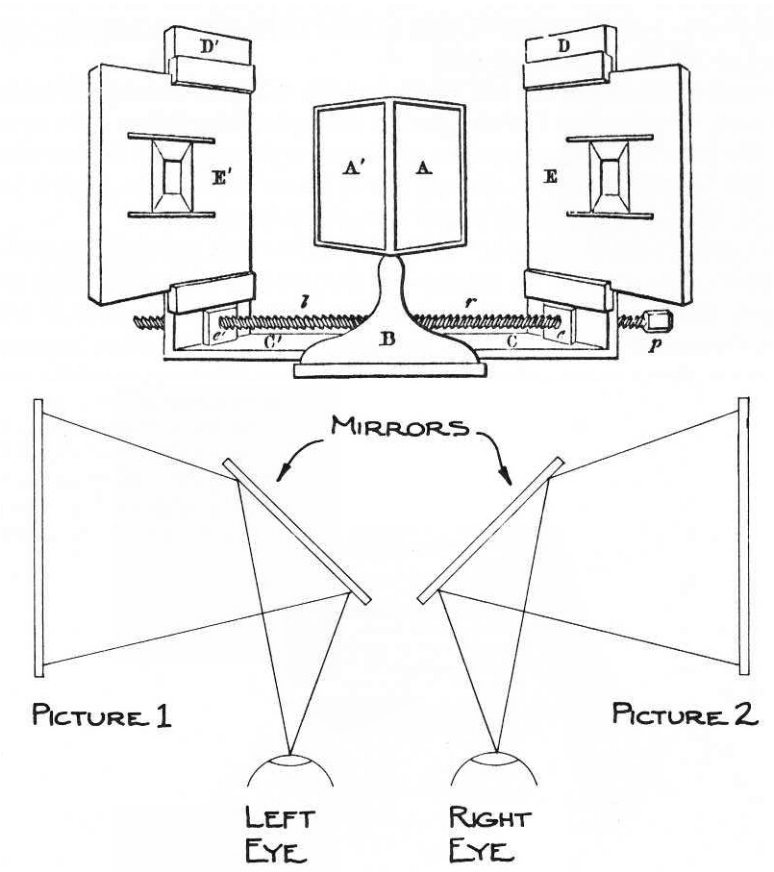


Figure 1.1: Wheatstone's mirror stereoscope: Two mirrors at A', A reflect the drawings at E', E and produce a 3D relief when viewed simultaneously from very close range. Reproduced from Wheatstone [1838].

The stereoscope, he invented, mainly consisted of two mirrors at right angles and two vertical picture holders (see figure 1.1). He then presented a series of line drawings, shown in figure 1.2, of simple objects with perspectives corresponding to right and the left eye,

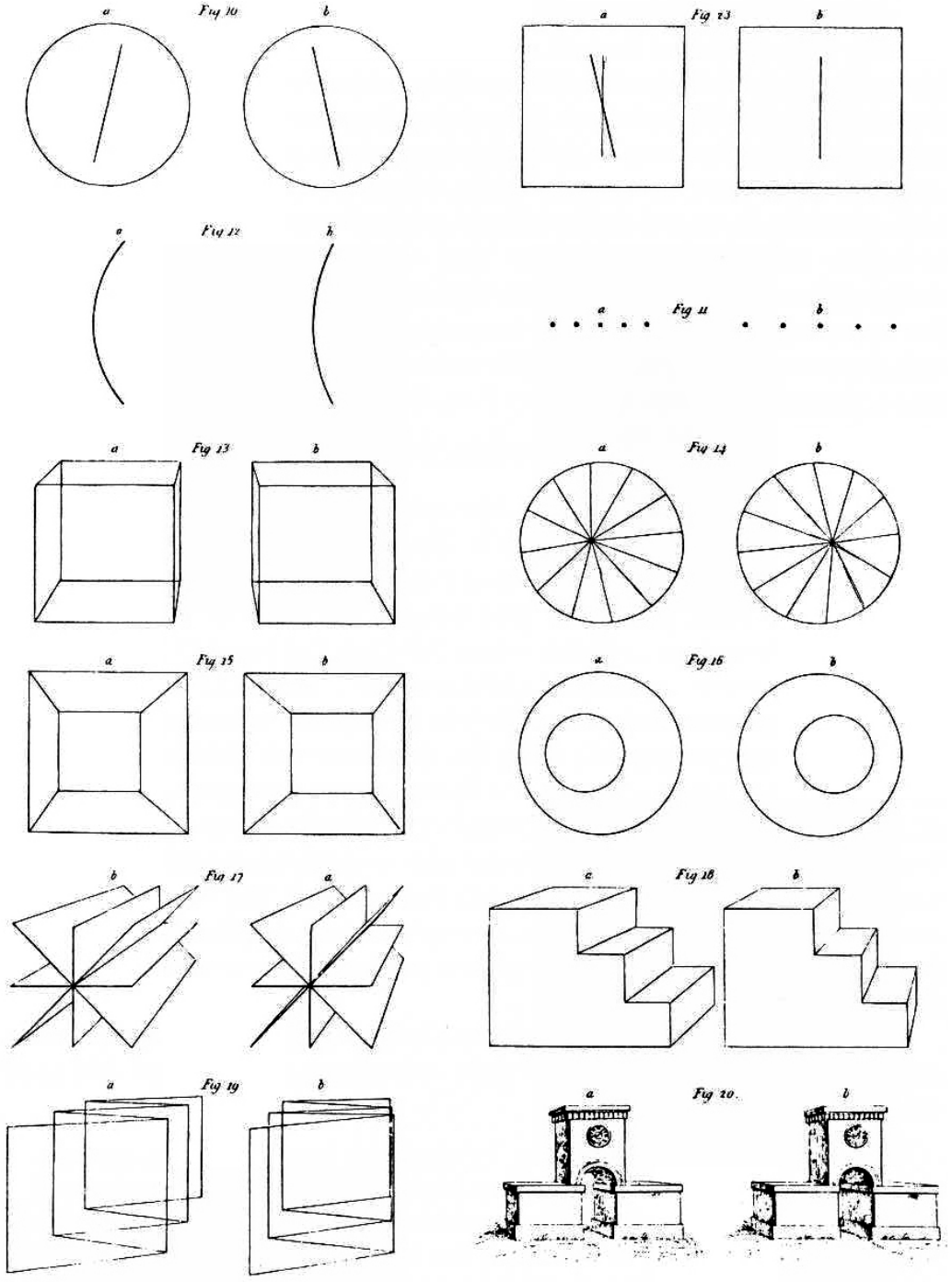


Figure 1.2: Line drawings presented each eye through the stereoscope. Reproduced from Wheatstone [1838].

separately to each eye. He observed that all these objects appeared (i) in three dimensions when the pictures for left and right eyes corresponded, (ii) appeared flat when pictures in the two eyes were the same and, (iii) appeared in reverse depth when the pictures with disparity were reversed to the two eyes.

These experiments showed that the brain has the capability of measuring the difference between the images from the two eyes and creating a sensation of depth. But how exactly does the brain measure these differences? Helmholtz (Helmholtz [1925]) conjectured that as the task of comparing the images from the two eyes is too complex, the brain first recognizes the forms of the objects (like the object contours) and compares them before extending to the entire scene. About 100 years after Julez [1959] (later in Julesz [1971]) through the use of computer generated random-dot stereo images, also referred to as stereograms, demonstrated that one could see stereo even in the absence of object contours. The random-dot stereograms he generated, consisted of slightly shifted hidden pattern that was not visible to each eye separately, but could be seen as 3D structure using a stereoscope. Julez therefore concluded that binocular disparity was in fact a low-level cue like object contours or edges. However, Ramachandran and Nelson [1976] showed that it was not the only cue used by the brain. They found that brain while it was capable of extracting depth without the aid of monocular cues like object boundaries, it sometimes did use such cues under noisy or camouflaged environment. While a lot of research has been carried out to understand the way the stereo processing is performed in the brain, it is still a very complex problem and we are very far from the true answer (Cumming and DeAngelis [2001]).

Alternatively, one can take a computational point of view to the problem of stereo vision. In computational stereo vision, instead of directly understanding the human visual system, the goal is to imitate it. This is done by using two cameras for the two eyes and a computer (that behaves like the brain) which interfaces with the cameras and processes the stereo images to finally provide the depth information. While the artificial stereo system does not entirely follow the human visual system it takes ideas from it which are useful in determining the solution. Even with 40 years of research in computational stereo vision there are still many unsolved problems which need to be addressed. In this thesis, our goal is to address some of these issues and to provide possible solutions.

Apart from trying to imitate the human visual system, computational stereo has many other applications. It has been used in robot navigation, to extract the depth of the objects in the scene for obstacle avoidance. One the most conspicuous examples of such a system is the Mars Rover (Goldberg et al. [2002]) which was equipped with a stereo camera to navigate through the Mars terrain. Stereo vision has also found its applications in topographic relief mapping, architecture, engineering, manufacturing and geology. More recently stereo vision is used for tracking of thousands of points on the human face or other surfaces for character animation. There is also a demand for real-time stereo vision systems to detect and track the pose of markers for surgical applications. Furthermore, research is also being done (Balakrishnan et al. [2007]) to create human-wearable stereo vision systems to assist the blind.

1.1 Computational stereo vision

Computational stereo vision, hereafter referred to as just stereo vision, has the following set up. It mainly consists of two cameras (the left and right cameras) which are placed next to each other with a lateral shift. A typical stereo camera system is shown in figure 1.3. computer which converts images into a digital form. These cameras capture the scene from two different viewpoints, which are then fed to a computer. The idea now is to find the locations in both the images that correspond to the same physical point in space. With this information, along with the geometry of the stereo set up, it is possible to determine the three-dimensional locations of all the points in the image. Primary problems that have to be addressed in stereo vision are the following: *calibration, stereo correspondence and* $3D$ *reconstruction.*

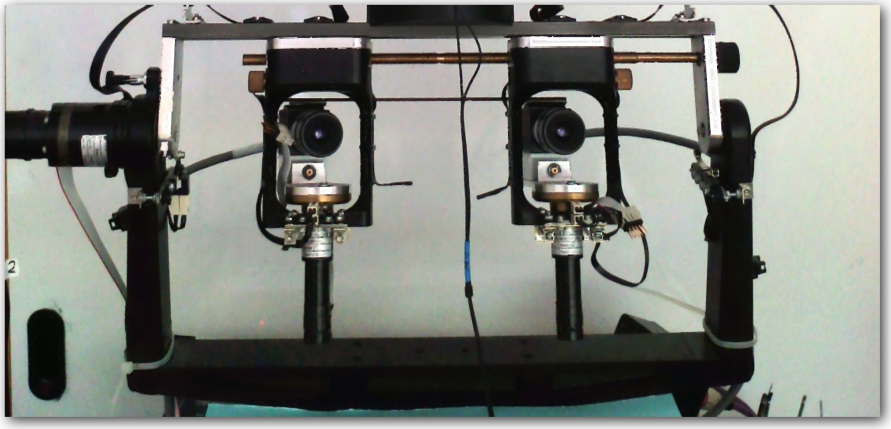


Figure 1.3: Two cameras are placed next to each other with a lateral shift to resemble the human eyes.

Camera calibration involves determining the intrinsic and extrinsic parameters of the camera. By intrinsic parameters, we mean the focal length, size of the pixels, and principal point (where the optical axis meets the image plane), and by extrinsic parameters, the relative positions and orientations of each camera. The location of the focal length, and principle point must be determined to relate the image pixel coordinates to positions in the image plane (see figure 1.4). The position and orientation of the cameras must be determined to relate image plane coordinates to the world coordinate system (or absolute coordinates) in which the camera resides (see figure 1.5). Therefore, the extrinsic parameters are required to determine the rigid body transformation between the two cameras and the intrinsic parameters to model the imaging process inside each camera. The problem of estimating the calibration (parameters) is, at this point, well understood and high-quality toolkits are available (e.g., Bouguet [2008] and links therein). For a more detailed discussion on this

topic, see Faugeras [1993], Hartley and Zisserman [2000] and Zhang [2000]. Throughout this thesis, we assume the camera calibration to be fixed and the parameters known.

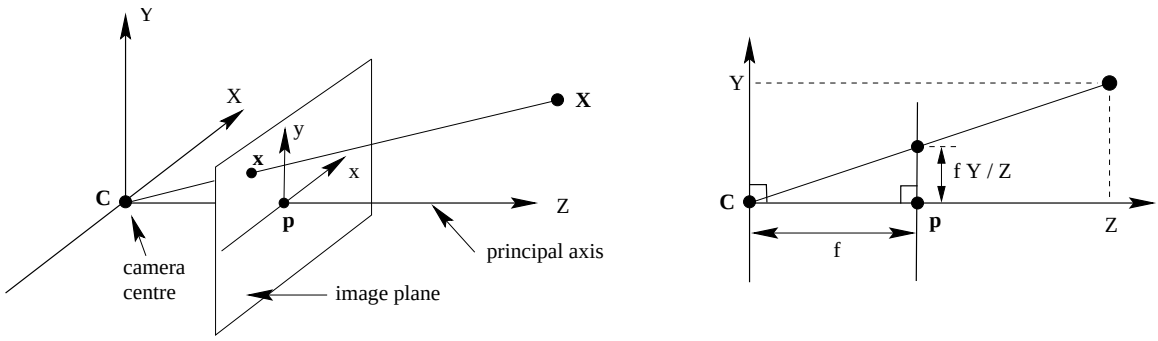


Figure 1.4: Intrinsic parameters involves the optical centre C, focal length f and the principle point **p**. Reproduced from Hartley and Zisserman [2000].

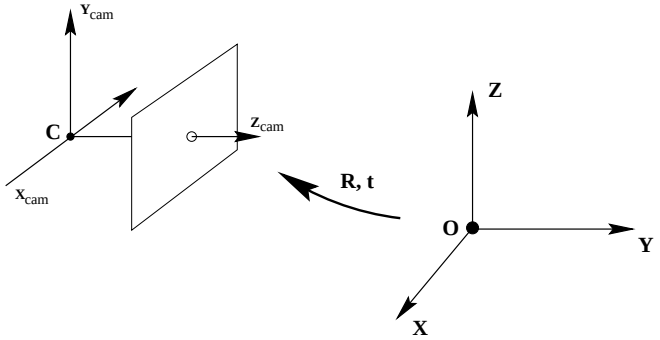


Figure 1.5: Extrinsic parameters relate the camera coordinate frame to world frame using rotation **R** and translation **t**. Reproduced from Hartley and Zisserman [2000].

The **correspondence problem** consists of determining the locations in each camera image that are the projection of the same physical point in 3D space. The search for the corresponding points is determined by the **epipolar geometry** of the cameras. The epipolar geometry is the intrinsic projective geometry between the two views. It depends on the (possibly unknown) intrinsic and extrinsic parameters of the two cameras. The camera parameters obtained using calibration can then be used along with correspondence information to reconstruct the 3D scene. This is done using **triangulation** and the whole process is referred as 3**D reconstruction**.

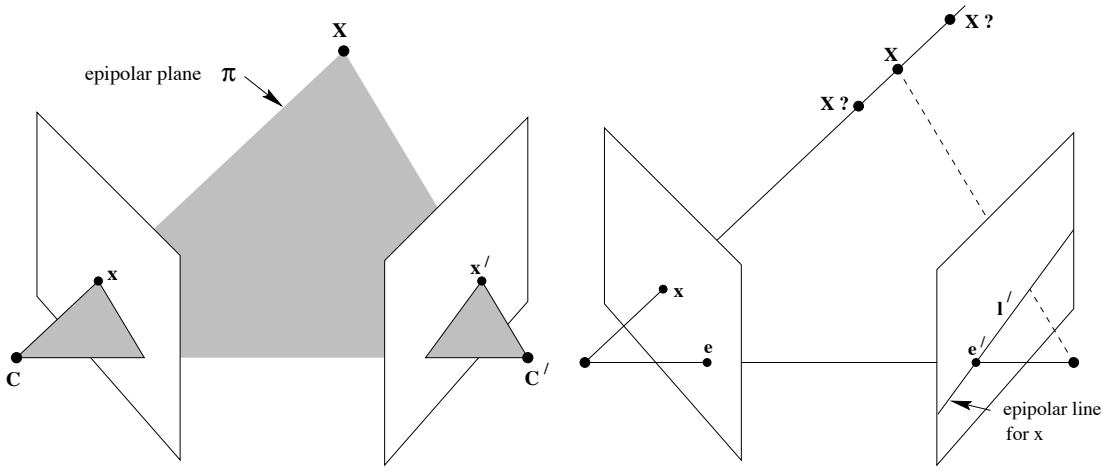
1.1.1 Epipolar constraint

When a point $\mathbf{X}$ in 3D space is imaged by a stereo camera pair, it projects as points $\mathbf{x}$ and $\mathbf{x}'$ in the left and right images respectively. As shown in the figure 1.6, the image points $\mathbf{x}$ and $\mathbf{x}'$, 3D point $\mathbf{X}$ and the camera centres $\mathbf{C}$ and $\mathbf{C}'$ are all co-planar. This plane is referred to as the *epipolar plane* and is denoted by π. The line joining the camera centres is referred to as the *baseline*. The point of intersection of the baseline and the image plane is called the *epipole* ($\mathbf{e}$ and $\mathbf{e}'$ shown in figure 1.6(b)). If we consider only the point $\mathbf{x}$ in the left image, its corresponding point must lie on the line $\mathbf{l}'$ where the epipolar plane intersects the right image plane, because of the co-planarity (see figure 1.6(b)). This line $\mathbf{l}'$ is called as the *epipolar line* corresponding to $\mathbf{x}$. Similarly, $\mathbf{l}$ is the epipolar line in the left image corresponding to $\mathbf{x}'$. All the epipolar lines $\mathbf{l}$ in the left image pass through the epipole $\mathbf{e}$ and similarly all epipolar lines $\mathbf{l}'$ in right image pass through $\mathbf{e}'$. In terms of stereo correspondence, the epipolar geometry constrains the search for the point corresponding to $\mathbf{x}$ to the line $\mathbf{l}'$. That is, the search is reduced from covering the entire image to a 1D search along the epipolar line $\mathbf{l}'$. This constraint imposed by the epipolar geometry is called the **epipolar constraint**

In general, the epipolar lines are slanted. Therefore, the search for corresponding points is time consuming as the pixels on skewed lines in the image space have to be compared (figure 1.8). This problem can be overcome by transforming the left and the right images such that the epipolar lines of the two are collinear and parallel to horizontal axis. This process is referred to as **rectification**. The pixels corresponding to point features from a rectified image pair will lie on the same horizontal scan-line and differ only in horizontal displacement. This horizontal displacement, or *disparity* between rectified feature points is related to the depth of the feature. recover 3D structure from 3D geometry notions like cameras.

1.1.2 Rectification

The goal of rectification is to re-sample the left and the right images such that the epipolar line in the re-sampled images run parallel and the disparities between the images are in the horizontal direction only. In order to do so a pair of $2D$ *projective transformations* are applied to the two images. In the calibrated case, the projective transformations simulate the effect of *rotating* the cameras. These rotations make both (the left and right camera) the optical axes perpendicular to the baseline. Let the projective space $\mathbb{P}^2$ represent the set rays passing through the origin of the 3D space. The points in this space, $\mathbb{P}^2$, are 2-dimensional represented as homogenous 3-vectors, for example a 2D point (x, y) is represented as $\mathbf{x} \simeq (x, y, 1)$, where $\simeq$ stands for equality upto a non-zero scale. Given a set of points $\mathbf{x}_i \in \mathbb{P}^2$ and corresponding points set of points $\mathbf{x}'_i \in \mathbb{P}^2$, the projective transformation H gives the mapping between the two, i.e., $\mathbf{x}' \simeq H\mathbf{x}$. Each one of the stereo images can be considered as a projective plane $\mathbb{P}^2$. The rectification process then involves finding two projective



(a) Epipolar plane (b) Epipolar line

Figure 1.6: Epipolar geometry images reproduced from the book by Hartley and Zisserman [2000].

transformations that maps the epipoles of each image to a point at infinity along the x-axis, i.e., $(1, 0, 0)^T$, subject to certain constraints. When the epipole is at infinity, all epipolar lines are horizontal. For more details on the exact procedure of rectification refer to Hartley and Zisserman [2000]. Other methods for rectification can also be found in Fusiello et al. [2000], Loop and Zhang [1999] and Faugeras [1993].

The rectification process therefore converts general camera configuration to a simplified one, as shown in the figure 1.7. We show the rectified images with parallel epipolar line corresponding to the figure 1.8 in figure 1.9. With the rectified camera configuration now the search for the corresponding points is only along the horizontal scan-line. Suppose $\mathbf{x} = (x_l, y_l)$ is point in the left image and $\mathbf{x}' = (x_r, y_r)$ is the corresponding point in the right image. The disparity in the configuration is the difference in the x-coordinates of the corresponding left and right pixel locations,i.e., $d = x_l - x_r$. We now show how this disparity is related to the depth in 3D space.

1.1.3 Disparity-depth relationship: Triangulation

In order derive relationship between disparity and the depth we refer to the figure 1.7. In this figure, $\mathbf{X}$ represents a scene point, which project on to $\mathbf{x} = (x_l, y_l)$ in the left image and $\mathbf{x}' = (x_r, y_r)$ in the right image. If coordinates (X, Y, Z) represent the point $\mathbf{X}$ in the left camera coordinates, then because of the rectification of the corresponding position in the right camera coordinate is $(X - b, Y, Z)$ where b is the baseline. The image points and

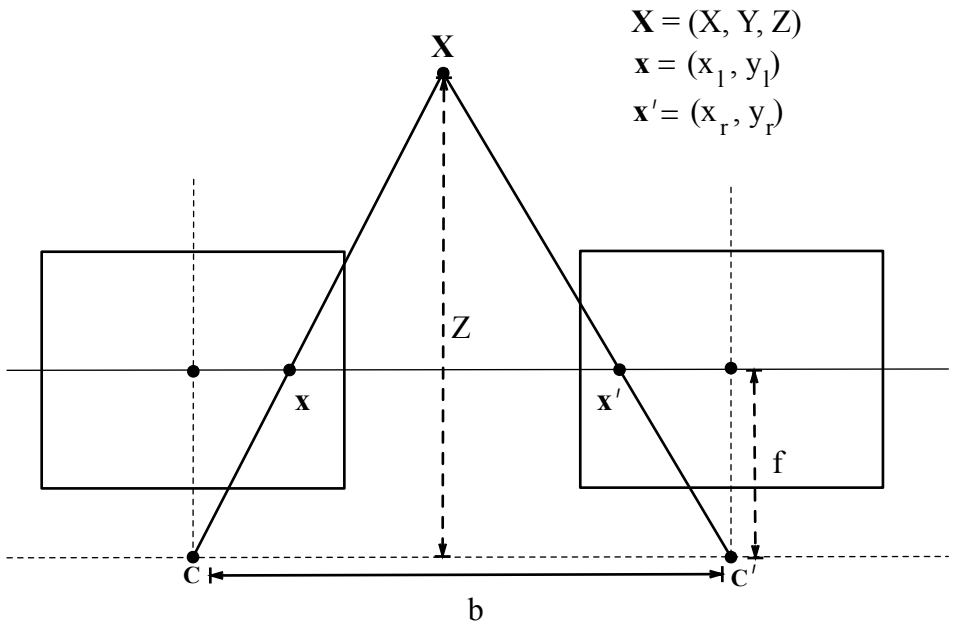


Figure 1.7: After rectification: The scene point $\mathbf{X}$ projected as $\mathbf{x}$ in the left image and $\mathbf{x}'$ in the right image. As can be seen because of rectification the shift between corresponding points $\mathbf{x}$ and $\mathbf{x}'$ is only along the horizontal scanline.

the camera coordinates are related as follows:

$$x_l = f\frac{X}{Z} \quad \text{or} \quad X = \frac{Zx_l}{f} \tag{1.1}$$

$$y_l = f\frac{Y}{Z} \quad \text{or} \quad Y = \frac{Zy_l}{f} \tag{1.2}$$

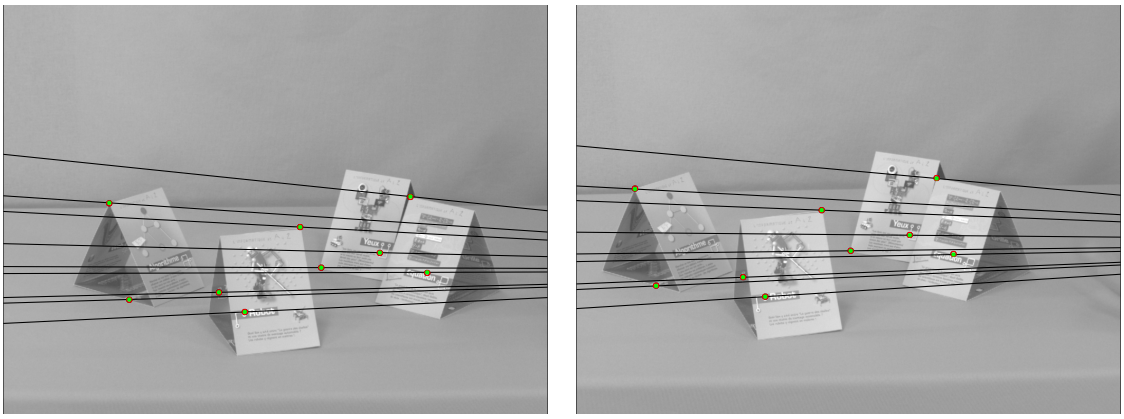
$$x_r = f\frac{X-b}{Z} \quad \text{or} \quad X - b = \frac{Zx_r}{f} \tag{1.3}$$

$$y_r = f\frac{Y}{Z} \quad \text{or} \quad Y = \frac{Zy_r}{f} \tag{1.4}$$

where f represents the focal length of the camera and is assumed to be the same for both. Using the above relationships, we can write:

$$\frac{x_l Z}{f} - b = \frac{x_r Z}{f} \tag{1.5}$$

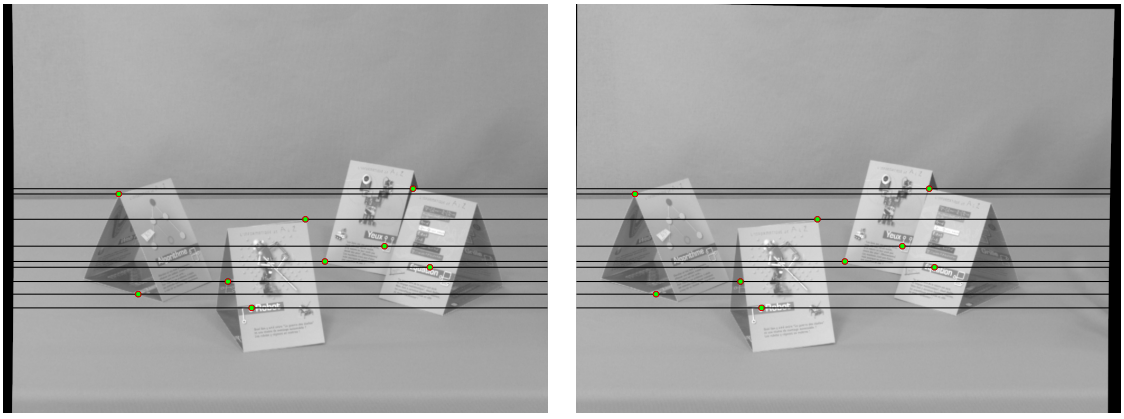
$$Z = f\frac{b}{x_l - x_r} \tag{1.6}$$



(a) Epipolar lines depicted on an unrectified left image

(b) Epipolar lines depicted on an unrectified right image

Figure 1.8: Epipolar lines on real stereo image pair. As the image shows the epipolar lines are in general slanted.



(a) Epipolar lines depicted on an rectified left image

(b) Epipolar lines depicted on an rectified right image

Figure 1.9: Rectification process allows to make epipolar line parallel and thereby reducing the search for corresponding points to be along scanlines

If we denote the *disparity* by $d = x_l - x_r$, then disparity is related to depth as follows:

$$Z = f\frac{b}{d} \tag{1.7}$$

From the above equation we see that the depth inversely proportional to disparity, for a known baseline b and focal length f. The baseline and the focal length are the intrinsic parameters of the camera and are found during camera calibration. Furthermore,

$$x_r = x_l + d \qquad y_l = y_r \tag{1.8}$$

If we can now find corresponding points in the two images, i.e., $\mathbf{x}, \mathbf{x}'$, then we can find the disparity d and estimate the depth Z to the corresponding 3D scene point. If the correspondences are found at all points of the image then we can find the disparity at every position in the image, this is referred to as a *disparity map*. Using disparity map and the intrinsic camera parameters, we can perform the 3D reconstruction of the scene. However, problem that remains to be addressed still is: *how to find the stereo correspondences?*

1.2 Methods for stereo correspondence

The goal is now to find corresponding points in the left and right images, given that they are rectified. After rectification, one compares the similarity of pixels at candidate corresponding locations $\mathbf{x} = (x_l, y_l)$ and $(x_r, y_r) = (x_l + d, y_l)$, by varying the values of d. However, simply comparing image intensities at two locations is not enough as there may be textureless regions, repetitive patterns and two images may have different illuminations which may lead to ambiguous matches.

Some of the early methods (Hannah [1974], Marr and Poggio [1979],Pollard et al. [1985], Baker and Binford [1981]) matched only certain features in the images like edges or corners or contours, as they are more robust to illumination changes and produced matches with high certainty. This was also because of limitations in the computational resources at that point of time. However, due to wide range applications in image rendering and 3 modelling, most of the techniques today focus on finding a dense set of stereo correspondences.

The problem of finding dense correspondences has been studied intensely. Some methods suggest comparing the image intensities in a local neighbourhood around every point in the two images. These methods make an implicit assumption of smoothness in choosing a neighbourhood. The distinction in these methods is the measures used for comparison. These methods are usually referred to as *local* methods.

Global methods, on the other hand, define an energy function which involves a cost based on image intensities of the two images and regularizing term which enforces the disparities in the neighbourhood to be smooth. As the correspondence problem is ill-posed, an explicit regularization is required to obtain a physically plausible solution. This energy function is then minimized with respect to disparity to obtain the final disparity map.

Most of the recent techniques use energy minimization-based techniques, which is motivated by probabilistic modelling. The idea in these techniques is to use *Markov Random Field* (MRF) along with Bayesian inference to model the disparities. While Markov Random Fields specify the local interaction between the disparities, the Bayesian inference allows this interaction to be included as a prior distribution. This prior distribution encodes the smoothness of the disparities and the cost from the stereo image intensities is introduced as the likelihood. The Bayes' theorem allows to compute the posterior distribution using the prior and the likelihood. The advantage of using a Bayesian approach is that it provides a promising approach to such ill-posed problems because it treats the task at hand as an inference problem, finding the optimal estimate. In such Bayesian-MRF approaches the objective then is to maximize the posterior probability over all possible disparity maps. Such maximization/minimization requires optimization techniques such as simulated annealing, Mean Field, Belief propagation or Graph cuts.

Even though such modelling techniques capture the local interactions between the neighbouring disparities and incorporate the stereo image information, some crucial problems still remain:

- Some areas in the scene that are visible in one image may be *occluded* in the other and this can lead to incorrect matches.
- Regularization term in the model could smooth over all disparities and lead to poor solutions at the object boundaries.
- Incorporating just the stereo image intensities and smoothness term would model disparities which may not be consistent with the geometric properties of the surface.

In order to tackle these issues some extra information or constraints are required in modelling the correspondence problem. *Monocular cues* such as gradient, edges or colour information pertaining a single image could be used within the model to provide better solutions for disparity. In addition some extra geometric constraints have to be incorporated to obtain surface-consistent solutions for the disparity.

1.3 Contributions of the thesis

In this thesis, we focus on extending the constraints on the stereo correspondence problem based on monocular cues and extra geometric constraints. To this end we propose the following:

Cooperative disparity estimation and object boundary extraction

The first method proposes to cooperatively estimate disparities and object boundaries in a joint probabilistic framework. The idea here is to tackle the problem of localizing discontinuities in disparity which correspond to the object boundaries in the real world, along with that of disparity estimation. This scheme involves **incorporation of gradient information from a single image as monocular cue**. While the disparities are detected

using the stereo cue (the left and the right images), the monocular cues help in correcting the disparity at the discontinuities and finding the object boundaries. We model both the stereo and monocular cues within a joint MRF-framework. This part of our contributions can be found in the paper Narasimha et al. [2008].

Estimating surface consistent disparities

The second method incorporates **surface differential geometric constraints into the disparity model**. These constraints are derived from the surface normals in disparity space. The idea is to model the disparities in such a way that they lie on the plane defined by the surface normals. This constraint leads to solutions that are consistent with the surface geometric properties of the scene. The idea is to simultaneously estimate the disparity and surface normals, considering explicitly the influence of one on the other. This is done by modelling both the disparities and the normal in a joint framework. This work was originally published in Narasimha et al. [2009] and Narasimha et al. [2010].

Joint probabilistic modelling using coupled random fields

The major challenge in both of the above mentioned methods is to incorporate these cues and constraints with in a single joint probabilistic setting, in which the relationship between the disparities and the variables under consideration (object boundary or surface normals) can be explicitly established. In this regard we use the idea of *coupled Markov Random Fields*, which allows to model more than one random variable. This kind of modelling permits the influence of one variable on the other to be made explicit within the model. Such a probabilistic set up also allows for separate optimization techniques to be used for maximizing the posterior distributions pertaining to each of the variables, providing further flexibility in modelling and optimization. An *Alternating Maximization* procedure is then used to achieve overall optimization.

1.4 Outline of the thesis

We will now provide a brief description of the subsequent chapters. The chapters detailing the main contributions of the thesis are indicated by $\star$. The organization of the thesis is as follows:

In **chapter 2**, we provide a brief survey of the literature in stereo matching. In this chapter, we also provide a introduction to Markov Random Fields (MRF) and focus mainly techniques that use MRF models for stereo matching. We provide a brief summary of optimization techniques for MRFs used to estimate the disparity map. In addition, we discuss some of the ambiguities of the stereo matching and the existing methods to overcome them. We show that the use of monocular and geometric cues

are important for stereo matching problem. This discussion provides us with the motivation for proposed methods of the thesis.

In **chapter 3** we describe the idea of coupled Markov Random Field (coupled-MRF) models in the context of many applications such as image restoration, boundary estimation, image segmentation, texture segmentation, as well as in stereo matching. We also present an optimization strategy called Alternation Maximization. These models and methods provide the platform on which the algorithms proposed in this thesis are built.

⋆ In **chapter 4**, we present a method for incorporating monocular cues in order to co-operatively estimate object boundaries and disparities within a single joint probabilistic framework. We use coupled-MRFs to model the stereo disparities and the object boundary information. We show that such a model allows mutual improvement in the estimation of both disparity and boundary estimates. We use the Alternation Maximization that provides an efficient method for estimating both disparities and boundaries. Finally, the results obtained using the proposed method are presented and discussed.

⋆ In **chapter 5**, we propose a method to obtain surface consistent solutions for disparity. We do so by extending the constraints on the stereo matching to incorporate geometric contextual information about the surface properties. We present the significance of such constraints and show that incorporation of these constraints involves the estimation surface normals. We show the relationship between the surface normals in the depth space and disparity space. We then propose a coupled-MRF model to simultaneously estimate disparities and surface normals in the disparity space. The alternation maximization procedure used to estimate the two variables is discussed and results are presented.

In **chapter 6**, we summarize the methods proposed in the thesis and highlight some of the significant aspects pertaining to them. We conclude the thesis by providing some future directions for research.

Chapter 2

State of the art : Stereo Matching

As discussed in the previous chapter, one of the most difficult problems in stereo is that of finding correspondences between points in the left and right images. This problem is often referred to as the problem of *stereo matching*. The difference or shift in the position of the matches between two images is referred to as *disparity* and when described over an entire image is referred to as a *disparity map*. In order to reduce the search space of the disparities, most methods make use of the epipolar constraint. In the previous chapter, we demonstrated that this constraint reduces the search space to be along the scanline after rectification. Furthermore, we showed how the disparity is inversely proportional to the depth.

In this chapter we will discuss the papers corresponding to stereo matching, understood as matching along the scanline taking account the epipolar constraint. The stereo matching literature can broadly divided into two parts: *dense stereo matching* and *sparse stereo matching*. Sparse stereo matching involves finding matches between image features such as edges, lines or contours, leading to a disparity for each feature. The dense stereo matching techniques, on the other hand, try to find disparities at every position in the image. As in this thesis we are interested only in dense disparities, we will focus on the literature survey of only such techniques. These methods use image intensity information and may or may not use a regularization to eventually find the disparities at every pixel of the image. Already, the research done in this topic itself is vast and cannot be covered completely in one review. Therefore, we provide a survey of some important and relevant papers in this chapter.

In the next section we will discuss briefly some of the early stereo algorithms and their limitations. We then discuss in some detail how the stereo matching problem can be formulated as an energy minimization problem (section 2.2.1). In section 2.2.2 we show the relation of this energy to a Markov Random Field using Bayesian statistics. Optimization of this energy is discussed in section 2.3, throwing light on some important techniques such

as Mean Field, Belief Propagation and Graph Cuts. We then discuss the importance of additional cues in stereo in section 2.4. In particular, we discuss papers that deal with three aspects of stereo matching, namely, occlusions, disparity discontinuities and fronto-parallel assumptions. Finally we provide the motivation for the methods proposed in this thesis in section 2.5.

2.1 Early stereo algorithms

The traditional methods used for dense disparity extraction are the area-based approaches. The basic idea in such methods is to match small patches in the two images, under the assumption that the disparity within a patch is almost constant. *Matching costs* such as the sum of absolute differences (SAD) (Kanade [1994]), the sum of squared differences (SSD) (Matthies et al. [1989]) and the normalized cross correlation (NCC) (Ryan et al. [1980]) are used to measure the compatibility of the left and right images with a candidate shift at every pixel. This cost describes the similarity between the patches in the two images in terms of image intensities. A constant offset (bias) of pixel intensity values is compensated by the zero-mean versions ZSAD, ZSSD, and ZNCC. For each pixel position, the final disparity in such approaches is calculated by finding the minimum [1] of this cost. As the search for the disparity is only along a scanline, the minimum of this cost describes where along this line the most similar patch occurs (See figure 2.1, page 18).

Mathematically, the area-based cost C is defined in a 3D disparity space, also referred to as *disparity space image* (DSI), having the width $x \in [0, p]$, height $y \in [0, q]$ (where p and q are the width and the height of the image) and disparity $d \in [d_{min}, d_{max}]$:

$$\begin{aligned} C : [0,p] \times [0,q] \times [d_{min}, d_{max}] &\longrightarrow \mathbb{R} \\ (x, y, d) &\longmapsto C_{x,y}(d). \end{aligned} \tag{2.1}$$

This function $C_{x,y}(d)$ assigns a cost for every disparity $d \in [d_{min}, d_{max}]$ associated with pixel position (x, y). Assuming that the images are rectified, each element of cost $C_{x,y}(d) \in C$ maps the pixel (x, y) of left image to pixel $(x + d, y)$ in the right image. The disparity attributed to every point (x, y) is then:

$$d(x, y) = \underset{d \in [d_{min}, d_{max}]}{\arg\min} \; C_{x,y}(d). \tag{2.2}$$

For a right and left image pair $\mathbf{I}_R, \mathbf{I}_L$, if $C_{x,y}(d)$ is defined as a SSD, the cost that is to be minimized corresponds to the following:

$$C_{x,y}(d) = \sum_{m=-w/2}^{m=w/2} \sum_{n=-h/2}^{n=h/2} \big(\mathbf{I}_L(x+m, y+n) - \mathbf{I}_R(x+d+m, y+n)\big)^2 \tag{2.3}$$

1. maximum in case of correlation

where the image patch, considered in the two images, is a 2D window of size $w \times h$. However, this cost is very sensitive to illumination differences and thus the zero mean version of the cost (ZSSD) is generally used, which can be written as:

$$C_{x,y}(d) = \sum_{m,n} \left(\left(\mathbf{I}_L(x+m, y+n) - \bar{I}_L(x,y)\right) - \left(\mathbf{I}_R(x+d+m, y+n) - \bar{I}_R(x,y)\right) \right)^2 \quad (2.4)$$

where $\bar{I}_L(x,y)$ and $\bar{I}_R(x,y)$ are the mean intensities within the window $w \times h$. If the we use the absolute difference in place of squared difference in the two equations above (2.3 and 2.4), they would then correspond to SAD and ZSAD respectively.

The normalized cross correlation (NCC) is defined as the product of the two intensity vectors normalized over the 2D window:

$$C_{x,y}(d) = \frac{\sum_{m,n} \left(\mathbf{I}_L(x+m, y+n)\mathbf{I}_R(x+d+m, y+n)\right)}{\sqrt{\sum_{m,n} \left(\mathbf{I}_L(x+m, y+n)\right)^2} \sqrt{\sum_{m,n} \left(\mathbf{I}_R(x+m, y+n)\right)^2}}. \quad (2.5)$$

The zero-mean normalized version (ZNCC) is usually preferred as it is not sensitive to the gain or to the offsets of the camera:

$$C_{x,y}(d) = \frac{\sum_{m,n} \left(\mathbf{I}_L(x+m, y+n) - \bar{I}_L(x,y)\right)\left(\mathbf{I}_R(x+d+m, y+n) - \bar{I}_R(x,y)\right)}{\sqrt{\sum_{m,n} \left(\mathbf{I}_L(x+m, y+n) - \bar{I}_L(x,y)\right)^2} \sqrt{\sum_{m,n} \left(\mathbf{I}_R(x+m, y+n) - \bar{I}_R(x,y)\right)^2}}. \quad (2.6)$$

All of the above methods assume constant disparity over the patches. This essentially enforces a fronto-parallel constraint on the disparity. Also, the size of the patch determines the smoothness of the output disparity. As can be seen in Figure 2.2, while a small window provides a noisy output, a very large window might over-smooth the output. Therefore the window size is an important parameter to be tuned. As the uniqueness of matches is enforced only in one image (reference image), points in the other image may get matched to multiple points in the reference image. Another limitation, which is widely addressed in area-based (and other) literature, is that of localizing disparity discontinuities, which such methods are incapable of handling. Variants of area-based methods have been suggested, such as using adaptive windows (Kanade and Okutomi [1994]) and shiftable windows (Bobick and Intille [1999]), to deal with this problem. Recently, Yoon and Kweon [2007] presented a method where the weight of the pixel within a given window is adjusted based on the colour similarity and spatial distance from the centre. However, such methods still retain the fronto-parallel constraint. Such techniques are also referred to as *local methods*, as only the local neighbourhood of the image is taken into consideration.

Among other local methods is the one suggested by Pollard et al. [1985]. Their paper described an important concept called the disparity gradient limit. The disparity gradient (DG) between two nearby points in a stereo image-pair is defined as the difference in their disparities divided by their separation in visual angle. Pollard et al. based their idea on the

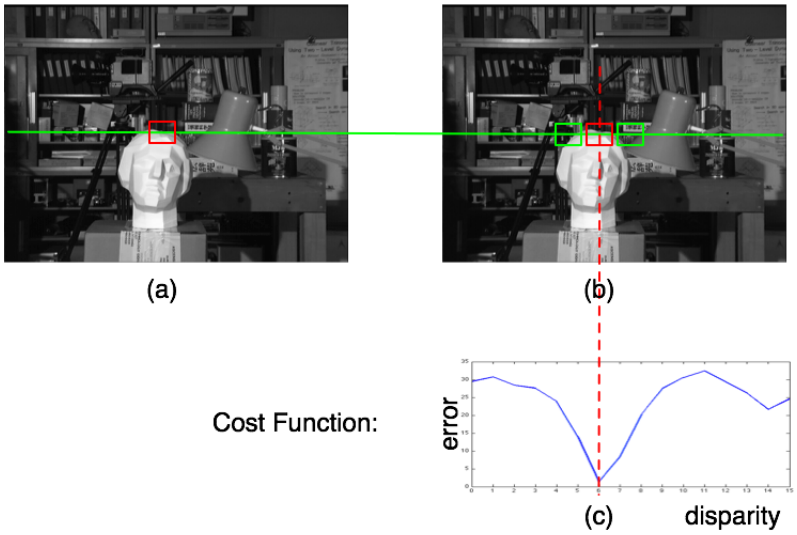


Figure 2.1: Area-Based Approach: Match pixel windows in one image to the other. (a) Shows the left image with image patch as 2D window (shown in red). (b) Shows the shifted windows (shown in green) where the window with the minimum cost is shown in red. (c) shows the cost function and how the minimum of the function is used to find the disparity.

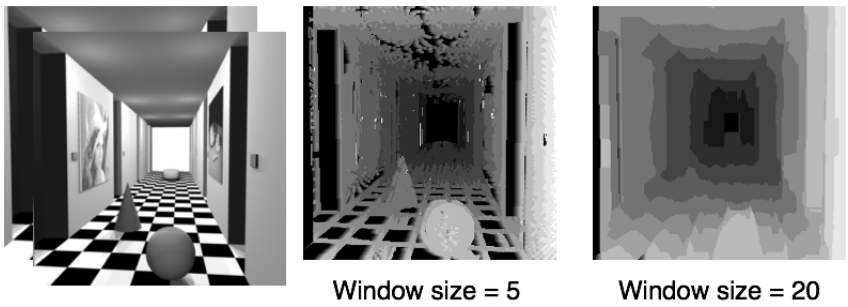


Figure 2.2: SSD approach: Window Size effects the smoothness or otherwise of the disparity map. While window size of 5×5 leads to noisy disparity map, a larger window size (of 20×20) leads to over-smoothing of disparities.

psychophysical studies (by Burt and Julesz [1980]) demonstrating that the fusional capacity of human vision breaks down when this gradient exceeds a critical value (approximately 1), namely the DG limit. They, therefore, formulated the matching cost by assigning a finite cost to all neighbours satisfying the DG limit and infinite to those beyond the limit. Similar, studies on the use of disparity gradient were independently made by Prazdny [1985] at the same time.

2.2 Energy minimization-based algorithms

Unlike the local methods, the *global* ones (Barnard [1989], Scharstein and Szeliski [1998], Boykov et al. [2001]) explicitly model the problem of stereo matching as an energy function and solve it as an optimization problem. In this framework, the estimated disparity map $\mathbf{d}$ is the minimum of a global energy. The energy E_{total}, also referred to as objective function has two terms: i) the **data term** E_{data} that measures if the input images (left and right) agree with the disparity and ii) the **smoothness term** E_{smooth} that encodes the interaction between the neighbouring disparities. Mathematically, we first define a finite set

$$\mathcal{S} = \{\mathbf{x} \mid \mathbf{x} = (x, y),\ x \in [0, p],\ y \in [0, q]\}, \tag{2.7}$$

where $p \times q$ is the size of the image [2], and $\mathbf{x}$ corresponds to the pixel positions. This set $\mathcal{S}$ has one-to-one correspondence with the image grid positions. A candidate disparity map can then be defined as

$$\mathbf{d} = \{d_{\mathbf{x}} \mid d_{\mathbf{x}} \in [d_{min}, d_{max}] \ \&\ \mathbf{x} = \in \mathcal{S}\}. \tag{2.8}$$

The above equation signifies that every position $\mathbf{x}$ in the grid is associated with a disparity $d_{\mathbf{x}}$ which takes its values from the set/interval $[d_{min}, d_{max}]$. The energy for this disparity map $\mathbf{d}$ can be written as,

$$E_{\text{total}}(\mathbf{d}) = E_{\text{data}}(\mathbf{d}) + \lambda\, E_{\text{smooth}}(\mathbf{d}) \tag{2.9}$$

where λ weights the influence of the two terms on the total energy. The minimization of this energy E_{total} with respect to all possible disparity maps $\mathcal{D} = [d_{min}, d_{max}]^{p \times q}$ gives the final disparity map $\mathbf{d}^*$:

$$\mathbf{d}^* = \arg\min_{\mathbf{d} \in \mathcal{D}} E_{\text{total}}(\mathbf{d}) \tag{2.10}$$

2.2.1 Energy function formulation

The data term ($E_{\text{data}}(\mathbf{d})$) is usually defined as a function of image intensities, at a given disparity:

$$E_{\text{data}}(\mathbf{d}) = \sum_{\mathbf{x}} V_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I}) \tag{2.11}$$

where $d_{\mathbf{x}}$ is the disparity at the position $\mathbf{x}$ in the image and the term $V_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I})$ penalizes the disparities which do not agree with the input image pair $\mathbf{I} = (\mathbf{I}_L, \mathbf{I}_R)$ (left and right images) at that position. Though often only a pixel-wise difference between the intensities two images is considered as a penalty, windowed differences as described in section 2.1 are also used. As in case of windowed differences, the function $V_{\mathbf{x}}$ resides in DSI (2.1, page 16), providing a cost of choosing for $d_{\mathbf{x}}$ at the position $\mathbf{x}$ as the disparity based on the input image

2. We assume the right and left images are of the same size

pair $\mathbf{I}$. However, the measures in section 2.1 do not take into consideration the sampling of the image. A measure quite commonly used to overcome this limitation is the one presented by Birchfield and Tomasi [1999a]. Here, instead of comparing pixel values shifted by integral amounts, they compare each pixel in the reference image to a linearly interpolated pixel in the other image. This measure was first introduced in context of dynamic programming and has been since used as a data cost in other energy minimization based approaches (For example, Sun et al. [2003], Yang et al. [2009]).

The second term in (2.9) ensures that the neighbouring disparities have coherent values and is therefore referred to as the smoothness term ($E_{\text{smooth}}(\mathbf{d})$). This smoothness term is similar to the continuity constraint suggested by Marr and Poggio [1976]. In order to formalize this term, we first model the interaction between the disparities at different positions. The *interaction* is usually defined as a function of differences between the disparities in nearby positions. This set of nearby positions is referred to as a *neighbourhood* ($\mathbf{N_x}$) of the position $\mathbf{x}$. Each pixel in $\mathbf{N_x}$ is called a neighbour of $\mathbf{x}$. The term $E_{\text{smooth}}(\mathbf{d})$ can be now written as:

$$E_{\text{smooth}}(\mathbf{d}) = \sum_{\mathbf{x}} \sum_{\mathbf{y} \in \mathbf{N_x}} V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}}). \tag{2.12}$$

$V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}})$ is normally referred to as an *interaction function*. The form of this function is an important aspect of the energy formulation and will be discussed in more detail in later sections. Another thing to notice is that equation (2.12) considers only pairwise interactions. This means that the model depends only on first order differences of disparities, thus favouring fronto-parallel disparities. This model has been widely used for its simplicity and because of the development of powerful optimization techniques such as Mean Field Approximation (Strecha et al. [2006], Yuille et al. [1990]), Belief Propagation (Yang et al. [2009], Xu and Jia [2008], Felzenszwalb and Huttenlocher [2006], Klaus et al. [2006], Sun et al. [2005, 2003]) and Graph Cuts (Boykov et al. [2001], Kolmogorov and Zabih [2001]).

The definition of the interaction function $V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}})$ is very important in determining the smoothness of the final disparity map. $V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}})$ could be chosen (Poggio et al. [1985]) as a monotonically increasing function such as:

$$V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}}) = |d_{\mathbf{x}} - d_{\mathbf{y}}|^{\alpha}, \quad \alpha = 1 \, or \, 2 \tag{2.13}$$

Such a function penalizes very large disparity differences through very large costs. This makes the final disparity map $\mathbf{d}$ very smooth and results in poor localization of disparity discontinuities. To improve the localization of discontinuities Gamble and Poggio [1987] weight the smoothness prior with intensity differences. This idea encourages the disparity discontinuities to coincide with intensity/colour edges.

Alternatively, the interaction function can be modelled as a robust function. The derivative of a robust function goes to zero for large $|d_{\mathbf{x}} - d_{\mathbf{y}}|$, for example, the Potts or the

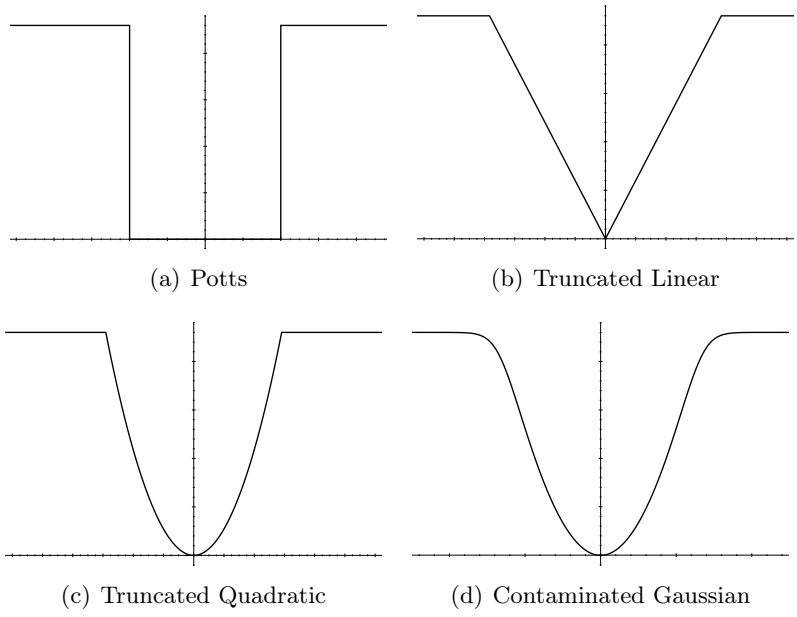
truncated-linear function (figure 2.3(a) and 2.3(b)) :

$$V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}}) = \begin{cases} 0 & \text{if } |d_{\mathbf{x}} - d_{\mathbf{y}}| < T \\ c & \text{otherwise} \end{cases} \qquad \text{(Potts)}$$

$$\text{or} \qquad (2.14)$$

$$V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}}) = \begin{cases} |d_{\mathbf{x}} - d_{\mathbf{y}}| & \text{if } |d_{\mathbf{x}} - d_{\mathbf{y}}| < T \\ c & \text{otherwise} \end{cases} \qquad \text{(Truncated linear)}$$

This function assigns a small cost if the difference disparity between the neighbours is small, and a constant penalty c for large differences. Parameter T determines the disparity differences beyond which a constant penalty is to be applied. Hence, this interaction function allows for disparity discontinues to occur, as the penalty for large disparity differences is limited. Some more examples of robust functions are shown in Figure 2.3(c) and 2.3(d). Further details on robust statistics can be found in the paper by Black and Rangarajan [1996]. As a large part of the literature uses this kind of interaction, we call this a *standard*



(a) Potts (b) Truncated Linear

(c) Truncated Quadratic (d) Contaminated Gaussian

Figure 2.3: Examples of robust functions for $V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}})$

model. We will discuss in the section 2.4.3 about papers using higher order derivative models for disparity.

We will show how the energy formulation in (2.9) can be given a Bayesian-MRF interpretation. We give a Bayesian statistics-based justification to the minimization of the selected energy function. We will show how the minimization of the energy function is equivalent to finding the *maximum a posteriori* estimate of a certain Markov Random Field (MRF).

2.2.2 Markov Random Fields and stereo

Geman and Geman [1984] were among the first to give a Bayesian interpretation to such energy functions based on the Markov Random Field (MRF) framework. We refer the readers to Li [2001] for more detailed introduction to MRFs. Like many problems in computer vision the disparity estimation task can also be formulated as a labelling problem. By a labelling problem, we mean assigning a label to every pixel position; in the present case the labels are the disparities.

2.2.2.1 Disparity estimation as Labelling Problem

As in described in (2.7), we consider the finite set $\mathcal{S}$ of $p \times q$ pixels on a regular 2D-grid, which are also referred to as *sites*. We define a neighbourhood system $\mathcal{N} = \{\mathbf{N_x} | \mathbf{x} \in \mathcal{S}\}$ on $\mathcal{S}$. The neighbourhood $\mathbf{N_x} \in \mathcal{N}$ satisfies two properties i) $\mathbf{x} \notin \mathbf{N_x}$ and ii) if $\mathbf{x} \in \mathbf{N_y}$ then $\mathbf{y} \in \mathbf{N_x}$. The pair $(\mathcal{S}, \mathcal{N}) = \mathcal{G}$ can be viewed as an undirected graph $\mathcal{G}$ where the nodes $\mathcal{S}$ are linked through a neighbourhood relationship $\mathcal{N}$. A *clique* $\mathcal{C}$ for $(\mathcal{S}, \mathcal{N})$ is defined as a subset of sites in $\mathcal{S}$, where each member of the set is a neighbour of all the other members. A more detailed study of such graphical models can be found in the book by Bishop [2007].

We denote by $\mathbf{D} = \{D_\mathbf{x}, \mathbf{x} \in \mathcal{S}\}$ the unknown disparity values at each pixel $\mathbf{x} = (x, y)$. The $D_\mathbf{x}$'s are considered as random variables that take their values in a finite discrete set of L disparity labels denoted by $\mathcal{L} = \{d_1, d_2, \ldots d_k, \ldots, d_L\}$. The disparity labels in $\mathcal{L}$ correspond to discrete set of values ranging from d_{min} to d_{max}. The *disparity field* $\mathbf{D}$, also referred to as *configuration*, takes its values in $\mathcal{D} = \mathcal{L}^{p \times q}$. We use small letter $\mathbf{d}$ to denote a specific realization of the random field $\mathbf{D}$. This realization $\mathbf{d}$ corresponds a disparity map and is similar to the one described in (2.8). The joint probability that a random variable $D_\mathbf{x}$ takes the value $d_k \in \mathcal{L}$ is denoted as $P(D_\mathbf{x} = d_k)$ and abbreviated as $P(d_k)$. Similarly, the joint probability $P(\mathbf{D} = \mathbf{d})$ is abbreviated by $P(\mathbf{d})$. $\mathbf{D}$ is said to be a Markov Random Field on $\mathcal{S}$ with respect to a neighbourhood system $\mathcal{N}$ if, and only if, the following two conditions are satisfied:

i) $P(\mathbf{d}) > 0, \quad \forall\, \mathbf{d} \in \mathcal{D}$

ii) $P(d_\mathbf{x} | d_{\mathcal{S}-\mathbf{x}}) = P(d_\mathbf{x} | d_{\mathbf{N_x}})$

where $\mathcal{S} - \mathbf{x}$ is the set difference and $d_{\mathbf{N_x}}$ denotes all labels of sites in $\mathbf{N_x}$. The first property is called positivity and the second one is referred to as the Markovian property. The Markovian property states that only neighbouring labels have direct interactions with each other. One of the most important theoretical results, the *Hammersley and Clifford theorem* (Besag [1974]), provides a mathematically tractable means of specifying the joint probability of an MRF, through a Gibbs Random Field (GRF). This theorem states the equivalence between MRFs and GRFs as: $\mathbf{D}$ *is an MRF on* $\mathcal{S}$ *w.r.t.* $\mathcal{N}$ *if and only if* $\mathbf{D}$ *is a GRF on* $\mathcal{S}$ *w.r.t.* $\mathcal{N}$. A GRF describes the random variables in terms of Gibbs distribution,

which takes the following form:

$$P(\mathbf{d}) = \frac{1}{Z} \exp\left(-\sum_{c \in \mathcal{C}} V_c(\mathbf{d}) \right) \tag{2.15}$$

where $V_c(\mathbf{d})$ is the clique potential defined over all the cliques $c \in \mathcal{C}$. Therefore, Hammersley and Clifford theorem establishes an equivalence between a MRF which is characterized by its local property (the Markovian property) and a GRF which is characterized by its global property (the Gibbs distribution). Considering only pairwise potentials and defining $V_c(\mathbf{d}) = V_{\mathbf{x},\mathbf{y}}(d_\mathbf{x}, d_\mathbf{y})$ (where $\mathbf{x}$ and $\mathbf{y}$ are neighbours) as in (2.12), defines a joint distribution for the disparity-MRF:

$$P(\mathbf{d}) = \frac{1}{Z} \exp\left(-\sum_{\mathbf{x}} \sum_{\mathbf{y} \in \mathbf{N_x}} V_{\mathbf{x},\mathbf{y}}(d_\mathbf{x}, d_\mathbf{y}) \right) \tag{2.16}$$

where Z is the normalizing factor.

2.2.2.2 MAP-MRF estimation

As the disparity field $\mathbf{D}$ is not directly observable, its realization $\mathbf{d}$ has to be estimated based on observations. For this reason $\mathbf{D}$ is also referred to as *latent* or *hidden* variable. The observed data, in our case the left and right images $\mathbf{I}_L$ and $\mathbf{I}_R$, are together referred to as $\mathbf{I}$. In other words, the disparity value at each location can be seen as a hidden variable and the actual data from the two images can be seen as observations and thus our MRF model can be represented graphically as in Figure 2.4. The most popular way of estimating

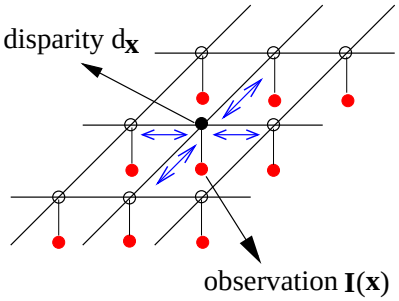


Figure 2.4: Disparity MRF: the observed data $\mathbf{I}(\mathbf{x})$ refers to the left and right image data and $d_\mathbf{x} \in \mathbf{d}$ refers to the disparity at the pixel position $\mathbf{x} = (x, y)$

the values of disparity-MRF is through *maximum a posterior* (MAP) estimation. The goal of the MAP-MRF is to estimate the realization $\mathbf{d}^{\text{MAP}}$ given the data $\mathbf{I}$. That is :

$$\mathbf{d}^{\text{MAP}} = \arg\max_{\mathbf{d} \in \mathcal{D}} P(\mathbf{d}|\mathbf{I}) \tag{2.17}$$

Using Bayes' rule the posterior can be written as:

$$P(\mathbf{d}|\mathbf{I}) = \frac{P(\mathbf{I}|\mathbf{d})\,P(\mathbf{d})}{P(\mathbf{I})} \tag{2.18}$$

Because $P(\mathbf{I})$ is a constant for fixed $\mathbf{I}$ we can write the above equation as

$$P(\mathbf{d}|\mathbf{I}) \propto P(\mathbf{I}|\mathbf{d})\,P(\mathbf{d}) \tag{2.19}$$

The *likelihood function* $P(\mathbf{I}|\mathbf{d})$ expresses how probable the observed data is for different settings of the disparity. By making an assumption that $P(\mathbf{I}|\mathbf{d})$ is conditionally independent at every pixel position $\mathbf{x}$ with respect to $\mathbf{d}$, we can write:

$$P(\mathbf{I}|\mathbf{d}) = \frac{1}{C}\exp\left(-\sum_{\mathbf{x}} V_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I})\right) \tag{2.20}$$

where C is the normalization constant. The $P(\mathbf{d})$ is referred to as *prior* probability and is defined as MRF (in the previous section). We can now write the posterior probability using the equations (2.20) and (2.16) as

$$P(\mathbf{d}|\mathbf{I}) \propto \exp\left(-\sum_{\mathbf{x}} V_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I}) - \lambda \sum_{\mathbf{x}} \sum_{\mathbf{y}\in\mathbf{N}_{\mathbf{x}}} V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}})\right) \tag{2.21}$$

where λ is weighting term between the likelihood and prior. Maximizing the above function is equivalent to minimizing:

$$E(\mathbf{d}) = \sum_{\mathbf{x}} V_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I}) + \sum_{\mathbf{x}} \sum_{\mathbf{y}\in\mathbf{N}_{\mathbf{x}}} V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}}) \tag{2.22}$$

We now note the equivalence of the above equation with the energy formulation in (2.9), where the functions $V_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I})$ and $V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}})$ can be defined in the same way as in (2.11) and (2.12), respectively. As $V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}})$ determines the smoothness of the disparity map and is part of the prior (2.16), it also referred to as *smoothness prior*.

The discussion until now has shown how we can formulate the energy function (section 2.2.1) and how it can interpreted using Bayesian-MRF framework. An important point to be noted is that the forms of the interaction function (2.12 or 2.16) and the data function (2.11 or 2.20) are very difficult to specify. This is due to inherent ambiguities of the stereo matching problem itself, including the depth discontinuities, occlusions, image noise as well as the complexity of the scene itself. Even if the forms of these functions are given, we still have the problem of optimizing the stereo model to find the MAP-MRF estimates. The problem of optimization is difficult because of the vast solution space $\mathcal{D} = \mathcal{L}^{p\times q}$. It is, therefore, necessary to make approximations both on the model and the optimization algorithm.

2.3 Optimization

The goal of optimization is to minimize the energy (2.22), to obtain the MAP-MRF estimates for the disparities. Geman and Geman [1984] suggested the use of Simulated Annealing (Kirkpatrick et al. [1983]) for optimizing such energy functions. While Simulated Annealing can be proven in theory to achieve the global minimum, it is prohibitively slow in practice, especially for stereo matching. In order to overcome this limitation Besag [1986] introduced a greedy method called the Iterated Conditional Modes. Though this method is computationally efficient, it is very sensitive to the initialization and thus ineffective in finding good disparities.

Over the last few years approximate inference techniques such as Mean Field (Strecha et al. [2006]), Belief Propagation (Yedidia et al. [2003], Sun et al. [2003]) and Graph Cuts (Boykov et al. [2001], Kolmogorov and Zabih [2001]) have gathered a lot of attention in stereo matching. Such algorithms allow fast approximate solution of MRF-based problems. While they do not ensure global optimality, they give substantially more accurate results than it were previously possible. In the next few sections, we will discuss in some detail Mean Field, Belief Propagation and Graph cut algorithms as they are relevant to this thesis.

2.3.1 Mean Field Approximation

Mean Field theory has its roots in statistical physics and was intended for approximating the behaviour of interacting spin systems in thermal equilibrium (Chandler and Percus [1988]). It has been widely used in computer vision for image segmentation by Geiger and Yuille [1991], Forbes and Fort [2007], blind image separation by Tonazzini et al. [2006], motion analysis and tracking (Hua and Wu [2006], Medrano et al. [2009]), as well as stereo disparity estimation by Yuille et al. [1990], Strecha et al. [2006].

The basic idea behind Mean Field is to approximate the true posterior distribution $P(\mathbf{d}|\mathbf{I})$ of the MRF by a tractable distribution $Q(\mathbf{d})$. This is done by assuming that the approximate distribution $Q(\mathbf{d})$ fully factorises over all the sites $\mathbf{x}$ as follows:

$$Q(\mathbf{d}) = \prod_{\mathbf{x}\in\mathcal{S}} Q_{\mathbf{x}}(d_{\mathbf{x}}) \tag{2.23}$$

where $Q_{\mathbf{x}}(d_{\mathbf{x}})$ is a distribution over L possible disparity values of $d_{\mathbf{x}} \in \mathcal{L}$ at the site $\mathbf{x}$. Solution of the Mean Field problem is sought by variational methods. Variational methods allow the transformation of the probabilistic inference task (2.21) into an optimization problem. The solution to the variational problem is often given in terms of *fixed point equations*.

In the case of Mean Field, the variational formulation of the problem is equivalent to minimizing the Kullback-Leibler (KL) divergence (Yedidia et al. [2003], Wainwright and Jordan [2005], Jaakkola [2000]) between the two distributions $P(\mathbf{d}|\mathbf{I})$ and $Q(\mathbf{d})$:

$$\mathrm{KL}\big(Q(\mathbf{d})\big\|P(\mathbf{d}|\mathbf{I})\big) = \sum_{\mathbf{d}\in\mathcal{D}} Q(\mathbf{d}) \log \frac{Q(\mathbf{d})}{P(\mathbf{d}|\mathbf{I})} \tag{2.24}$$

The KL divergence is always positive and zero only if the *variational distribution* $Q(\mathbf{d})$ is equal to the true posterior distribution $P(\mathbf{d}|\mathbf{I})$. By substituting the equations (2.23) and (2.21) in the equation for KL divergence we get:

$$\begin{aligned}\mathrm{KL}\big(Q(\mathbf{d})\big\|P(\mathbf{d}|\mathbf{I})\big) = \; & - \sum_{\mathbf{x}} \sum_{d_{\mathbf{x}} \in \mathcal{L}} Q_{\mathbf{x}}(d_{\mathbf{x}}) V_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I}) \\ & - \sum_{\mathbf{x}} \sum_{\mathbf{y} \in \mathbf{N}_{\mathbf{x}}} \sum_{d_{\mathbf{x}}, d_{\mathbf{y}} \in \mathcal{L}} Q_{\mathbf{x}}(d_{\mathbf{x}}) Q_{\mathbf{y}}(d_{\mathbf{y}}) V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}}) \\ & + \sum_{\mathbf{x}} \sum_{d_{\mathbf{x}} \in \mathcal{L}} Q_{\mathbf{x}}(d_{\mathbf{x}}) \log Q_{\mathbf{x}}(d_{\mathbf{x}}) \end{aligned} \tag{2.25}$$

Finding the marginals $Q_{\mathbf{x}}(d_{\mathbf{x}})$ from the above equation is not straight-forward and requires iterative re-substitution. As the $Q(\mathbf{d})$ is fully factorized (2.23), the above equation can be optimized one marginal component at a time.

The update equations (also the fixed point equations) are derived by minimizing the KL divergence (2.25) with respect to each marginal $Q_{\mathbf{x}}(d_{\mathbf{x}})$ in $Q(\mathbf{d})$ while keeping the remaining marginals fixed. This is equivalent to taking the partial derivative $\frac{\partial \mathrm{KL}(\cdot)}{\partial Q_{\mathbf{x}}(d_{\mathbf{x}})}$ and setting it to zero. It can be easily verified that, the minimization of (2.25) with respect marginals $Q_{\mathbf{x}}(d_{\mathbf{x}})$ leads to the following:

$$Q_{\mathbf{x}}(d_{\mathbf{x}}) \leftarrow \frac{1}{Z} \exp\left(- \left(V_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I}) + \sum_{\mathbf{y} \in \mathbf{N}_{\mathbf{x}}} \sum_{d_{\mathbf{y}} \in \mathcal{L}} Q_{\mathbf{y}}(d_{\mathbf{y}}) V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}}) \right) \right) \tag{2.26}$$

These update equations (collectively for all $\mathbf{x}$) are also referred to as *Mean Field equations*. The normalization Z ensures that $\sum_{d_{\mathbf{x}} \in \mathcal{L}} Q_{\mathbf{x}}(d_{\mathbf{x}}) = 1$. Successive application of the updates correspond to iteratively enforcing different Mean Field equations. These equations indicate that the updates are made "locally" by averaging the neighbours with respect to the other component distributions. In other words, each $\mathbf{x}$ only sees the "mean effect" of the neighbours $\mathbf{N}_{\mathbf{x}}$.

Yuille et al. [1990] were among the first to use Mean Field Approximation in the context of stereo. They formulated their energy function on psycho-physical grounds, incorporating discontinuities and the matching of different primitives instead of just intensities. However, as mean-field approximates the true distribution with a much simpler form with marginals over a single variable, it does not provide a good approximation of the true posterior. More complicated forms for $Q(\cdot)$ can be used instead of the fully factorised version in (2.23) (Wainwright and Jordan [2005] and Jaakkola [2000]). Nevertheless, Strecha et al. [2006] used this approximation within an Expectation-Maximization framework, estimating both disparities and occlusions, and showed results comparable to the state-of-the-art. Mean Field approximation can also be interpreted as parallel message-passing algorithm, in the spirit of message passing algorithms such as Belief Propagation. Here, each site $\mathbf{x}$ sends a message $Q_{\mathbf{x}}(d_{\mathbf{x}})$ to its neighbours, which is in turn based on the message it received from its neighbours in the previous time step.

2.3.2 Belief Propagation

Belief Propagation (BP) was first introduced by Pearl [1986], and is best understood in the context of probabilistic graphical models (See Section 2.2.2.1). BP is part of the most popular class of algorithms called the *message passing* algorithms, where information or "message" is passed from one site to another along the edges of a graph until convergence. The BP is exact on trees (graphs with no cycles), but in graphs with cycles, like the MRF shown in figure 2.4, there are no guarantees on convergence.Surprisingly, despite these issues, BP provides good approximate results on such cyclic graphs. In particular good results are obtained when BP is applied to stereo matching, where the graph $\mathcal{G}$ (See section 2.2.2.1) with cycles represents the MRF model. When the BP is applied to graphs with cycles or *loops*, it is often referred to as Loopy Belief Propagation (LBP) [3].

In order to explain the message passing in LBP algorithm we define the following notations: $b_{\mathbf{x}}(d_{\mathbf{x}})$ is the belief at the site $\mathbf{x}$, and is equivalent to the approximate distribution $Q_{\mathbf{x}}(d_{\mathbf{x}})$ in the previous section $m_{\mathbf{x},\mathbf{y}}(d_{\mathbf{y}})$ is the message sent by site $\mathbf{x}$ to $\mathbf{y}$ as to what state the variable $d_{\mathbf{y}}$ should be in. Note that both $b_{\mathbf{x}}(d_{\mathbf{x}})$ and $m_{\mathbf{x},\mathbf{y}}(d_{\mathbf{y}})$ are vectors of length L corresponding to the number of states in the set $\mathcal{L} = d_1, d_2, \dots d_k, \dots, d_L$. To improve readability, we re-write the posterior distribution in (2.21) as follows:

$$P(\mathbf{d}|\mathbf{I}) \propto \prod_{\mathbf{x}} \psi_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I}) \prod_{\mathbf{x}} \prod_{\mathbf{y} \in \mathbf{N_x}} \phi_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}}) \tag{2.27}$$

where $\psi_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I}) = \exp\big(-V_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I})\big)$ and $\phi_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}}) = \exp\big(-V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}})\big)$. The belief at every site $\mathbf{x}$ is proportional to the local evidence at that position, in our case the data or the likelihood term $\psi_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I})$ and to all the messages coming in to the $\mathbf{x}$:

$$b_{\mathbf{x}}(d_{\mathbf{x}}) = k\, \psi_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I}) \prod_{\mathbf{y} \in \mathbf{N_x}} m_{\mathbf{y},\mathbf{x}}(d_{\mathbf{x}}) \tag{2.28}$$

where k is the normalizing constant ensuring that $\sum_{d_{\mathbf{x}} \in \mathcal{L}} b_{\mathbf{x}}(d_{\mathbf{x}}) = 1$. The messages are updated iteratively as follows:

$$m_{\mathbf{x},\mathbf{y}}(d_{\mathbf{y}}) \leftarrow \max_{d_{\mathbf{x}} \in \mathcal{L}} \psi_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I})\, \phi_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}}) \prod_{\mathbf{z} \in \mathbf{N_x} \setminus \mathbf{y}} m_{\mathbf{z},\mathbf{x}}(\mathbf{x}) \tag{2.29}$$

The right-hand-side of the above equation takes into account the data term, the interaction term and all the messages that are coming into $\mathbf{x}$, except the one from $\mathbf{y}$. The updates for the two equations are diagrammatically shown in the the figure 2.5. These equations represent what is called as the *max-product* BP. If the max-operation in (2.29) is replaced by a sum it becomes *sum-product* BP. While, the max-product BP algorithm finds the MAP estimate or minimum energy associated with the MRF, the sum-product algorithm can be used to approximate the posterior probability $p(\mathbf{d}|\mathbf{I})$ of each label (in our case disparity) for each pixel.

3. Note that in this thesis we use LBP and BP interchangeably.

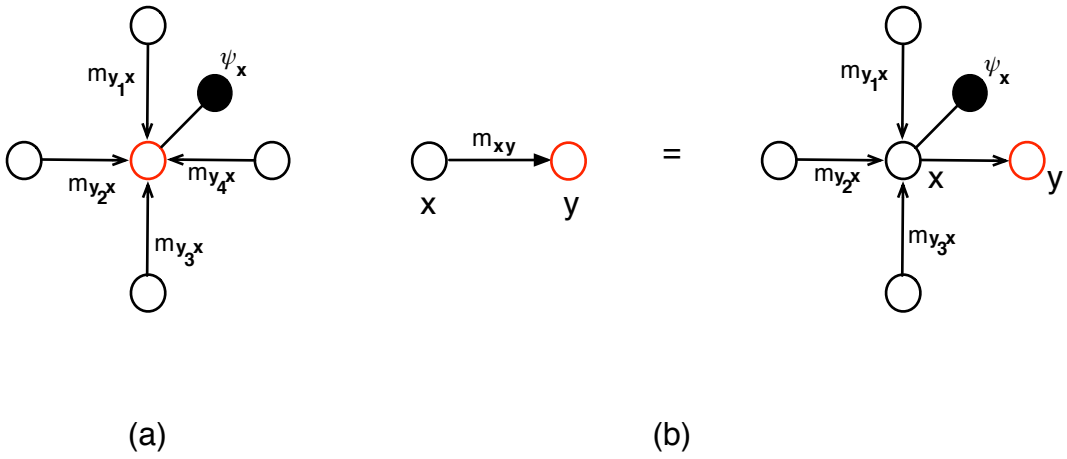


Figure 2.5: Update equations for Belief Propagation (a) shows diagrammatically the equation (2.28) where belief at $\mathbf{x}$ (shown in red) is calculated taking into account the messages from the neighbours ($m_{\mathbf{yx}}$) and the data $\psi_{\mathbf{x}}$ and (b) shows the equation (2.29) the message $m_{\mathbf{xy}}$ passed from $\mathbf{x}$ to $\mathbf{y}$ (shown in red). $m_{\mathbf{xy}}$ includes messages passed from the neighbours in $\mathbf{N_x}$ to $\mathbf{x}$ and the evidence $\psi_{\mathbf{x}}$.

The equations (2.28) and (2.29) merely provide the update equations for BP. The principle of BP, however, can be explained using variational methods. Yedidia et al. [2003] showed that fixed points of the BP (update equations) actually, correspond to the minimum of the *Bethe approximation*, and to the global minimum if there are no loops. Unlike the Mean Field approximation, where only the single site distributions are considered (2.23), the Bethe approximation considers both pair-wise and single site distributions. A more detailed and in-depth comparison of the Mean Field and Belief Propagation is provided in the paper by Weiss [2001]. He found that despite its non-convergence, BP finds better local minima than Mean Field approaches.

In the context of stereo matching, LBP was first used by Sun et al. [2003] to compute disparities. However, if used directly this method is still computationally demanding. Felzenszwalb and Huttenlocher [2006], therefore, provide methods for speeding up the LBP approach. In particular, Felzenszwalb and Huttenlocher suggested the use of a multi-grid approach where the computations are performed in a coarse to fine manner, thus providing substantial speed up in stereo matching application. A large number of variants of BP have also been developed over the years, among which the sequential-tree re-weighted BP (sequential-TRW) by Kolmogorov [2006] has been noticeable. This is because sequential-TRW achieves lower bound on energy, thus ensuring convergence. A comparison of sum-product, max-product and sequential-TRW in context of stereo and other vision applications such as photomontage and binary image segmentation is done by Szeliski et al. [2008]. This

comparative study showed that the sequential-TRW consistently provided better results than both forms of LPB.

2.3.3 Graph Cuts

Graph Cuts from combinatorial optimization can be used to minimize the MRF energy function (2.22). Graph Cuts include the max-flow/min-cut algorithms in combinatorial optimization. These techniques were first used in computer vision by Greig et al. [1989] in the context of binary image restoration. Greig et al. showed that global minimum of the energies of the form (2.22) could be achieved when the number of labels is limited to two (binary labelling). Graph Cuts, while widely used in computer vision applications, are in general limited to only binary labelling problems.

In order to apply Graph Cuts to multi-label problems (such as stereo correspondence), Ishikawa [2003] transformed the graph to an equivalent binary problem. However, in their framework the interaction functions were constrained to be convex. The problem with having convex interaction function is that the resulting method performs poorly at the disparity discontinuities. Alternatively, Boykov et al. [2001] proposed to apply the Graph Cuts repeatedly on pairs of labelling within their inner-loops. They suggested two graph-cut algorithms, namely *expansion move* and *swap move* algorithms. For a label $\alpha \in \mathcal{L}$, expansion move allows any set of sites to change their labels to α. The local minimum is found, if no expansion move for any label α yields a lower energy. The swap move algorithm interchanges the labels of some subset of sites labelled α to β and vice versa. Similar to expansion move the swap move finds the local minimum such that no other swap move will produce a lower energy. These two algorithms produce very stable minima, and allow the interaction function to be metric in the case of expansion move and semi-metric in case of swap move algorithms. Kolmogorov and Zabih [2002b] showed further classes of interaction functions that could be considered within the Graph Cuts framework.

The expansion and swap move algorithms have been applied to stereo correspondence problem by Boykov et al. [2001], Kolmogorov and Zabih [2001] and Kolmogorov and Zabih [2002a] with some success. The comparative study done by Tappen and Freeman [2003] of BP and Graph Cuts showed that the two algorithms produced comparable results under identical parameter settings. Tappen and Freeman, however, used the simple Potts model for comparing the two algorithms. The extended comparison of BP, Graph Cuts and sequential-TRW by Szeliski et al. [2008] showed that sequential-TRW and expansion move algorithms performed the best for stereo matching. While they compared the algorithms based on the energy, they found that some energies calculated were lower than that of the ground-truth. This illustrates the need for more accurate modelling of the problem itself. The figure 2.6 shows the output disparity map obtained by using Mean Field, standard-BP and Graph-cut on the *map* image from the Middlebury database. A coarse to fine strategy was used in case of both Mean Field and BP based on the method suggested by Felzenszwalb and Huttenlocher [2006]. As can be see the performance on this simple image is comparable in

all the three cases.

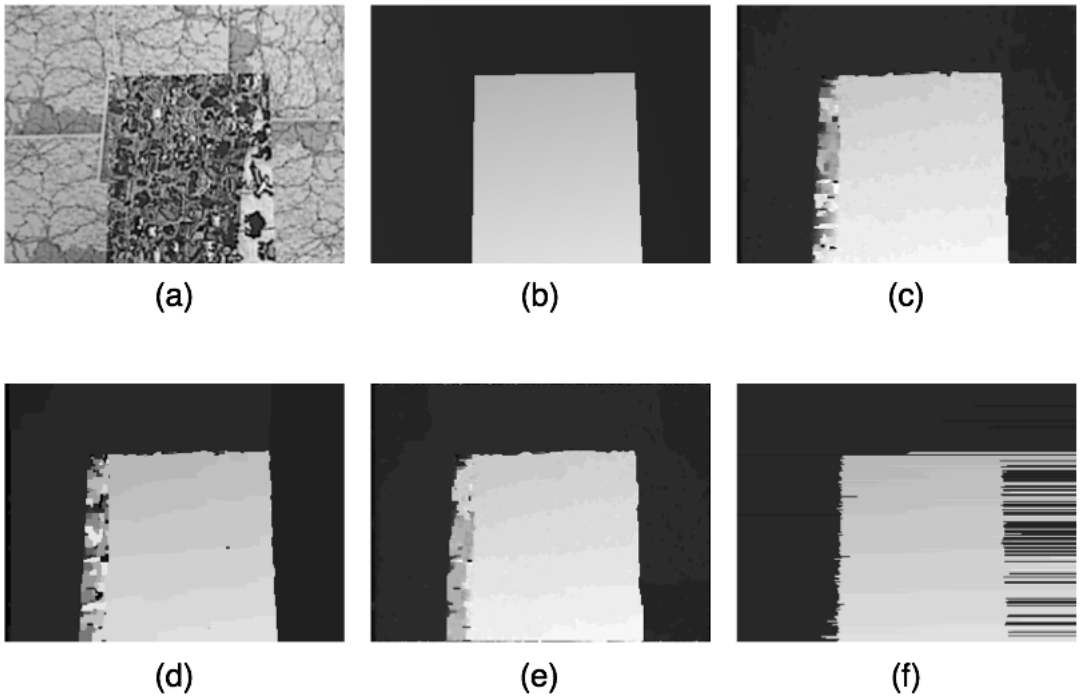


Figure 2.6: (a) Shows Left image of the stereo pair *map*. (b) the disparity ground-truth. (c), (d) (e) and (f) show the disparity maps estimated using Mean Field, BP, Graph Cuts and Dynamic Programming respectively. (f) shows the "streaking" caused by the Dynamic Programming algorithm (discussed in section 2.3.4).

2.3.4 Other methods

Among other discrete optimization methods to minimize (2.22, page 24) is Dynamic Programming (DP) (Birchfield and Tomasi [1999a], Ohta and Kanade [1985], Torr and Criminisi [2004]). While (2.22) is defined on a 2D-grid, the DP algorithm finds the global minimum for independent scanlines (owing to the epipolar constraint). However, when used in stereo matching DP suffers from "streaking" effect (See figure 2.6(f)). This is because of the difficulty in enforcing the inter-line consistency. Moreover, DP makes use of the ordering constraint, which may not always be true. Kolmogorov et al. [2006] suggested an extension called *layered* DP to overcome the problem interline consistency.

Alternatively, the stereo problem can be expressed in continuous form with a continuous disparity range. In such cases variational methods like the one suggested by Alvarez et al. [2000] can be used to minimize the energy functional. Alvarez et al. derive the Euler-

Lagrange equations, which are partial differential equations (PDE), whose solutions are the fixed points of the energy functional. They seek the solution of these underlying PDEs using a gradient descent method. In order to reduce the risk of being trapped in some irrelevant local minima during the iterations, the authors use a focusing strategy based on a linear scale-space. Faugeras and Keriven [1998] suggested the use of Level-Set methods to deform an initial set of surfaces obtained the Euler-Lagrange PDEs. Convex programming methods were also suggested by Miled and Pesquet [2006] and Pock et al. [2008]. Bhusnurmath and Taylor [2008] used the interior point method to solve the matching problem by approximating the data term with a convex function and using Laplacian terms for interaction. While these methods allow computation of sub-pixel disparities directly, they suffer from a major drawback in that they require the energy functional to be convex and require the minimization to be carried out on a convex set. This means that the final disparities produced will be smooth across the boundaries. It is also more difficult to introduce occlusion handling, given the restrictions on the energy function. Furthermore, methods using PDEs require discretization which is not straightforward owing to numerical instability.

Another technique which allows the direct measurement of sub-pixel disparities is the phase-based technique (Sanger [1988], Fleet et al. [1991]). This method finds a solution to the correspondence problem by using the differences in the phase of local spatial frequency components. Odd and even one-dimensional spatial Gabor filters, at different spatial frequencies are convolved with the stereo pair. The difference between the phase at corresponding points in the two images is used to find the stereo disparity. Two main limitations of this method are: (1) the maximum disparity is limited by the filter width, therefore allowing only small disparity range in practice. (2) phase information is very sensitive to local image characteristics. This is mainly because of singularities in the phase-signal, which are often by caused textureless regions in the images.

Finally there are methods like the one suggested Zitnick and Kanade [2000] which do not fall into the category of either global or local methods. Zitnick and Kanade's algorithm is variant of Marr and Poggio [1976]'s and is inspired by human stereo vision. This algorithm performs local computations by imposing constraints on uniqueness of matches and smoothness. Furthermore, it uses a 3D support window to allow for slanting surfaces. While the computations are carried out locally, the optimization is done iteratively using nonlinear operations. This results in a behaviour similar to that of global optimization algorithms.

As stated by Szeliski et al., while there are number of methods to optimize, the main requirement is actually to have an energy which is representative of the scene. However, having an accurate function is not necessarily good as it may be too complex to optimize. As result there is trade-off on what a good energy is and how this function could be optimized. We will now continue our discussion how the model can be improved by incorporating additional cues.

2.4 Additional cues and constraints

The energy in (2.9), re-written below, is referred to as a basic stereo matching model:

$$E(\mathbf{d}) = \sum_{\mathbf{x}} V_{\mathbf{x}}(d_{\mathbf{x}}, \mathbf{I}) + \sum_{\mathbf{x}} \sum_{\mathbf{y} \in \mathbf{N_x}} V_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x}}, d_{\mathbf{y}})$$

As mentioned earlier, there are inherent ambiguities in the stereo model that need to be taken into account:

- **Occlusions**: Some points in the scene are visible in one image but not in the other, as a result there may be pixels in one image that do not have a correspondence in the other. If this information is not incorporated in the energy model, then erroneous results will be obtained in such regions.
- **Depth discontinuities**: While most algorithms use a discontinuity preserving function in the interaction term, to allow for disparity jumps, proper localization of the discontinuities is still a problem.
- **Fronto-parallel assumption**: Most methods formulate the energy function in such a way that they make an implicit or explicit assumption that the disparity over a region is constant. This assumption leads to inconsistency in disparity values with respect to the real surfaces in the scene.

In order to account for these issues, the energy model introduced in section 2.2.1 (and in section 2.2.2.1) must be adapted. This is usually done by incorporating both binocular and monocular cues (like colour, gradient etc., from the reference image), and taking into consideration the geometrical properties of the scene itself. We will now discuss some of the methods used to tackle each of these problems.

2.4.1 Occlusion handling: Additional binocular cues

The data term in (2.11) (or the likelihood term in (2.20)) usually assumes that a point on the surface has the same colour when viewed from different angles. However, this is not true as some points may be occluded by other parts in the scene. This means that some points in the reference image may not actually have corresponding point in the other image. In order to overcome this problem, one of the solutions is to *cross check* if disparities from the left image to right are consistent with disparities from the right image to left (shown in figure 2.7). This technique was first introduced in Bolles and Woodfill [1993] using correlation, but some recent papers such as the ones by Yang et al. [2009], Hirschmüller [2008], Xu and Jia [2008] and Sun et al. [2005] have used similar checks within their MRF framework.

The figure 2.8, shows how the region (shown in red) occurring around the discontinuity, is visible only in the right image not in the left. Thus illustrating that the occluded regions occur mainly at the depth discontinuities. Belhumeur and Mumford [1992], Geiger et al. [1995] and Bobick and Intille [1999] use this information within their energy minimization

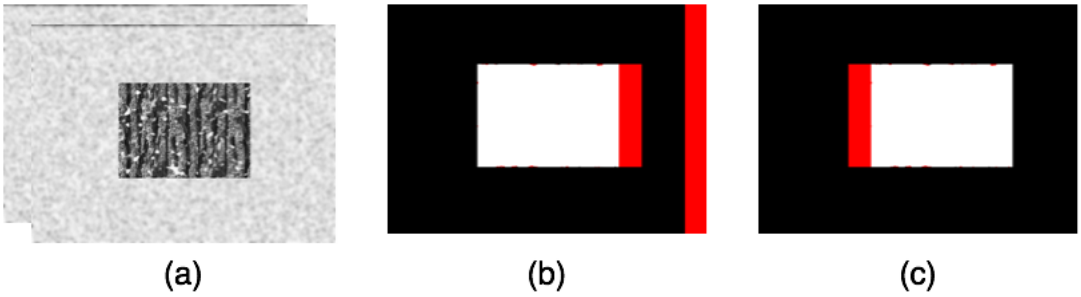


Figure 2.7: Cross Check applied to the texture images in (a) shows the occluded regions in red. (b) is the disparity map with right image as reference and (c) is the one with left as reference.

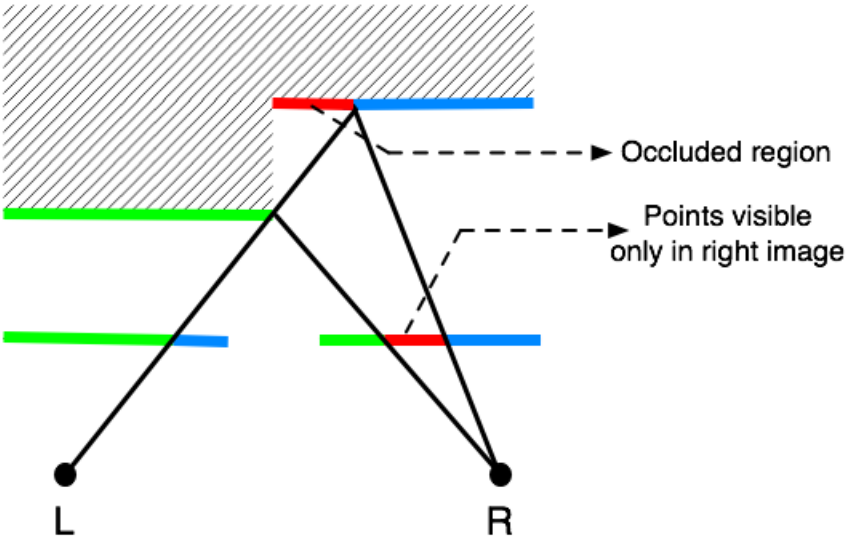


Figure 2.8: L and R represent the left and right cameras. Regions shown in green and blue in the scene are visible both in the images. However, a part of scene near the discontinuity (shown in red), is visible only to the right camera but not to the left.

algorithms to handle occluded regions. The model presented by Belhumeur and Mumford [1992] uses horizontal and vertical *line processes* to represent the discontinuities and use them as strong clue for modelling occlusions. Line processes are binary variables located at sites between the disparity lattice (figure 2.4) that indicate the presence or absence of a discontinuity between two adjacent disparities. Unlike Belhumeur and Mumford, Geiger et al. [1995] formulate the problem in the matching space. The disparity map is represented by a path in the matching space, which is broken either when a discontinuity is detected or when an occluded region occurs. These algorithms make extensive use of the ordering

constraint, which states that if an object is to the left of another in one stereo image, it is also to the left in the other image. Alternatively, Zitnick and Kanade [2000] use a uniqueness constraint that enforces a one-to-one mapping between the two images in 3D disparity image space. Such constraints are not always true, depending on the scene geometry.

Most of recent algorithms model the occlusions taking into account the visibility of the point in the two images. That is each correspondence is associated with a binary label indicating if a given point is visible in the other image or not. Sun et al. [2005], Hirschmüller [2008] and Yang et al. [2009] make the visibility explicit through cross checking. Sun et al. [2005] iterate alternately between disparity and occlusion variables during optimization to determine the two simultaneously. In contrast, Yang et al. and Hirschmüller infer the disparities for the occluded pixels through the unoccluded ones. While Yang et al. use plane-fit and hierarchical BP to propagate the disparities from the non-occluded to the occluded ones, Hirschmüller extrapolates the disparities from the background into the occluded regions. Strecha et al. [2006] model the occlusion as an outlier process by constructing a colour model for all the pixels which have no correspondences established. They model the visibility and depth jointly as a Hidden-MRF and obtain the statistics of both inlier (disparity) and outlier process using Expectation-Maximization. Kolmogorov and Zabih [2001] and Ishikawa [2003] use a binary visibility constraint in Graph Cuts framework. Ishikawa treats the stereo images symmetrically and enforces the uniqueness constraint. However, as their interaction term is convex they do not produce good results at discontinuities. On the other hand, Kolmogorov and Zabih handle occlusions along with a robust interaction term within the Graph Cuts framework.

Xu and Jia [2008] use an outlier confidence approach, to deal with occlusions. Instead of assigning a binary label to each pixel (occluded or unoccluded), they associate a confidence measure to determine if a pixel is an outlier or not. They first estimate an initial disparity map using BP. The outlier confidence is then estimated on this initial disparity map, based on the beliefs generated and the inter-frame consistency. This estimated outlier confidence is subsequently used in an overall global optimization again based on BP to obtain the final disparity map. Methods such as that of Sun et al. [2005], Xu and Jia [2008] and Yang et al. [2009] deal with left and right images symmetrically in order to handle both right and left occlusions simultaneously. In contrast Min and Sohn [2008] handle the occlusions asymmetrically, that is only the left (or right) disparity field is used to estimate the occluded pixels. This asymmetric occlusion detection is done using geometric and photometric constraints, and is then used in a BP framework to determine the final disparity map.

Once the occluded regions are determined, the problem now is to assign disparities to these regions. While some of the algorithms described above provide solutions for this, the important aspect that needs to be taken care of is that the disparity discontinuities are retained. We will now discuss few algorithms along with the ones discussed here in detail, to see how the disparity discontinuities are handled by existing approaches.

2.4.2 Localizing disparity discontinuities: Colour and gradient cues

As we have seen in section 2.2.1, the interaction term (2.12) if modelled as a robust function (see figure 2.3, 21) allows for jumps in disparities. This means that it does not smooth across very different disparities. Despite this, the disparities are often wrongly estimated in the neighbourhood of discontinuities. In other words, these functions do not necessarily ensure proper localization of discontinuities as seen in figure 2.9. This figure further illustrates that the disparity discontinuities are closely related to occluded region.

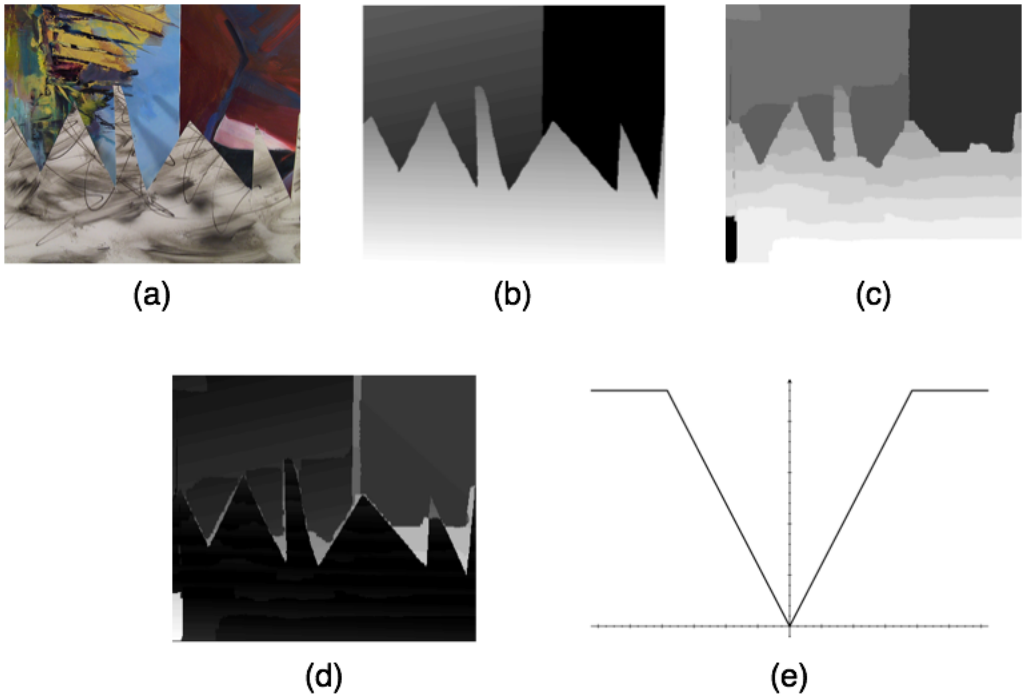


Figure 2.9: (a) Shows the original image. (b) Disparity Ground-truth. (c) Disparity estimated using BP and using (e) as interaction function. (d) Absolute Error between (b) and (c). Large errors have higher grey-values (lighter intensity values) and are concentrated mainly at the disparity discontinuities. (e) Truncated Linear function.

As discussed in the previous section there are methods (Belhumeur and Mumford [1992], Geiger et al. [1995] and Bobick and Intille [1999]) which deal with the problem of localizing disparity discontinuity simultaneously with occlusions using line process. These methods required another variable, i.e. line process, to be estimated along with disparity. Another way of estimating disparities properly at discontinuities is to use cues, such as colour, edges. Some of the early algorithms such as those by Baker and Binford [1981], Ohta and Kanade [1985] used edge information from the two images. Baker and Binford [1981] performed

the stereo matching based on the edges and then extended the initial solution to get the disparities in the areas between the edges based on intensity information. In contrast, Ohta and Kanade estimated dense disparity maps using DP and then used edge matching to propagate the disparities across scan-lines. The popular technique, however, is to use colour segmentation of the reference image and to associate smooth/planar disparity surface with each segment, thereby aligning the disparity discontinuities with colour edges. In most cases the reference image is over-segmented, that is segmented into large number of regions irrespective of the whether it belongs to an object. This is done to ensure that the region edges correspond to the colour edges by taking into account all discontinuities. Tao et al. [2001] were among the first to use such a method. Their algorithm used a greedy search to generate disparity hypotheses for each segment from its neighbouring segments. Each of these hypotheses were then tested using a quality measure ascertained by warping of the images. Bleyer and Gelautz [2004] used a similar approach to obtain proper disparities at the discontinuities. However, they used a *layered* representation providing more robust solutions. A layered representation allows segments with similar disparities to be approximated by the same planar equation.

Sun et al. [2003] incorporate colour segmentation information in a BP framework as an additional prior energy term. In contrast to the robust surface fitting techniques, this energy simply penalizes disparities within a segment if they are different. As a result, this constraint is not strong enough, and shows only marginal improvement in localizing the discontinuities properly. Hong and Chen [2004] formulated the stereo matching problem as an energy minimization problem in the segment domain instead of the traditional pixel domain. They then used Graph Cuts to assign the corresponding disparity plane to each segment. While this approach produced good results at discontinuities, it does not give satisfactory results for highly textured images such as the *map* image in figure 2.6(a). This is because in highly textured images over-segmentation does not ensure the inclusion of all the discontinuities. In order to overcome this, Sun et al. [2005] used segmentation as a soft constraint. They estimate the plane parameters in each segment using a robust plane fitting algorithm. These plane parameters are then used as a soft constraint within their stereo model, which is optimized using BP. Klaus et al. [2006] suggested the use of a data measure which incorporated not only the intensity differences between the right and left images but also the gradient difference between the two. With this as the data term they used a similar energy formulation as Hong and Chen [2004] in the segment domain, and used BP for optimization.

In an attempt to consider both colour segmentation and disparity at same time, Chang et al. [2007] suggested a technique to estimate the disparity map in two steps: The first step involves estimating a dense disparity map using a few *reliable disparity values*. This problem is formulated as an energy minimization and optimized using Graph Cuts; The second step involves finding denser reliable disparity values through cross check and modified mean-shift filtering. The mean-shift filtering (see Comaniciu and Meer [2002] for details) is modified to incorporate not only the colour and position but also the disparities. The two steps are

iterated until convergence.

Zitnick and Kang [2007] use segmentation of both images and match the segments instead. Furthermore, they do not use over-segmentation by mean-shift filtering like most of the other segmentation-based approaches. Instead, they first partition the images into grids of equally sized segments and refine the shape and size of each segment using iterative K-means algorithm. With these segments as nodes of the MRF, disparity is estimated for each segment. They assume constant disparity over each of these segments, and do not consider any planar model. In same lines as the segment-based optimization techniques is the one suggested by Wang and Zheng [2008]. The basic idea is to optimize the disparity plane parameters of all the regions using a inter-regional cooperative optimization. The initial plane parameters are determined by a window-based disparity estimation and plane fitting within each segment. Xu and Jia [2008] perform a BP optimization at pixel-level, but they introduce the segmentation and plane-fit in their data-term as a soft constraint, as in Sun et al. [2005], to improve their results at discontinuities.

Recent work by Yang et al. [2009] involves the combination of several techniques: colour-reweighted correlation suggest by Yoon and Kweon [2007], BP as suggested by Felzenszwalb and Huttenlocher [2006], cross-check by Bolles and Woodfill [1993], colour segmentation by Comaniciu and Meer [2002] and finally plane-fitting by Fischler and Bolles [1987]. They first find initial right and left disparity maps using BP and then perform a cross-check. The cross-check classifies the pixel disparities into stable. unstable, and occluded. A plane-fit is performed on segments obtained by colour segmentation using only the stable pixels. The data term is then modified to incorporate the plane-fit disparities as in Sun et al. [2005]. With this modified data term the overall BP optimization is carried out to estimate the disparity map.

In summary, most of the segmentation-based stereo algorithms do the following: perform colour segmentation as a preprocessing on either/or both of the reference images, determine the initial disparity map using known algorithms such as window-based, BP, Graph Cuts, perform plane fitting with each of the segment, and finally find the final disparity map by optimizing either over the plane parameters or by using plane-fit disparity as a soft constraint.

2.4.3 Disparity surfaces: Geometric constraints

As mentioned in section 2.2.1, a large part of the literature considers only the first order difference in disparities, that is pairwise terms, in their interaction function (2.12). Using such a function enforces a *fronto-parallel* assumption on the model. Such an assumption supposes that the scene under consideration can be approximated by a set of fronto-parallel planes (on which the disparity is constant) and thus biases the results towards "staircase" solutions. The figure 2.10(a) shows the 3D view of the ground-truth disparity for the stereo-pair in figure 2.9(a). The staircase solution produced by standard BP algorithm is shown in figure 2.10(b) in 3D and corresponds to the disparity map in figure 2.9(c).

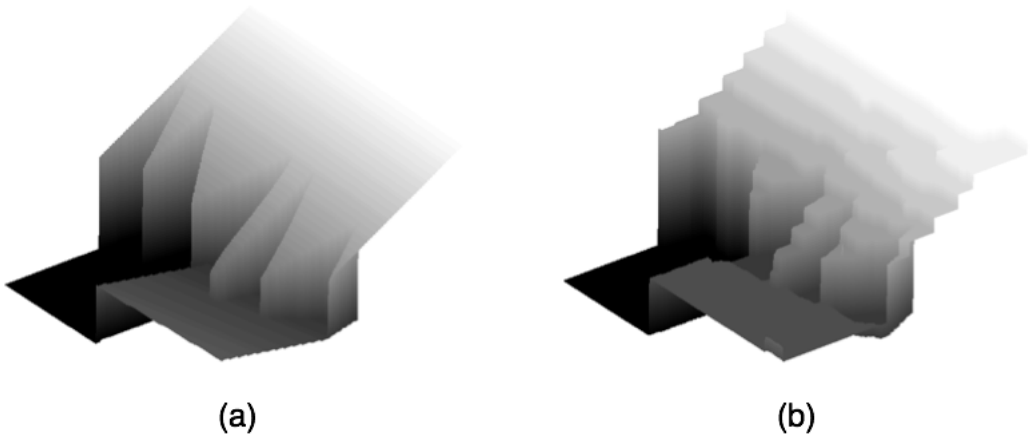


Figure 2.10: (a) 3D image of the Ground-truth disparity in figure 2.9(b). (b) 3D image showing the staircase effect due to fronto-parallel assumption.

Few attempts have been made to move beyond the fronto-parallel assumption in the stereo correspondence problem. In order to address this problem, Devernay and Faugeras [1994] proposed to extend the classical correlation method to estimate both the disparity and its derivatives directly from the image data. This is achieved by deforming the correlation window in one image based on the disparity and its derivatives. The derivatives are first initialized using plane-fitting and then a classical minimization method such as Levenberg-Marquardt is used to refine both the disparities and its derivatives directly. They then relate these derivatives to differential properties of the surface such as those encoded by normals and curvatures. However, in this paper the authors encounter numerical instability when computing second-order or higher order disparity derivatives.

Birchfield and Tomasi [1999b] and Lin and Tomasi [2004] propose to perform iteratively segmentation and correspondence. While Birchfield and Tomasi model the disparities within each segment as a plane, Lin and Tomasi model them as splines. The limitation of this approach is that the proposed algorithm is likely to get stuck in local minima in the presence of untextured surfaces. Furthermore, these methods do not consider the geometrical properties of the surface itself. The segmentation based approaches mentioned in the previous section allow the slanted surfaces to be recovered due to the plane-fitting used in their algorithms. Although most of them perform very well on the Middlebury data set [4], these approaches cannot handle curved surfaces.

Alternatively, Li and Zucker [2006a,b] explicitly take into account the differential geometric properties of the surface. Li and Zucker introduce the notion of geometric consistency between nearby matching pairs using both depth (position disparity) and surface normals. They measure the consistency of the normals by transporting them along the

4. http://www.middlebury.edu/stereo, by D. Scharstein, and R. Szeliski.

surface. They show that contextual information can be enriched with extra geometric constraints for stereo correspondence. As a consequence, they propose to use surface normals and curvatures to guide the disparity estimation towards a geometrically consistent map. The initial disparities and its derivatives (which are related to the normals) are computed using a method similar to that described by Devernay and Faugeras. However, in order to overcome numerical instability issues encountered by Devernay and Faugeras, they perform all the derivative computations in the depth space. The optimization is performed using cooperative algorithms by Li and Zucker [2006a] and using BP by Li and Zucker [2006b]. One main limitation of this algorithm is that it precomputes the local surface normals.

As the smoothness prior takes into account only the first-order differences (pairwise terms), one solution would be to use second-order (triple cliques) or higher-order priors. An important issue in considering the higher-order cliques is that there not many techniques that can optimize such functions as it becomes computationally infeasible. Bhusnurmath and Taylor [2008] used an interior point method to find the optimal disparity from an objective function that included first-order as well as second-order differences. However, in order to apply this method, both the data term and interaction term are required to be convex.

Recently, Woodford et al. [2009] used a second-order smoothness prior to encourage planar disparities in their Conditional Random Field (CRF) model. A CRF (Lafferty et al. [2001]), like an MRF is an undirected graphical model, where the conditional probabilities of the label sequence can depend on arbitrary, non-independent features of the observation sequence, without forcing the model to account for the distribution of those dependencies. To optimize the second-order CRF model (as applied to the correspondence problem), Woodford et al. used an extension of Graph Cuts, namely the Quadratic Pseudo Boolean Optimization (QPBO) algorithm. The QPBO algorithm only provides a partial labelling, that is only subset of sites will be assigned disparities, leaving the rest unlabelled. Woodford et al. therefore used an extension developed by Rother et al. [2007] to QPBO to solve for the labels of unlabelled sites. This optimization allows for the use of triple cliques by decomposing them into unary and pairwise cliques through the addition of a latent variable. This method was suggested by Kolmogorov and Zabih [2002b] to take into account triple cliques in Graph Cuts. While second-order priors account for planar surfaces, higher-order priors are required to model curved surfaces. Although QPBO can handle triple cliques, it is unclear how this can be extended to higher-order priors.

One of the most recent attempts to recover surfaces with different orders of smoothness is by Smith et al. [2009]. They move away from the MRF-stereo model and propose a non-parametric model for stereo matching. The main idea in their paper is to consider each pixel as a feature vector composed of position and colour, and each image as a point cloud in the the feature space. For each image they induce a dense graph with weighted edges. The dense graph is then approximated by a sparse one using only the dominant weights. They use a data term based on colour alone and the regularization is modelled as a Gaussian mixture with respect to disparity, where the weights for the Gaussian are

determined by both colour and position, i.e, from the feature vector. These regularization weights are also the ones that determine the sparsity of the graph itself. Graph Cut is then used to match the sparse graphs and to determine the disparities. The authors relate the networks within the graph to pixel grouping hierarchy, which allow the pixels with similar features to be grouped together, thus avoiding image segmentation. However, the density of the graph connections plays a key role as it determines the smoothness or otherwise of the final disparity map.

2.5 Motivation

As can be seen in the above sections, it is essential to use additional cues to overcome some of the problems of stereo matching. In this thesis we deal with the two issues discussed above, namely localization of disparity discontinuities and adding additional geometric constraints to relax the fronto-parallel assumption. We motivate our thesis as follows:

- We see that most of algorithms use preprocessing such as segmentation and plane-fitting to localize disparity discontinuities. In our work, we make use of an important observation that *the disparity discontinuities (which are also 3D depth discontinuities) occur usually at object boundaries.* Thus, object boundaries are an important cue for the disparity discontinuities and vice versa. We make use of this information to cooperatively estimate both disparity and object boundaries (chapter 4).
- Recovering binocular disparities in accordance with the surface properties of the scene under consideration involves either using higher-order neighbourhoods or including the geometric contextual information in the energy function. As optimization of higher-order neighbourhoods involves using "non-standard" techniques of optimization, we take inspiration from Li and Zucker's work where the geometric constraints are included within pairwise MRF frame-work. While their work allows the use of standard optimization techniques, the differential geometric constraints, which are included as surface normals, are precomputed. We on the other hand, model the disparity and normals in such a away that each can be estimated taking into account the other. Therefore, simultaneously updating the two until, both disparities and surface normals are accurately estimate (chapter 5).

The next chapter introduces *Coupled-Markov Random Fields* to model the problems stated above. In both of the above two cases, we use two random fields to model each variable, that is one for disparity and the other for either discontinuities or normals, but at same time includes information about one-another. Thus, we model two posterior distributions for each variable by including the information from the other as an observation. We then find the MAP estimates for each of these sets of variables using an *Alternating Maximization* framework.

Before we go into the details of each of these models, we first provide a brief state-of-the-art on Coupled Random Fields in the next chapter. We discuss some aspects of alternating maximization and how it can be used in our context to determine the MAP estimates of the system under consideration.

Chapter 3

Coupled Markov Random Fields

As mentioned in the previous chapter, we address two problems in this thesis, namely, localization of disparity (depth) discontinuities and recovering disparities in accordance with the surface properties of the scene. We see that, in stereo:

- Brightness edge may indicate the presence of depth edge and vice versa;
- The depth values of the surface should agree with the surface properties, such as surface normals, of the scene.

Thus, the stereo matching problem is related to two problems, namely, the one of boundary extraction and the other surface normal estimation. Intuitively, the solutions of each of these individual problems could impose constraints on the other, thereby simultaneously improving the results of the two. Such a cooperative setup can be modelled using *Coupled Markov Random Fields* (coupled-MRFs). The main advantage of modelling a problem based on coupled-MRFs is that it allows simultaneous estimation of several related functions from one or more modalities of observations.

Let $\mathbf{A}$ and $\mathbf{B}$ be two MRFs representing related quantities and associated with observations $\mathbf{I}_1$ and $\mathbf{I}_2$. Coupled-MRFs allows to model $\mathbf{A}$ and $\mathbf{B}$ given the observation using the Gibbs distribution as:

$$P(\mathbf{a}, \mathbf{b}|\mathbf{I}_1\mathbf{I}_2) = \frac{1}{Z} \exp\big(- E(\mathbf{a}, \mathbf{b}|\mathbf{I}_1\mathbf{I}_2)\big) \tag{3.1}$$

where Z is the normalization factor, $\mathbf{a}$ and $\mathbf{b}$ are realizations of $\mathbf{A}$ and $\mathbf{B}$ respectively, and $E(\mathbf{a}, \mathbf{b}|\mathbf{I}_1\mathbf{I}_2)$ represents the energy involving the two random variables and the observations.

In this chapter, we try illustrate the basic idea behind coupled-MRFs through the numerous applications it has been used, such as: image segmentation, motion, boundary detection, and stereo. In the next section we discuss related work concerning coupled-MRFs as applied to different problems. We summarize some aspects are of the coupled-MRFs in section 3.2. We then discuss one of the methods to optimize such coupled-MRFs, namely Alternating

Maximization, in section 3.3. Finally in section 3.4, we present the highlights of the thesis with reference to coupled-MRFs and Alternating Maximization.

3.1 Related work

The main idea behind coupled-MRFs, is to model the two or more related problems using respective number of MRFs which are coupled. This coupling is nothing but the mutual constraints imposed by the solutions of each individual problem and is incorporated within the energy function(s) associated with the MRF. The hope is then to find an integrated cooperative solution that significantly improves individual results of each sub-problem. We now illustrate some applications where the coupled-MRFs are used to illustrate this idea.

3.1.1 Line process-based coupled-MRFs

In this section, we discuss papers that use the concept of *line process*, which was first introduced by Geman and Geman [1984]. The line process variables are located at sites between MRF lattice, which is based on image pixel grid, and indicate the presence or absence of a discontinuity between two adjacent elements. The line process-MRF can be visualized as the "dual" of the lattice where the edges become sites, and vice versa (See figure 3.1).

Surface interpolation

Among the earliest algorithms to use such coupled-MRFs, is the one suggested by Marroquin [1984]. He proposed an approach for surface interpolation from noisy sparse data taking into account the discontinuities. Marroquin modelled the behaviour of a piecewise smooth surface using two MRFs that are coupled: a continuous-valued one that corresponds to the depth at each location, and a binary one composed of horizontal and vertical line processes (See figure 3.1). Referring to the example (3.1) in the previous section $\mathbf{A}$ corresponds to the depth, $\mathbf{B}$ to the binary line process and the noisy sparse depth data corresponds the the observation $\mathbf{I}_1$ (given only single modality of observation). At every global iteration, all the line and depth sites are visited sequentially. When a line site is visited, its state is updated using the Metropolis algorithm (Kirkpatrick et al. [1983]). The depth values are updated using a gradient descent algorithm. Thus, Marroquin created two processes that are decoupled, where the continuous field finds its equilibrium instantaneously after the update of the line process. From a conceptual viewpoint, this was first algorithm to illustrate the advantage of performing the boundary detection and interpolation tasks at the same time. It showed that, it was possible include prior knowledge about the smoothness of the surface and about the geometry of the discontinuities, as well as the information provided by the observations in the same energy minimization framework. However, this method had the disadvantage that the line process was only determined by the sparse depth data. As

consequence of this the estimated discontinuities were offset and ragged as compared to the true ones.

Geiger and Girosi [1991] approached used a similar model as Marroquin for surface reconstruction. But instead of using Metropolis algorithm for line process and gradient descent for surface interpolation, Geiger and Girosi use a deterministic Mean Field approximation (Yuille et al. [1990]) to compute the estimate of the mean values of the surface field and the discontinuity field (line-process). Geiger and Yuille [1991] extended the work done by Geiger and Girosi for image segmentation. Furthermore, Geiger and Yuille instead of considering two line process fields (horizontal and vertical), used a single continuous line process field.

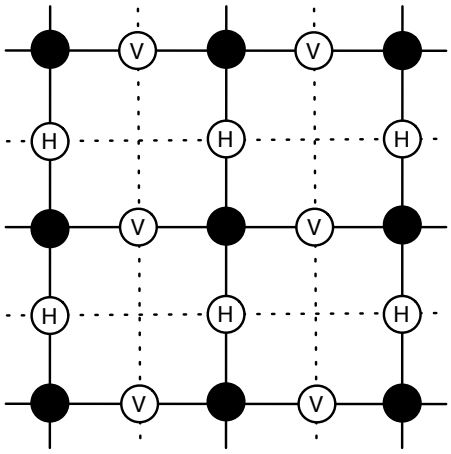


Figure 3.1: The continuous valued MRF lattice is shown with solid lines and black dots. The horizontal line-process and the vertical line-process variables are represented by "H" and "V" within circles respectively, both of which lie between the continuous lattice. Note that the line-process MRF can be seen as a dual lattice (shown using dotted lines) where the continuous lattice edges become sites, and vice versa

Estimating discontinuities from several vision modules

Gamble and Poggio [1987] proposed a scheme integrating colour, texture, motion and stereo, together referred to as *vision modules*, to find surface discontinuities. Each of these vision modules were modelled using a coupled-MRF consisting of a continuous process and a line process as done by Marroquin [1984], i.e., within a single energy function. The discontinuity output of each of these modules are coupled to intensity edges obtained directly from the image data. These outputs are then provided as an input to a simple linear classifier to obtain the final discontinuity map. While they provided a method to integrate different vision cues to find the discontinuities, they did not model the interaction between the line processes of different vision modules. The absence of cooperation between the vision

modules made it difficult to find reliable discontinuities.

Günsel et al. [1996] use an idea similar to Gamble and Poggio [1987] to find discontinuities, but instead of using different visual cues as coupled-MRFs, they use the image information at different scales. That is the vision modules in the case of Günsel et al. consist of coupled-MRFs representing intensity and line process at different scales. Each of these modules, along with the term that couples the line processes at different scales, are then integrated into a single MRF. The posterior distribution corresponding to this MRF is maximized using Thresholded Posterior Mean (TPM), instead of the Maximum A Posteriori (MAP) estimate. In order to find the TPM estimate they use a modified Simulated Annealing procedure.

Optical flow estimation

Similar to the previous approaches Heitz and Bouthemy [1993] handle the problem of optical-flow (velocity vectors) estimation and discontinuity processing simultaneously in a coupled-MRF framework. They use dense gradient and sparse edge-based measurements to estimate both velocity and discontinuity maps. They use continuous-MRF for velocity field and binary-MRF composed of line processes (both horizontal and vertical) to represent discontinuities that are located between velocity vectors. The additional feature of this model is that it takes into account occluded regions corresponding to the discontinuities, by considering not only the intensity edges but also the direction of the motion of the edges. The optimization of the MRFs is carried out sequentially by updating the different sites of velocity and discontinuity fields, using Iterated Conditional Modes (ICM introduced by Besag [1986]). The main disadvantage of this approach was that the estimated discontinuities were locally broken. The reason for this is the absence of any smoothness term associated with line processes, to ensure their continuity.

Processing acoustic images

Murino and Trucco [1998] use a three-fold process for reconstruction, restoration and segmentation of acoustic images. An acoustic camera transmits an acoustic signal and collects the returns from targets. These are then processed in such a way that range information and acoustical intensities can be retrieved for several viewing directions. This information is usually viewed as (acoustic) image data, i.e., noisy range and confidence (acoustical intensities) images. While the range data represent the depth, the confidence data associated with each depth measure denotes its reliability. The idea is then to reconstruct the 3D images using the noisy input data. Murino and Trucco use three MRFs associated with confidence, range and line process. The process of reconstruction consists of segmentation of the images using the line process and confidence field. The estimated confidence values are used to smooth inside the object surfaces taking into account the discontinuities using the line process. They express the three-fold process in a single energy function and optimize the

fields using Simulated Annealing by sequentially updating the corresponding fields.

Image restoration

In line with the previous methods, Bedini et al. [2001] suggests a coupled-MRF approach for edge-preserving image restoration. Here again, the image intensities are modelled as a continuous field and the line process as a binary field. However, unlike previously mentioned methods, Bedini et al. perform a mixed annealing where the intensities and the binary variables are updated alternately. The strategy assumes that the energy function with respect to intensity is convex and quadratic, when the line process configuration is fixed. So, they reformulate the continuous energy function as a least squares minimization and use deterministic techniques such as conjugate gradient techniques. The line process variables are estimated using Simulated Annealing, by drawing the samples conditioned on the intensity process. This kind of mixed annealing process does not guarantee overall convergence. However, as the Simulated Annealing procedure is used only for binary-line processes, and not for the continuous valued MRF, the computational efficiency is improved.

Elimination of line process

Blake and Zisserman [1987] suggested a way of eliminating the line process, by replacing it with a *weak continuity constraint*, when used in conjunction with a continuous variable. This weak continuity constraint imposes a fixed penalty for discontinuities within the continuous energy function. Thereby allowing the discontinuities to occur and eliminating the need for line process. Furthermore, Blake and Zisserman showed how this constraint allows to retain all the properties of the line process. While they described a general energy model over two variables; one continuous and the other line process, they did not consider a probabilistic MRF framework. Black and Rangarajan [1996], further extended these findings and showed that the such line process-based discontinuity models could be replaced by robust functions (see chapter 2, section 2.2.1).

The idea of using robust functions instead of line process has been used widely in MRF-based techniques, for example by Strecha et al. [2006], Felzenszwalb and Huttenlocher [2006], Boykov et al. [2001]. In particular, Sun et al. [2003] and Xue et al. [2008] model the problem of disparity estimation and multiple target tracking in coupled-MRF framework respectively. Both Sun et al. and Xue et al. use one discrete multi-label-MRF and two binary-MRFs. In the case of Sun et al., the three MRFs are used to represent disparity, discontinuity line-process and occlusions. In contrast, Xue et al. use the MRFs to represent joint state of multi-targets, existence of each target and occlusions between adjacent targets. Both Sun et al. and Xue et al. eliminate the two binary-MRFs by introducing two robust functions in their place. The main advantage of this elimination is that it allows for estimation only the desired MRFs. Sun et al. and Xue et al. used Belief Propagation to estimate disparity and target state for tracking respectively.

3.1.2 Coupled-MRFs without line process

Most of the methods suggested above use a implicit/explicit binary discontinuity-MRF within their coupled-MRF. Furthermore, the line process-MRF is interwoven with the continuous valued MRF (see figure 3.1). In the following section we discuss applications which use coupled-MRF model without the interwoven line process.

Image segmentation

Wu and Chung [2007] propose a method for segmentation of images coupled with a boundary model. The couple-MRF therefore consists of a label field for segmentation and binary field for the boundary. The boundary field of Wu and Chung is different from the normal line process field in that it does not exist on a dual lattice between the pixels, but on the pixel sites directly. Their main contribution lies in modelling the interaction between the boundary and label field. Unlike other approaches where the line process works implicitly and is simple (horizontal and vertical), this method takes into account complex boundary patterns. They first define all possible boundary patterns within a 3×3 window and express the interaction between the two MRFs based on them. In other words, the energy representing the interaction between label and boundary MRFs allows only the defined patterns to exist, penalizing heavily the undefined boundary configurations. They argue that defining boundary configuration within a window makes boundaries less sensitive to noise, as compared to dealing with individual line processes (as done by Heitz and Bouthemy [1993]). Optimization of the model is performed using ICM and is compared to the models suggested by Geiger and Girosi [1991] and Geman and Geman [1984]. While Wu and Chung's method shows improvement over the others, it still requires the construction of all possible boundary configuration within a window.

Texture segmentation

Xia et al. [2006] perform adaptive segmentation of textured images using coupled-MRFs. The model consists of two mutually dependent components: one for estimating the feature vectors (gray-level statistics and local texture content over a window) from an image and other to label the image to achieve segmentation. Like Wu and Chung, Xia et al. describe a coupled-MRFs, where each (feature vectors and labels) MRF is described on the same set of image sites. In other words the feature-MRF resides on the same lattice as that of the label-MRF lattice. Even though they do not use a binary line process like the other methods, in this paper the features and the labelling are defined using a single energy function representing the posterior distribution. While a simple MRF model would just find the posterior probability of the segmentation labels given the feature vectors, Xia et al. estimate the feature vectors themselves. The features are considered to be random vectors and each of these vectors are modelled as an independent Gaussian distribution. The segment labels are used to not only constrain the mean and variance of the Gaussian feature

field, but also to account for the pairwise relationship between the neighbouring labels. This pairwise relationship in the label field is modelled to favour neighbouring sites to have same labels. The optimization is carried out alternately between the feature and label field using Simulated Annealing. The labelling is assumed to be known when estimating features and vice versa.

Cooperative estimation of optical flow and disparity

The methods discussed until now (including section 3.1.1) use a *single* energy with two variables describing the same visual process in their coupled MRF model, for example surface interpolation, optical flow, stereo and their respective corresponding discontinuities or texture and their associated features. In contrast to such methods Nasrabadi et al. [1989] and Sudhir et al. [1995] model two different visual processes, namely optical flow and disparity, using coupled-MRFs. While Nasrabadi et al.'s method suggested only the improvement of disparity information using optical flow, Sudhir et al. introduced a cooperative mechanism between the two MRFs. This mechanism involves estimating both stereo disparity and optical flow, but constraining the estimation of each by the information provided by the other. Unlike the other described methods, Sudhir et al. define two energies; one for estimating disparities given the optical flow and the other for estimating optical flow. Referring to the example equation (3.1), if optical flow is represented by variable $\mathbf{A}$, disparity by $\mathbf{B}$ and the observed images represented by $\mathbf{I}_1$ and $\mathbf{I}_2$. Let $\mathbf{I}_1$ and $\mathbf{I}_2$ represent stereo pairs are different time instances. Now the energy function in (3.1) can be split into:

$$E(\mathbf{a}, \mathbf{b}|\mathbf{I}_1\mathbf{I}_2) = E_1(\mathbf{a}|\mathbf{b}\mathbf{I}_1\mathbf{I}_2) + E_2(\mathbf{b}|\mathbf{a}\mathbf{I}_1\mathbf{I}_2) \tag{3.2}$$

with $\mathbf{a}$ and $\mathbf{b}$ representing the realizations of $\mathbf{A}$ and $\mathbf{B}$ respectively. Each of these energies E_1 and E_2 incorporate discontinuity information within disparity and optical flow estimation respectively, based on the consistency between the right and left images given at two different time instances ($\mathbf{I}_1$ and $\mathbf{I}_2$). Sudhir et al. carry out the optimization of the two field alternately using ICM for a fixed number of iterations.

Estimation of disparity and occlusion

Similarly, Sun et al. [2005] use a coupled-MRF approach to estimate disparities and occlusions. While disparity is treated as a multi-label discrete field, the occlusions are modelled as a binary one. In this paper, they propose an algorithm that iteratively performs these two steps: (1) infer the disparities in one view considering the occlusions of the other view, and (2) infer the occlusions in one view from the disparities of the other view. They call this model symmetric, as it takes into account the observation from both views. The energy function for this model is similar to one in (3.2) with $\mathbf{A}$ standing for disparities, $\mathbf{B}$ for occlusions and only one pair images $\mathbf{I}_1$ as observation. The optimization is performed iteratively by alternating between disparity and occlusion estimation using Belief Propagation.

3.2 Summary basic concepts of coupled-MRFs

From the examples provided in the previous section, we see that coupled-MRFs provide a flexible way in which two related variables could be modelled and estimated simultaneously. Depending on the structure of the energy function it is also possible to model each variable separately, but taking into account the influence of the other. In the previous section, we categorized the coupled-MRFs into two one which use line processes (section 3.1.1) and the other which do not (section 3.1.2). The ones using line process with their framework usually modelled a interwoven coupled-MRF, that is the line process existed between the sites of the MRF-lattice. One exception to this is the work by Wu and Chung [2007], where the line-process existed on the sites. On the other hand, methods which do not use line process have both variable defined on the same MRF lattice. The table 3.1 provides brief summary of most of the applications and the associated methods of optimization presented in the previous section. There is one other way to categorize the methods presented in the previous section that is based on overall optimization:

1) Methods that optimize the energy function by sequentially updating each variable. Such techniques usually define a single energy function for all the variables, for example Marroquin [1984], Gamble and Poggio [1987] (each vision module is described by one energy function with two variables), Heitz and Bouthemy [1993], Murino and Trucco [1998] and Günsel et al. [1996].

2) Methods that eliminate one of the variables, in particular line process, and optimize the energy function over a single variable using standard techniques. The effect of the line process is included as a robust function, like in Sun et al. [2003] and Xue et al. [2008]

3) Methods that employ alternation to update each variable. In such cases, the each energy function incorporates the influence of the other as an observation (Bedini et al. [2001],Xia et al. [2006],Sun et al. [2005] and Sudhir et al. [1995]). These methods employ a strategy where the optimization of the fields are carried out alternately to estimate the each of the random variables involved. This procedure is often referred to as *Alternating Maximization*[1]. In such a procedure, the optimization of one of the variables is carried out assuming that the other(s) is(are) known. Alternating between the variables therefore allows finding estimates for the each variable involved in the coupled-MRF. Also, its iterative nature allows the mutual improvement of the solutions to the MRF-based problems involved.

In the next section we will discuss some of the aspects of this Alternation Maximization procedure. We also show the mathematical soundness of this procedure, when involving the probabilistic evaluation of the two variables.

1. Alternating Minimization if the corresponding energy is minimized, instead of maximizing the posterior.

Table 3.1: Summary of Methods using Coupled-MRFs

Applications	Coupled-MRF Variables	No. of Energy functions	Overall optimization	Optimization technique(s)
Surface interpolation Marroquin [1984]	Line process(LP) and Surface depth	one	Sequential	Metropolis algorithm and gradient descent
Surface reconstruction Geiger and Girosi [1991], Geiger and Yuille [1991]	LP and Surface depth	one	Sequential	Mean Field
Discontinuity estimation Gamble and Poggio [1987]	LP and image features like colour, motion, stereo	one/module	Sequential	Metropolis algorithm for each module
Optical flow estimation Heitz and Bouthemy [1993]	LP and Optical flow	one	Sequential	ICM
Acoustic image processing Murino and Trucco [1998]	LP, confidence and range	one	Sequential	Simulated Annealing
Image segmentation Wu and Chung [2007]	LP (not interwoven) and segment labels	one	Sequential	ICM
Disparity Estimation Sun et al. [2003]	Occlusion-LP, Discontinuity-LP and Disparity	one	Eliminate Line processes	BP
Muti-target tracking Xue et al. [2008]	existence of target-LP, Occlusion-LP and Multiple targets	one	Eliminate Line processes	BP
Image restoration Bedini et al. [2001]	LP and image intensity	one	Mixed annealing Like Alternation	Simulated Annealing and Least square minimization
Texture segmentation Wu and Chung [2007]	features vectors (not binary or interwoven) and segment labels	one	Alternation	Simulated Annealing
Optical flow and Disparity estimation Sudhir et al. [1995]	Optical flow and Disparity (not binary or interwoven)	two	Alternation	BP
Disparity Estimation Sun et al. [2005]	Occlusion (binary but not interwoven) and Disparity	two	Alternation	BP

3.3 Alternating Maximization

The Alternating Maximization procedure can be used to solve optimization problems over more than one variable in an iterative set-up. In order to understand the procedure, we commence with a simple example where the goal is to solve a maximization problem over two variables of the following form: given $\mathcal{P}$, $\mathcal{Q}$ and a function $f : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$, maximize f over $\mathcal{P} \times \mathcal{Q}$. That is, find

$$(P^*, Q^*) = \max_{P \in \mathcal{P}, Q \in \mathcal{Q}} f(P, Q). \tag{3.3}$$

Often maximizing over both variables simultaneously is not straightforward. However, maximizing with respect to one variable while keeping the other one fixed is in general, easy and sometimes possible analytically. In such a situation, the Alternating Maximization algorithm described next is well suited: start with an arbitrary initial point $Q^0 \in \mathcal{Q}$; for iteration t, compute

$$P^{(t+1)} = \arg\max_{P \in \mathcal{P}} f(P, Q^{(t)}) \tag{3.4}$$

$$Q^{(t+1)} = \arg\max_{Q \in \mathcal{Q}} f(P^{(t+1)}, Q). \tag{3.5}$$

In other words, instead of solving the original maximization problem over two variables, the Alternating Maximization algorithm solves a sequence of maximization problems over only one variable. The converged values are declared the solution to the original problem. This example can be extended to multiple variables where the maximization of the function $f(P, Q, R, \ldots)$ jointly over all variables is carried out by Alternating Maximizations over individual subsets of variables $P, Q, R, \ldots$.

We now explain this Alternating Maximization procedure in a probabilistic set up. Suppose $\mathbf{A}$ and $\mathbf{B}$ are two MRFs. These example random fields may represent disparity, motion, discontinuities or any such variables discussed above. The goal is to estimate realizations of $\mathbf{A}$ and $\mathbf{B}$ that are consistent with a joint probabilistic model and the observed data, like image(s) intensities or intensity gradients depending on the problem at hand. Ideally we are interested in finding the MAP (Maximum A Posteriori) estimates of $\mathbf{A}$ and $\mathbf{B}$,

$$(\mathbf{a}^{\mathrm{MAP}}, \mathbf{b}^{\mathrm{MAP}}) = \arg\max_{\mathbf{a},\mathbf{b}} \; p(\mathbf{a}, \mathbf{b}|\mathbf{I})\,. \tag{3.6}$$

where $\mathbf{a}$ and $\mathbf{b}$ are specific realizations of $\mathbf{A}$ and $\mathbf{B}$ respectively, and $\mathbf{I}$ represents the observed data. If we consider $\mathbf{a}$ to be disparities and $\mathbf{b}$ to be discontinuities; then the solution space of the above becomes too large to be computationally tractable. As a result, this global optimization problem has in general no straightforward solution. Thus, we consider the Alternating Maximization approach to solve the above equation, where the posterior probability is alternately maximized in the first and second variable. Starting from current

estimates $\mathbf{a}^{(t)}$ and $\mathbf{b}^{(t)}$ at iteration t, we consider the following updating,

$$\mathbf{a}^{(t+1)} = \arg\max_{\mathbf{a}} p(\mathbf{a}, \mathbf{b}^{(t)}|\mathbf{I}) \tag{3.7}$$

$$\mathbf{b}^{(t+1)} = \arg\max_{\mathbf{b}} p(\mathbf{a}^{(t+1)}, \mathbf{b}|\mathbf{I})\,. \tag{3.8}$$

It follows easily that,

$$p(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}|\mathbf{I}) \geq p(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}|\mathbf{I})\,. \tag{3.9}$$

While the above equation shows that the alternation procedure increases the posterior probability at each step, there is a possibility that it gets stuck in a local maximum owing to the non-convexity of the problem. Therefore, there is no guarantee that such a procedure leads to a global maximum. However, it is more adept at bypassing local maxima than the other approaches which sequentially update all the random variables involved. Each restricted iteration of Alternating Maximization is typically global, and is therefore able to "hop" great distances through the reduced variable space in order to find an optimal iterate. By using Bayes' theorem it is very easy to show that the alternation in equation (3.7) is equivalent to the following one,

$$\mathbf{a}^{(t+1)} = \arg\max_{\mathbf{a}} p(\mathbf{a}|\mathbf{b}^{(t)}, \mathbf{I}) \tag{3.10}$$

$$\mathbf{b}^{(t+1)} = \arg\max_{\mathbf{b}} p(\mathbf{b}|\mathbf{a}^{(t+1)}, \mathbf{I})\,. \tag{3.11}$$

So the above equation shows that the inference of each of the variables $\mathbf{a}$ and $\mathbf{b}$ can be performed by the using just the conditional distributions $p(\mathbf{a}|\mathbf{b}, \mathbf{I})$ and $p(\mathbf{b}|\mathbf{a}, \mathbf{I})$ respectively. This greatly simplifies the modelling as there is no need to specify a completely joint model. A specific model for each variable encoding the influence of the other in there energy term would be sufficient to fully describe the problem involved. Furthermore, each sub-problem can optimized using different techniques, thereby allowing more freedom in type of optimization technique to be used.

3.4 Coupled-MRF in the proposed approach

As can be seen from the discussion in section 3.1 the coupled-MRF approach has been used to tackle issues such as discontinuities and occlusion in stereo matching (chapter 2). As discussed in chapter 2, the main focus of this thesis is; (1) to estimate disparities in conjunction with boundaries;(2) to use surface geometric constraints to obtain disparities and surface normals. In both of these cases we use two random fields, one for disparity and the other for boundaries or surface normals. These two fields form parts of a coupled-MRF (section 3.1), where the two MRFs reside on the same lattice (unlike the interwoven line process-MRF shown in figure 3.1). In addition, each of these MRFs are described by separate energy functions incorporating the influence of the other as a observation. In

other words, we define conditional distributions for each MRF variable taking into account the influence of the other, which are then optimized using the Alternating Maximization procedure (section 3.3, Equation (3.10)). The advantages of using such a set up are as follows:

- The consideration of two separate random fields increases modelling flexibility i.e, the energy terms for two fields can be made more dependent or independent according to the information to be incorporated.
- Local constraints on each random field can be independently applied.
- The use of a well based statistical estimation framework (Alternating Maximization) resulting in cooperative estimation and mutual improvement of both the involved variables.
- There is no restriction on the type of optimization used for each step of the Alternating Maximization procedure.

These features allow us to model the stereo matching algorithm in a cooperative framework where the disparities can be mutually improved taking into account the boundaries or surface normals and vice versa. In the following chapters, which are the main contributions of the thesis, we will present in detail two models in which disparity is estimated cooperatively, one with boundaries and the other with surface normals.

Chapter 4

Cooperative Disparity Estimation and Object Boundary Extraction

In chapter 2, we demonstrated that additional cues such as colour and image gradient are essential in order to efficiently model the stereo problem. One of the issues we discussed was the localization of disparity discontinuities in the stereo matching problem. We saw that most of the existing algorithms (see chapter 2 section 2.4.2) have a preprocessing step in which they perform colour segmentation on either/or both of the reference images. They then determine the initial disparity map using known algorithms such as Sun et al. [2003] or Boykov et al. [2001] and find the final disparity map either by further optimizing over the disparity-plane parameters or by plane-fitting the disparities within each of the segments found during preprocessing.

In this chapter, we propose an approach which eliminates the need for preprocessing and uses the relationship between the disparity discontinuities and object boundaries to cooperatively estimate the two. We build on standard approaches for dense disparity estimation, and propose an original approach which simultaneously corrects disparity and finds the object boundaries. While these are usually considered as two separate tasks, in our approach they are dealt with cooperatively, i.e. the presence of disparity discontinuities aids the detection of object boundaries and vice versa. The proposed method relies on two assumptions:

(i) The discontinuities in depth are usually at object boundaries (which is true for natural images).

(ii) The disparity discontinuities obtained from naive disparity estimation are usually at the vicinity of actual depth discontinuities.

Thus, if we locate the object boundaries which are in the vicinity of the disparity discontinuities, we can correct the disparity values so that they fit more closely to the object boundaries. As mentioned in chapter 3, we use the coupled Markov Random Field

(coupled-MRF) to jointly model disparities and object boundaries and estimate them using the alternation maximization procedure [1]. In the next section we give an overview of our approach. This joint model is defined in section 4.2. The optimization techniques used and the alternation procedure are discussed in section 4.3 and section 4.4 respectively. Experimental results obtained using the proposed method are presented in section 4.5. Finally, we discuss the features and limitations in section 4.6.

4.1 Overview of the proposed disparity-boundary estimation

In this chapter, we propose a method to carry out cooperatively both disparity and object boundary estimations by setting the two tasks in a coupled-MRF framework. We define two MRFs, one for disparity and the other one for *displacement*. The displacement field models the corrections that need to be applied at disparity discontinuities in order to align them to the object boundaries. In this way, the disparity discontinuities can be then assumed to represent the object boundaries. In other words, the displacement field acts as an auxiliary field which provides a feedback from the boundary estimation to the disparity estimation. It is to be noted that we use the term *displacement* as it is associated with both a magnitude and a direction. While magnitude is set to one, it is the direction component of the displacement that is used for the correction of disparity discontinuities. The function of the displacement model is therefore to estimate the *directions* in which the discontinuities have to be moved. This information is incorporated in the disparity model so that the disparity values at discontinuities are corrected, such that they represent the actual object boundaries.

Note that depth discontinuities are a subset of image discontinuities. This is because if the appearance of two overlapping objects is uncorrelated, then the boundary between them will be evident in both the image and disparity map. There may however, be texture boundaries within each object that are not associated with any disparity discontinuity. In addition to that, the standard algorithms for disparity estimation do not localize the discontinuities properly. This limitation can be attributed to smoothness constraint imposed on the disparity modelling. In absence of explicit occlusion handling, even with use of robust interaction function, the disparity discontinuities occur at "improper" locations. As previously mentioned we assume that these "improper" disparity discontinuities occur in the vicinity of the actual object boundaries. This assumption enables us to search of near by gradient maxima which are most likely to be associated with the object boundary. Thus justifying our approach to find simultaneously the disparities and object boundaries.

Our coupled-MRF framework is similar to that of Wu and Chung [2007] in that we *do not* use a line process for boundary representation (Heitz and Bouthemy [1993], Geman et al. [1990]). This means that the field representing the boundary does not exist between the pixels (as in line processes) but on the pixel locations themselves. However, in contrast to

1. This work was originally published in the paper Narasimha et al. [2008].

Wu and Chung, we do not focus directly on the boundaries but provide the directions toward which discontinuities must be displaced, based on observed image gradient values. We have seen that the coupled-MRFs models based on line processes (chapter 3) define only a single energy for both line process and associated visual processes such as surface interpolation Marroquin [1984], Geiger and Girosi [1991] or optical flow Heitz and Bouthemy [1993]. Unlike such methods, in our coupled-MRF we define one energy function associated with each MRF which includes the influence of the other MRF. While this allows for flexibility in modelling each function separately, it also enables to cooperatively estimate both disparity and boundary.

The figure 4.1 gives a pictorial overview of the proposed method. It shows how the information in the displacement field is used to determine the disparity values at the boundary locations. The gradient map of the reference image (in our case the left image) is used as evidence for the object boundaries. The displacement values give the direction in which the disparity discontinuities have to be moved in order to be aligned with nearest maxima of the gradient magnitude. If the direction at particular location is non-zero it means that the disparity at that location may be wrong and should be corrected. Intuitively, from the figure 4.1(a) we see that disparity at such locations can be corrected by replacing the disparity at the current position by that of the neighbour in the opposite direction of the displacement. The figure 4.1(b) show the that the disparity model applies the disparity correction using estimated displacement values defined at each pixel location. This allows us to obtain corrected disparity and object boundaries at the same time. The resulting procedure involves alternation between estimation of disparity and displacement fields in an iterative framework. The overall procedure for simultaneously estimating disparities and displacements is as follows:

(i) Estimate an initial disparity map using a standard technique. In such a map disparity discontinuities occur at improper locations.

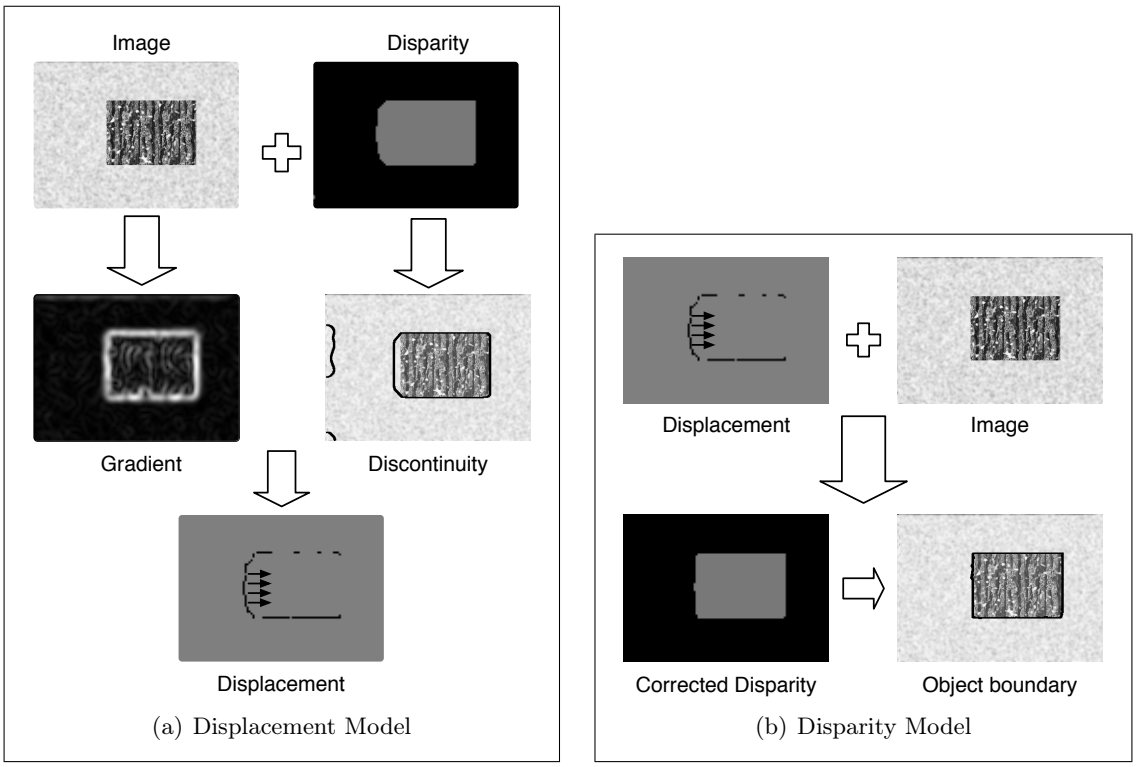
(ii) Extract these locations and estimate the displacement field. This field mainly estimates directions in which the discontinuities have to be moved to align with the nearby gradient maxima.

(iii) Re-estimate the disparities using disparities considering the influence of the displacement field. This allows to rightly estimate the disparities in the location where the corrections have to be applied.

(iv) Alternate between step (ii) and (iii) repeatedly until no corrections need to be applied.

(v) The disparity discontinuities now represent the object boundaries and the disparities are corrected.

The coupled-MRF model we propose results in a posterior distribution for both the disparity and displacement fields. This means that the fields can the be estimated according to the Maximum A Posteriori (MAP) principle. We use an Alternation Maximization pro-



(a) Displacement Model

(b) Disparity Model

Figure 4.1: The figure shows how the *displacement* information is used for disparity correction and vice versa

cedure, described in chapter 3, to estimate the two variables. This results in two conditional distributions; one for disparity and the other for displacement. While the disparities are modelled as an MRF, the model reduces to a Markov chain when considering displacement variables. The estimation of both disparities and displacement is alternated and done at multiple scales. The disparity-MRF is then optimized using variational Mean Field and the exact optimization of the Markov chain for displacement is carried out using the Viterbi algorithm.

4.2 Joint disparity and displacement model

We consider a finite set $\mathcal{S}$ of $p \times q$ pixels on a regular 2-dimensional grid. The observed data consists of the left and right images, $\mathbf{I}_L$ and $\mathbf{I}_R$, which are together referred to as $\mathbf{I}$. In our setting, the left image is taken as the reference image and, in particular, object boundaries are defined for this image. We denote by $\mathbf{D} = \{D_{\mathbf{x}}, \mathbf{x} \in \mathcal{S}\}$ the unknown disparity values at each pixel position $\mathbf{x} = (x, y)$. The values of $D_{\mathbf{x}}$ are considered as

random variables that take their values in a finite discrete set $\mathcal{L}$. $\mathbf{D}$ is referred to as the disparity field or disparity map and takes its values in $\mathcal{D} = \mathcal{L}^{p \times q}$. In addition, we consider a *displacement* field denoted by $\mathbf{A} = \{A_{\mathbf{x}}, \mathbf{x} \in \mathcal{S}\}$ where each $A_{\mathbf{x}}$ is made of two random variables denoted by $M_{\mathbf{x}}$ and $E_{\mathbf{x}}$. The variables $M_{\mathbf{x}}$ take their values in $\{-1, 0, 1\}$ and the variables $E_{\mathbf{x}}$ in the set of unit two-dimensional vectors denoted by $\mathcal{U}$. Mathematically, $E_{\mathbf{x}} \in \mathbb{R}^2$ and $|E_{\mathbf{x}}| = 1$, or equivalently, $E_{\mathbf{x}} \in \mathcal{U}$. For each pixel location $\mathbf{x}$, the $M_{\mathbf{x}}$ and $E_{\mathbf{x}}$ random variables are related to object boundaries since they give respectively the direction and the orientation in which the disparity discontinuity must be moved from its current position to a new position so that it matches the object boundaries. More specifically, $E_{\mathbf{x}}$ can be interpreted as the normal to the "disparity discontinuity contour" and a non-zero $M_{\mathbf{x}}$ indicates the direction of displacement along this normal. This direction is the normal direction if $M_{\mathbf{x}} = 1$, it is the opposite direction if $M_{\mathbf{x}} = -1$. If $\mathbf{x} \in \mathcal{S}$ does not correspond to a disparity discontinuity location, then $M_{\mathbf{x}}$ is assigned to 0. In what follows, we will write $\mathbf{M} = \{M_{\mathbf{x}}, \mathbf{x} \in \mathcal{S}\}$ with $\mathbf{M} \in \mathcal{M} = \{-1, 0, 1\}^{p \times q}$ and $\mathbf{E} = \{E_{\mathbf{x}}, \mathbf{x} \in \mathcal{S}\}$ with $\mathbf{E} \in \mathcal{E} = \mathcal{U}^{p \times q}$.

To explicitly take into account the fact that disparities and object boundaries are related, we propose to define a joint probabilistic model, namely $p(\mathbf{d}, \mathbf{a}|\mathbf{I})$. This is nothing but the probability of a disparity and displacement fields given the observed images. Note that we use $\mathbf{d}$ and $\mathbf{a} = (\mathbf{m}, \mathbf{e})$ to denote specific realizations of the random fields $\mathbf{D}$ and $\mathbf{A} = (\mathbf{M}, \mathbf{E})$ respectively. Our goal is to estimate the realizations of $\mathbf{d}$ and $\mathbf{a}$ that are consistent with our joint probabilistic model and the observed data $\mathbf{I}$. Ideally we are interested in finding the MAP (Maximum A Posteriori) estimates of $\mathbf{D}$ and $\mathbf{A}$,

$$(\mathbf{d}^{MAP}, \mathbf{a}^{MAP}) = \underset{\mathbf{d},\mathbf{a} \in \mathcal{D} \times \mathcal{M} \times \mathcal{E}}{\arg\max}\ p(\mathbf{d}, \mathbf{a}|\mathbf{I}). \tag{4.1}$$

Using the alternation maximization procedure described in chapter 3, the above equation can be estimated using two conditional distributions for $\mathbf{d}$ and $\mathbf{a}$ as follows, at any given iteration i:

$$\mathbf{d}^{(i+1)} = \underset{\mathbf{d} \in \mathcal{D}}{\arg\max}\ p(\mathbf{d}|\mathbf{a}^{(i)}, \mathbf{I}) \tag{4.2}$$

$$\mathbf{a}^{(i+1)} = \underset{\mathbf{a} \in \mathcal{M} \times \mathcal{E}}{\arg\max}\ p(\mathbf{a}|\mathbf{d}^{(i+1)}, \mathbf{I}) \tag{4.3}$$

Our task now reduces to defining the conditional distributions $p(\mathbf{d}|\mathbf{a}, \mathbf{I})$ and $p(\mathbf{a}|\mathbf{d}, \mathbf{I})$. The figure 4.1, in fact illustrates the mechanisms of the two equations above. It is also worth noting that, dealing with two such conditional distributions to account for cooperation mechanisms between $\mathbf{D}$ and $\mathbf{A}$ is easier and more tractable than trying to define directly a single joint distribution.

4.2.1 Displacement conditional disparity model

We first specify the disparity distribution conditional on the displacement field and the observed data, $p(\mathbf{d}|\mathbf{a}, \mathbf{I})$. The model is expressed as a Markov random field with an energy

function consisting of two terms. The first one corresponds to a data dependent term and the other to a regularizing or interaction term. The data term is similar to those in classical disparity-MRF models (Felzenszwalb and Huttenlocher [2006]), whereas the interaction term is modified to incorporate the displacement information. In a standard Markov random field modelling, we have to consider a neighbourhood system $\mathcal{N} = \{\mathcal{N}_{\mathbf{x}}, \mathbf{x} \in \mathcal{S}\}$ where $\mathcal{N}_{\mathbf{x}}$ is the set of neighbours of $\mathbf{x}$ and for which the reciprocity condition (4.4) below is satisfied,

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{S}, \qquad \mathbf{x} \in \mathcal{N}_{\mathbf{y}} \Leftrightarrow \mathbf{y} \in \mathcal{N}_{\mathbf{x}}. \tag{4.4}$$

In particular, we consider $\mathcal{N}_{\mathbf{x}}$ as the set of the eight nearest pixels of pixel $\mathbf{x}$. We define, $\forall \mathbf{a} \in \mathcal{M} \times \mathcal{E}$, the distribution $p(\mathbf{d}|\mathbf{a}, \mathbf{I})$ as a MRF on $\mathbf{d}$ with the following distribution, $\forall \mathbf{d} \in \mathcal{D}$,

$$p(\mathbf{d}|\mathbf{a}, \mathbf{I}) \propto \exp\left(-\sum_{\mathbf{x}\in\mathcal{S}} U_d(d_{\mathbf{x}}, \mathbf{I}) - \beta_d \sum_{\mathbf{x}\in\mathcal{S}} \sum_{\mathbf{y}\in\mathcal{N}_{\mathbf{x}}} V_d(d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{a}) \right) \tag{4.5}$$

where U_d and V_d are the data and the interaction terms respectively. The parameter β_d is an interaction parameter which allows to balance the effects of the data and interaction terms.

Data term.

The first term U_d in (4.5), referred to as the data term, assigns a cost for each disparity value chosen at location $\mathbf{x}$ based on the intensity difference between the left and the right images. However to account for noise, U_d is formulated as a truncated linear function (see chapter 2) depending on two scalar parameters λ_1 and T_1. $\mathbf{I}_L(\mathbf{x})$ (respectively $\mathbf{I}_R(\mathbf{x})$) denotes the pixel value at location $\mathbf{x}$ in the left (resp. right) image:

$$U_d(d_{\mathbf{x}}, \mathbf{I}) = \min\left(\lambda_1 \left|\mathbf{I}_L(\mathbf{x}) - \mathbf{I}_R(\mathbf{x}')\right|, T_1\right) \tag{4.6}$$

for pixel positions $\mathbf{x} = (x, y)$ and $\mathbf{x}' = (x, y + d_{\mathbf{x}})$, with candidate disparity $d_{\mathbf{x}} \in \mathcal{L}$.

Interaction term.

The second term V_d in (4.5) is called the interaction term. This term defines how the disparity at a location is influenced by its neighbours. In usual disparity-MRF models, all the neighbours interact and their influence depends on their disparity values. In the displacement conditional model, this influence depends also on the displacement field $\mathbf{A}$. For a given pixel location $\mathbf{x}$, only some of the neighbours in $\mathcal{N}_{\mathbf{x}}$ actually interact with $\mathbf{x}$, depending on the displacement values. This is because some locations in $\mathbf{y} \in \mathcal{N}_{\mathbf{x}}$ that may correspond to disparity discontinuities which may have to be corrected depending on the value of displacement $(m_{\mathbf{y}}, e_{\mathbf{y}})$ at that location as shown in figure 4.2. We therefore, model this function in such a way that only neighbours $\mathbf{y} \in \mathcal{N}_{\mathbf{x}}$ which do not require any correction

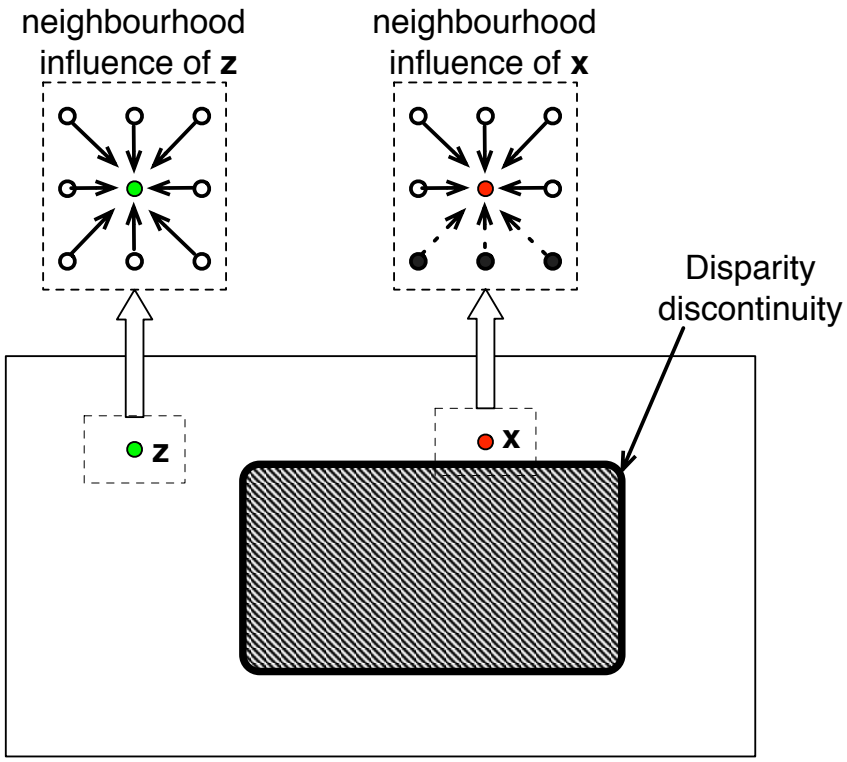


Figure 4.2: If a point $\mathbf{z}$ (shown in green) *does not lie near* any disparity discontinuity then all the neighbours interact. The interacting neighbours are shown as empty circles, with arrows indicating their influence. On the other hand, for a point $\mathbf{x}$ (shown in red) which *lies near* a disparity discontinuity (shown by thick black line) some of the neighbours *may not* interact. The neighbours which constitute disparity discontinuities are shown as solid circles, with the dashed arrows representing that their influence depends on the value of $(m_{\mathbf{y}}, e_{\mathbf{y}})$ for $\mathbf{y} \in \mathcal{N}_{\mathbf{x}}$.

influence the pixel $\mathbf{x}$. An intermediate *Active neighbourhood* field $\mathbf{H}(\mathbf{A}) = \{\mathbf{H}_{\mathbf{x}}(\mathbf{A}), \mathbf{x} \in \mathcal{S}\}$ is then built to encode this specificity. For each pixel location $\mathbf{x}$, the active neighbourhood $\mathbf{H}_{\mathbf{x}}(\mathbf{A})$ denotes the set of locations in $\mathcal{N}_{\mathbf{x}}$ that interact with $\mathbf{x}$. For a given realization $\mathbf{a} = (\mathbf{m}, \mathbf{e})$ of field $\mathbf{A} = (\mathbf{M}, \mathbf{E})$, it is defined as follows. Let $\mathcal{N}_{\mathbf{x}}^0$ denote the neighbours of $\mathbf{x}$ for which the displacement field is 0, *i.e.* $\mathcal{N}_{\mathbf{x}}^0 = \{\mathbf{y} \in \mathcal{N}_{\mathbf{x}}, m_{\mathbf{y}} = 0\}$. Then, define: $\forall \mathbf{a} \in \mathcal{M} \times \mathcal{E}$,

$$\mathbf{H}_{\mathbf{x}}(\mathbf{a}) = \begin{cases} \mathcal{N}_{\mathbf{x}}^0 & \text{if } m_{\mathbf{x}} = 0 \\ \left\{ \lfloor \mathbf{x} - m_{\mathbf{x}} \mathbf{e}_{\mathbf{x}} \rfloor \right\} \cap \mathcal{N}_{\mathbf{x}}^0 & \text{if } |m_{\mathbf{x}}| = 1 \end{cases} \tag{4.7}$$

where $\mathbf{e}_\mathbf{x}$ is the unit vector at location $\mathbf{x}$ when $\mathbf{E} = \mathbf{e}$. Note that $\lfloor \mathbf{x} - m_\mathbf{x}\mathbf{e}_\mathbf{x} \rfloor$ denotes the closest point in $\mathcal{S}$, when $\mathbf{x} - m_\mathbf{x}\mathbf{e}_\mathbf{x}$ does not belong to the grid. In particular, $\mathbf{H}_\mathbf{x}(\mathbf{a})$ can be the empty set if $|m_\mathbf{x}| = 1$ and $|m_{\lfloor \mathbf{x} - m_\mathbf{x}\mathbf{e}_\mathbf{x} \rfloor}| = 1$. The equation above shows there are three possible scenarios for the neighbourhood:

Case I: The first case in (4.7) can be interpreted as follows. If the displacement value $m_\mathbf{x}$ is zero at location $\mathbf{x}$, only those neighbours that have displacement values of zero will interact with $\mathbf{x}$. Intuitively, a zero displacement value means that at this location the disparity value agrees with the current object boundary estimation, and is reliable in that sense. In terms of active neighbourhoods, the case corresponds to $\mathbf{x} \in \mathbf{H}_\mathbf{y}(\mathbf{a})$ and $\mathbf{y} \in \mathbf{H}_\mathbf{x}(\mathbf{a})$, if $\mathbf{y} \in \mathcal{N}^0_\mathbf{x}$. This means that, when there is no evidence that a displacement is required at a given location, then the only *reliable* neighbours are those which themselves do not require a displacement (refer to figure 4.3, case I). For this case, the interaction function is defined as a truncated linear:

$$V_d(d_\mathbf{x}, d_\mathbf{y}, \mathbf{a}) = \min\Big(\lambda_2 |d_\mathbf{x} - d_\mathbf{y}|, T_2\Big) \tag{4.8}$$

where λ_2 and T_2 are scalar parameters playing a role similar to λ_1 and T_1 in (4.6).

Case II: If the displacement value is 1, i.e., $|m_\mathbf{x}| = 1$, then it indicates that disparity at this location has to be changed in order to better agree with object boundaries. Moreover, it indicates what should be the value of this new disparity value. It follows that in the second line of equality (4.7), only one neighbour at most should interact with $\mathbf{x}$. This neighbour, whose location is $\mathbf{y} = \mathbf{x} - m_\mathbf{x}\mathbf{e}_\mathbf{x}$, corresponds to the one which is located in the opposite direction to the normal $\mathbf{e}_\mathbf{x}$ located at $\mathbf{x}$. This is depicted pictorially in figure 4.3 case II. In such a case the interaction function enforces the disparity at $\mathbf{x}$ to be replaced by that of $\mathbf{y}$. That is:

$$V_d(d_\mathbf{x}, d_\mathbf{y}, \mathbf{a}) = T_2\Big(1 - \mathbb{1}_{\{D_\mathbf{x} = D_\mathbf{y}\}}(d_\mathbf{x}, d_\mathbf{y})\Big) \tag{4.9}$$

where T_2 is scalar same as the one in (4.8). The function $\mathbb{1}_{\{D_\mathbf{x} = D_\mathbf{y}\}}(d_\mathbf{x}, d_\mathbf{y})$ is an indicator function which takes the value 1 when $d_\mathbf{x} = d_\mathbf{y}$ and 0 otherwise. This indicator function enforces that the disparity $d_\mathbf{x}$ be the same as $d_\mathbf{y}$. This function stems from the intuition that the disparity at a location $\mathbf{x}$ where $m_\mathbf{x}$ is non-zero, could be corrected by replacing its value by disparity of its neighbour $d_\mathbf{y}$ in the opposite direction $-m_\mathbf{x}$, i.e., $\mathbf{y} = \mathbf{x} - m_\mathbf{x}\mathbf{e}_\mathbf{x}$.

Case III: The last case may happen when the displacement value $|m_\mathbf{x}| = 1$ and that of the selected active neighbour has itself been assigned a non zero displacement value. This means that the neighbour's corresponding disparity value is likely to change and therefor should not be used (see figure 4.3, case III). In that case, the set $\mathbf{H}_\mathbf{x}(\mathbf{a})$ in the second line of (4.7) is empty and the only reliable evidence is the data. The interaction potential is set to 0:

$$V_d(d_\mathbf{x}, d_\mathbf{y}, \mathbf{a}) = 0. \tag{4.10}$$

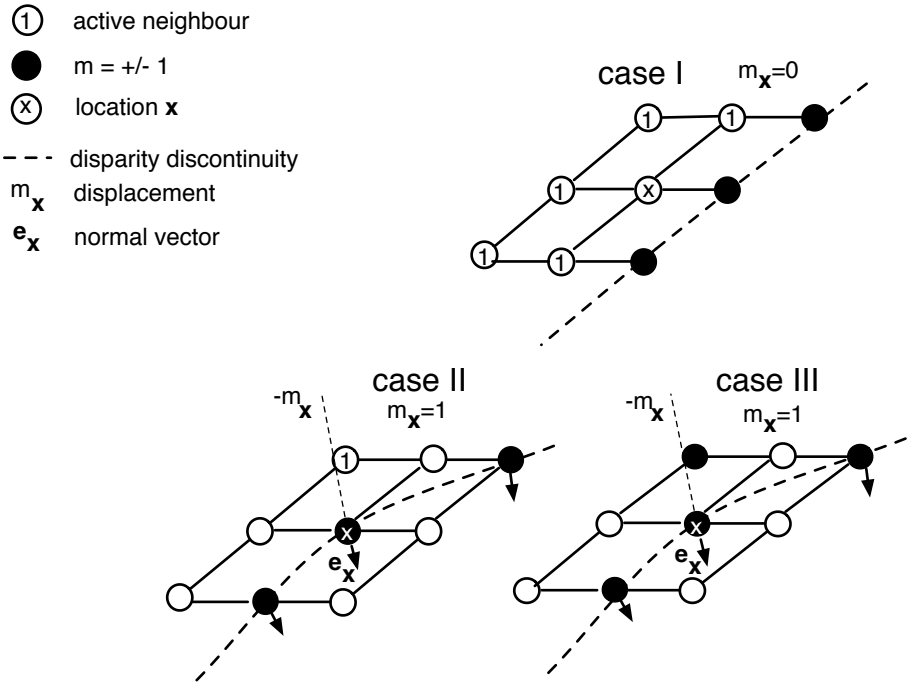


Figure 4.3: The figure shows how the neighbourhood is activated based on $\mathbf{A} = (\mathbf{M}, \mathbf{E})$. Active neighbours are represented as circles with 1 written inside. Inactive neighbours are shown as solid circles and the disparity discontinuity is indicated by dashed lines. Case I: When $m_{\mathbf{x}} = 0$, all the neighbours with $m_{\mathbf{y}} = 0$, $\forall\, \mathbf{y} \ \in \mathcal{N}_{\mathbf{x}}^{0}$, contribute to the interaction term. Case II: When $|m_{\mathbf{x}}| = 1$ the neighbour $\mathbf{y}$ in the opposite direction, $-m_{\mathbf{x}}$, is considered if $m_{\mathbf{y}} = 0$. Case III: When $|m_{\mathbf{x}}| = 1$ and the neighbour $\mathbf{y}$ in the opposite direction, $-m_{\mathbf{x}}$, is $|m_{\mathbf{y}}| = 1$, the interaction term is set to 0.

From the above discussion we can now write the interaction function V_d, using equations (4.8), (4.9) and (4.10). As we can see, V_d involves the displacement field $\mathbf{A}$ through the active neighbourhood field $\mathbf{H}(\mathbf{A})$, for all $\mathbf{a} \in \mathcal{M} \times \mathcal{E}$, and $d_{\mathbf{x}}, d_{\mathbf{y}} \in \mathcal{L}^2$ as follows,

$$V_d(d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{a}) = \begin{cases} \min(\lambda_2 |d_{\mathbf{x}} - d_{\mathbf{y}}|, T_2) & \text{if } \mathbf{x} \in \mathbf{H}_{\mathbf{y}}(\mathbf{a}) \ \& \ \mathbf{y} \in \mathbf{H}_{\mathbf{x}}(\mathbf{a}) \\ 0 & \text{if } \mathbf{x} \notin \mathbf{H}_{\mathbf{y}}(\mathbf{a}) \ \& \ \mathbf{y} \notin \mathbf{H}_{\mathbf{x}}(\mathbf{a}) \\ T_2\big(1 - \mathbb{1}_{\{D_{\mathbf{x}} = D_{\mathbf{y}}\}}(d_{\mathbf{x}}, d_{\mathbf{y}})\big) & \text{otherwise} \end{cases} \tag{4.11}$$

The corresponding MRF is then defined over the standard eight-neighbourhood system $\mathcal{N}$, although it behaves as one built on the *active neighbourhood* system given by $\mathbf{H}(\mathbf{a})$. The definition in (4.11) has the advantage that it allows all active neighbourhood systems, even those not satisfying the reciprocity condition (4.4). This idea of building an additional random field to deal with sets of active neighbours is similar to that in Le Hégarat-Mascle et al. [2007]. While they consider non-stationary neighbourhoods, Le Hégarat-Mascle et al. defined a neighbourhood system that does not necessarily satisfy the reciprocity condition.

4.2.2 Disparity conditional displacement model

The distribution of displacement field $\mathbf{A}$, conditional on disparity $\mathbf{D}$ and image information $\mathbf{I}$, will now be modelled. This requires the following structures to be defined. It requires the definition of *discontinuity chains* representing the locations at which a disparity discontinuity occurs considering a current value $\mathbf{d}$ of the disparity map $\mathbf{D}$. Disparity discontinuity extraction is done by using the robust function defined in the first line of (4.11). If the difference between the disparity at $\mathbf{x}$ and the disparity at any one of the neighbouring locations lies above the threshold value T_2, $\mathbf{x}$ is retained as a location at which a disparity discontinuity occurs. This procedure gives a binary map of disparity discontinuity locations. This map is then converted into a set of discontinuity chains denoted by $\mathcal{C}(\mathbf{D})$. More specifically, the set $\mathcal{C}(\mathbf{D})$ is made of a number $T(\mathbf{D})$ of connected components,

$$\mathcal{C}(\mathbf{D}) = \{\mathcal{C}^1(\mathbf{D}), \ldots, \mathcal{C}^{T(\mathbf{D})}(\mathbf{D})\} \tag{4.12}$$

which can be of different sizes. The figure 4.4 shows a cartoon example of how the discontinuity chains set $\mathcal{C}(\mathbf{D})$ is formed. For $t = 1, \ldots, T(\mathbf{D})$, each set $\mathcal{C}^t(\mathbf{D})$ is referred to as a discontinuity chain but is itself made of two sets $\mathcal{C}^t(\mathbf{D}) = \{\mathcal{S}^t(\mathbf{D}), \mathcal{W}^t(\mathbf{D})\}$ where:

$$\mathcal{S}^t(\mathbf{D}) = \{\ \mathbf{x}_1^t, \ldots,\ \mathbf{x}_K^t\} \subset \mathcal{S} \tag{4.13}$$

is a set of K locations which are connected, namely for all $k = 2, \ldots K-1$, $\mathbf{x}_k^t$ is a neighbour of $\mathbf{x}_{k-1}^t$ and $\mathbf{x}_{k+1}^t$ while the first and last locations have only one neighbour. For simplicity, we use notation K for the size of the t^{th} chain omitting the possible dependency on t and $\mathbf{D}$. The set $\mathcal{W}^t = \{\mathbf{w}_1^t, \ldots,\ \mathbf{w}_K^t\}$ represents the normals to the chain associated to each location $\mathbf{x}_k^t$. More generally, using notation:

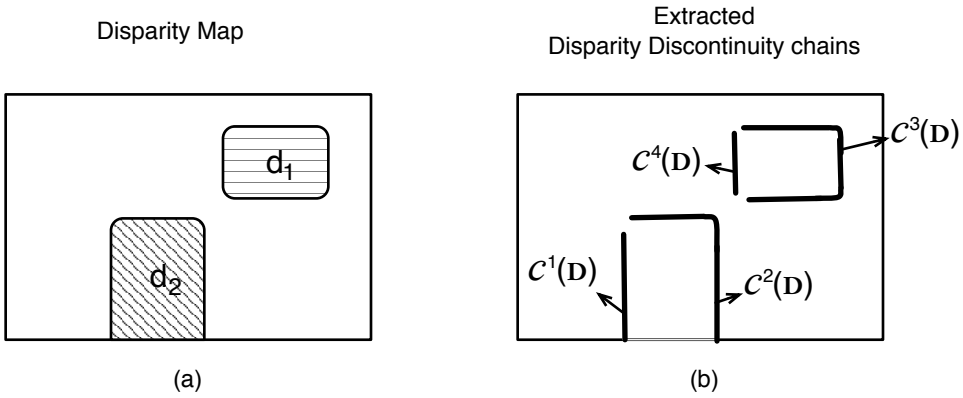


Figure 4.4: Cartoon example showing the formation of discontinuity chain set $\mathcal{C}(\mathbf{D})$. (a) represents a disparity map with two objects, one at disparity d_1 and the other d_2 (shaded regions). (b) shows the extracted disparity discontinuity chains $\mathcal{C}^t(\mathbf{D})$ of different lengths. In this example $\mathcal{C}(\mathbf{D}) = \{\mathcal{C}^1(\mathbf{D}), \mathcal{C}^2(\mathbf{D}), \mathcal{C}^3(\mathbf{D}), \mathcal{C}^{4(\mathbf{D})}(\mathbf{D})\}$.

$$\mathcal{S}(\mathbf{D}) = \cup_{t=1}^{T(\mathbf{D})} \mathcal{S}^t(\mathbf{D}) \tag{4.14}$$

and

$$\mathcal{W}(\mathbf{D}) = \cup_{t=1}^{T(\mathbf{D})} \mathcal{W}^t(\mathbf{D}), \tag{4.15}$$

for $\mathbf{x} \in \mathcal{S}(\mathbf{D})$, we will denote by $\mathbf{w_x} \in \mathcal{W}(\mathbf{D})$ the normal at location $\mathbf{x}$. Thus $\mathcal{C}^t(\mathbf{D})$ provides the positions and normals for all the points in the t^{th} chain. figure 4.5 illustrates the construction of the discontinuity chains.

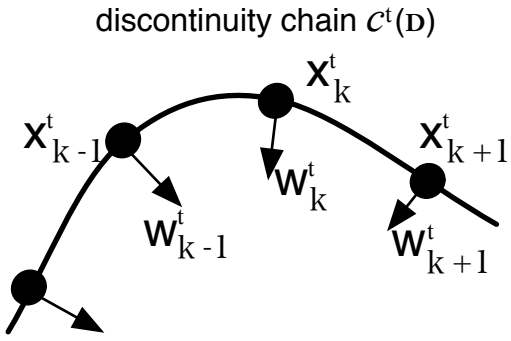


Figure 4.5: Illustration of discontinuity chain construction

The conditional distribution of field $\mathbf{A}$ given $\mathbf{D}$ and $\mathbf{I}$ is defined using the discontinuity chains $\mathcal{C}(\mathbf{D})$: $\forall \mathbf{a} = (\mathbf{m}, \mathbf{e}) \in \mathcal{M} \times \mathcal{E}, \mathbf{d} \in \mathcal{D}$,

$$p(\mathbf{a}|\mathbf{d}, \mathbf{I}) \propto \underbrace{\mathbb{1}_{\{E_{\mathbf{x}}=\mathbf{w}_{\mathbf{x}}, \forall \mathbf{x} \in \mathcal{S}(\mathbf{d})\}}(\mathbf{e})}_{\text{normals}} \underbrace{p(\mathbf{m}|\mathbf{d}, \mathbf{I})}_{\text{displacement}} \tag{4.16}$$

In (4.16), the first term (called *normals* in the equation) indicates that, at the discontinuity locations, the displacement field normals are the same as the discontinuity chains normals, with probability one conditionally to $\mathbf{D} = \mathbf{d}$. Once the normals are fixed, we are now interested in finding the direction along this normal that the chains should be moved in order to align with the object boundary. The probability distribution $p(\mathbf{m}|\mathbf{d}, \mathbf{I})$ in (4.16) encodes exactly this information. As the corrections need to be applied only at the disparity discontinuities and not on the entire grid $\mathcal{S}$, the distribution $p(\mathbf{m}|\mathbf{d}, \mathbf{I})$ is defined as follows:

$$p(\mathbf{m}|\mathbf{d}, \mathbf{I}) \propto \mathbb{1}_{\{M_{\mathbf{x}}=0, \forall\, \mathbf{x} \in \overline{\mathcal{S}}(\mathbf{d})\}}(\mathbf{m}) \prod_{t=1}^{T(\mathbf{d})} p\big(\mathbf{m}^t|\mathcal{C}^t(\mathbf{d}), \mathbf{I}\big) \tag{4.17}$$

where $\overline{\mathcal{S}}(\mathbf{d})$ is the complement of set $\mathcal{S}(\mathbf{d})$ and $\mathbf{m}^t = \{m_{\mathbf{x}}|\mathbf{x} \in \mathcal{S}^t(\mathbf{d})\}$. However, for the sake of clarity, the displacement at location $\mathbf{x}_k^t$ is denoted by m_k^t such that $\mathbf{m}^t = \{m_1^t, \ldots, m_K^t\}$.

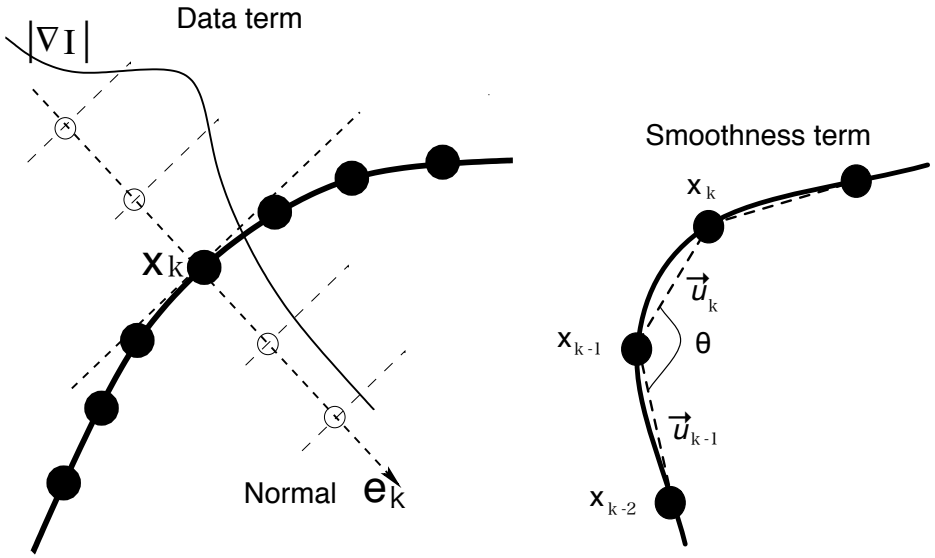
The first term in (4.17) ensures that the non-zero displacements can occur only at discontinuity locations with probability one. The second term in the right-hand-side of (4.17), is the product of probabilities defined on the discontinuity chains. These discontinuity chains are assumed to be independent and therefore the probability distribution of each of these chains can be individually expressed as a second order Markov chain as follows: $\forall \mathbf{m}^t \in \{-1, 0, 1\}^K, \mathbf{d} \in \mathcal{D}$,

$$p(\mathbf{m}^t|\mathcal{C}^t(\mathbf{d}), \mathbf{I}) = \prod_{k=3}^{K} p\Big(m_k^t|m_{k-1}^t, m_{k-2}^t, \mathcal{C}^t(\mathbf{d}), \mathbf{I}\Big) P\Big(m_1^t, m_2^t|\mathcal{C}^t(\mathbf{d}), \mathbf{I}\Big) \tag{4.18}$$

Until now we have shown how the displacement field is reduced to a Markov chain. We will now describe the chain distributions for each of these chains. As each of the chain distributions will be defined in a similar manner, from now on we drop the superscript t. The first terms in the right-hand-side of (4.18) are defined using a data term and an interaction term as specified below,

$$p(m_k|m_{k-1}, m_{k-2}, \mathcal{C}(\mathbf{d}), \mathbf{I}) \propto \exp\Big(- U_c\big(m_k, \mathcal{C}(\mathbf{d}), |\nabla \mathbf{I}_L|\big) - \beta_c\, V_c\big(m_k, m_{k-1}, m_{k-2}, \mathcal{C}(\mathbf{d})\big)\Big) \tag{4.19}$$

where β_c is an interaction parameter acting as a weight between the two terms U_c and V_c. The data term U_c tries to move the chains towards the highest gradients in the image, whereas the interaction terms V_c, enforces the chains to be smooth. These two terms U_c and V_c are described in detail in what follows.



(a) Data term (b) Interaction term

Figure 4.6: The figure 4.6(a) indicates how the data term favours a location in direction of gradient maximum. The figure 4.6(b) shows the vectors $\vec{u}_{k-1}$ and $\vec{u}_k$. The interaction term assigns a score based on the angle between the vectors $\vec{u}_{k-1}$ and $\vec{u}_k$.

Data term.

The data term U_c associates a cost for moving the point $\mathbf{x}_k$ on either side of the normal $\mathbf{w}_k$. In order to determine this cost, we first choose points on both sides of the normal using a range of locations distant from $-\epsilon$ to ϵ where ϵ is an integer value to be fixed (in the figure 4.6(a) $\epsilon = 2$). Then, we determine the difference in the gradient magnitude between the current position and the points chosen along the normal. If this difference is strictly negative, it means that moving the current chain point $\mathbf{x}_k$ to that position along the normal leads it a higher gradient position. Hence, the direction of motion towards this position is favoured. For example, in the figure 4.6(a) this positions of higher gradient are $\mathbf{x}_k + \epsilon\,\mathbf{w}_k$ where $\epsilon = -1$ or -2 and therefore the preferred direction of motion is negative i.e., $m_k = -1$ is preferred. This can be written as,

$$U_c\Big(m_k, \mathcal{C}(\mathbf{d}), \big|\nabla \mathbf{I}_L\big|\Big) \;\; = \;\; 1 - 2\,\mathbb{1}_{\left\{M_k = s_k\left(|\nabla \mathbf{I}_L|, \mathbf{w}_k\right)\right\}}(m_k) \tag{4.20}$$

where

$$s_k\left(\big|\nabla \mathbf{I}_L\big|, \mathbf{w}_k\right) = \operatorname{sgn}\left(\arg\min_{\ell \in [-\epsilon, \epsilon]} \left(|\nabla \mathbf{I}_L(\mathbf{x}_k)| - \big|\nabla \mathbf{I}_L(\mathbf{y}_\ell)\big|\right)\right) \tag{4.21}$$

where $\mathbf{y}_\ell = \mathbf{x}_k + \ell\,\mathbf{w}_k$ and $|\nabla \mathbf{I}_L|$ is the gradient magnitude in the reference image $\mathbf{I}_L$. Also, sgn denotes the function that is 1 when its argument is strictly positive, -1 when it is

strictly negative and 0 otherwise. The last term in (4.18) is evaluated based only on the data term.

Interaction term

The interaction term enforces smoothness by favouring those m-values which make the angle between the vectors defined by the positions $(\mathbf{x}_{k-2}, \mathbf{x}_{k-1})$ and $(\mathbf{x}_{k-1}, \mathbf{x}_k)$ as close to zero as possible. The position vectors defined by $\mathbf{x}_{k-2}, \mathbf{x}_{k-1}, \mathbf{x}_k \in \mathcal{C}(\mathbf{d})$ are denoted by $\vec{u}_{k-1}$ and $\vec{u}_k$ as shown in figure 4.6(b) and expressed as follows:

$$\vec{u}_{k-1} = (\mathbf{x}_{k-1} + m_{k-1}\mathbf{w}_{k-1}) - (\mathbf{x}_{k-2} + m_{k-2}\mathbf{w}_{k-2}) \tag{4.22}$$

$$\vec{u}_k = (\mathbf{x}_k + m_k\mathbf{w}_k) - (\mathbf{x}_{k-1} + m_{k-1}\mathbf{w}_{k-1}) \; . \tag{4.23}$$

The V_c term therefore assigns a score based on the angle between vectors $\vec{u}$ and $\vec{v}$ and is defined as,

$$V_c\Big(m_k, m_{k-1}, m_{k-2}, \mathcal{C}(\mathbf{d})\Big) = -\frac{\langle \vec{u}_{k-1}, \vec{u}_k \rangle}{|\vec{u}_{k-1}||\vec{u}_k|} \tag{4.24}$$

where $\langle \; . \; , \; . \; \rangle$ denotes the scalar product (see figure 4.6(b)). Note that (4.19) defined a second-order Markov chain which can be easily turned into a first-order Markov chain, so that optimal displacement values can be found using the standard Viterbi algorithm.

Furthermore, we consider an additional heuristic that prevents any discontinuity chain point already at a gradient peak to move from its position. This ensures that the contour chains do not move inside the object boundaries. This could happen as there may be positions with higher gradient within the object depending how textured the image is. As termination of the overall algorithm (Alternation Maximization, section 4.4) is determined by the number of moving points, i.e., non-zero m-values, this heuristic aids in attaining faster convergence. The displacement labels obtained at all chain positions are then embedded back in to image. This means that every pixel location is given a displacement label of zero except at the location where the discontinuity chain was formed. The pixel locations now consist of the following information: direction of movement, which is the normal to the discontinuity-chain at that point and magnitude set to 1 or 0, in order to curb fast movement to wrong gradient maxima. This completes the description our coupled-MRF model for the disparity and displacement fields. We have shown how each field ($\mathbf{D}$ or $\mathbf{A}$) uses the information from the other within their model. In the next section, we propose an optimization scheme to find estimates of the unknown fields $\mathbf{D}$ and $\mathbf{A}$ which are consistent with the observed image set $\mathbf{I}$.

4.3 Optimization

From discussion in section (4.2), we use directly the conditional distributions $p(\mathbf{d}|\mathbf{a}, \mathbf{I})$ and $p(\mathbf{a}|\mathbf{d}, \mathbf{I})$ (equations (4.2) and (4.3)) defined in the previous section, within our alter-

nating maximization framework. Also as discussed in section 4.2.1 the distribution involved in (4.2) is the MRF model defined by (4.5). For this distribution a direct optimization is intractable. Maximization in (4.2) can therefore performed approximately using algorithms such as Mean Field or Belief Propagation (BP).

As discussed in chapter 2 section 2.3.1, in order to compute the conditional in $p(\mathbf{d}|\mathbf{a},\mathbf{I})$ in a tractable manner using Mean Field, it is approximated by a simpler, factorisable distribution $Q(\mathbf{d}) = \prod_{\mathbf{x}\in\mathcal{S}} Q_{\mathbf{x}}(d_{\mathbf{x}})$. The task then is to find $Q(\mathbf{d})$, which is as close as possible to the true posterior in (4.5), where the distance between both distributions is measured by the Kullback-Leibler (KL) divergence. The minimization of the KL divergence with respect marginals $Q_{\mathbf{x}}(d_{\mathbf{x}})$ is done iteratively. This leads to the following set of update equations defined for all $\mathbf{x} \in \mathcal{S}$, also referred to as Mean Field equations:

$$Q_{\mathbf{x}}(d_{\mathbf{x}}) \leftarrow \frac{1}{Z} \exp\Bigg(-\Big(U_d(d_{\mathbf{x}},\mathbf{I}) + \sum_{\mathbf{y}\in\mathcal{N}_{\mathbf{x}}} \sum_{d_{\mathbf{y}}\in\mathcal{L}} Q_{\mathbf{y}}(d_{\mathbf{y}}) V_d(d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{a})\Big)\Bigg) \tag{4.25}$$

where Z is the normalizing constant.

If the same optimization is performed using BP instead of Mean Field, then as in chapter 2 section 2.3.2, the beliefs at every position $\mathbf{x}$ is computed as follows:

$$Q_{\mathbf{x}}(d_{\mathbf{x}}) = \frac{1}{Z} \exp\Big(- U_d(d_{\mathbf{x}},\mathbf{I})\Big) \prod_{\mathbf{y}\in\mathcal{N}_{\mathbf{x}}} m_{\mathbf{y},\mathbf{x}}(d_{\mathbf{x}}) \tag{4.26}$$

where Z is the normalizing constant. Note that we use the same notation $Q_{\mathbf{x}}(d_{\mathbf{x}})$ as in (4.25) to denote the beliefs in the equation above. The messages m are updated iteratively by taking into account the data term, the interaction term and all the messages that are coming into $\mathbf{x}$, except the one from $\mathbf{y}$, as follows:

$$m_{\mathbf{x},\mathbf{y}}(d_{\mathbf{y}}) \leftarrow \max_{d_{\mathbf{x}}\in\mathcal{L}} \exp\Big(- U_d(d_{\mathbf{x}},\mathbf{I})\Big) \exp\Big(- V_d(d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{a})\Big) \prod_{\mathbf{z}\in\mathcal{N}_{\mathbf{x}}\setminus\mathbf{y}} m_{\mathbf{z},\mathbf{x}}(d_{\mathbf{x}}) \tag{4.27}$$

For both optimization it is to be noted that the affect of displacement $\mathbf{a}$ is seen in the neighbourhood system $\mathcal{N}$ that is built on the active neighbourhood system given by $\mathbf{H}(\mathbf{a})$. The disparity value at every position can then be found as follows:

$$d^*_{\mathbf{x}} = \arg\max_{d_{\mathbf{x}}\in\mathcal{L}} Q_{\mathbf{x}}(d_{\mathbf{x}}) \tag{4.28}$$

Alternatively, we could compute the expected disparity as:

$$d^*_{\mathbf{x}} = \sum_{d_{\mathbf{x}}\in\mathcal{L}} d_{\mathbf{x}}\, Q_{\mathbf{x}}(d_{\mathbf{x}}) \tag{4.29}$$

The equation (4.28) and (4.29) are equivalent, if $Q_{\mathbf{x}}(d_{\mathbf{x}})$ has a single symmetric isolated mode at $d^*_{\mathbf{x}}$. In all our experiments we have used (4.28) to compute the disparity values at

every position. While the optimization for the disparity MRF in (4.5) is an approximate one, the one for finding displacement values in (4.16) is exact. This is because during our construction (section 4.2.2) we used discontinuity Markov chains instead of the 2D field. The exact inference of this Markov chain is done using Viterbi algorithm. We now provide brief discussion implementation of the Viterbi algorithm in our case.

4.3.1 Viterbi algorithm

As described in section 4.2.2, the Markov chain constructed to find the displacement field values is of the second order. There are two ways to infer the values for this chain; one is to use a second-order Viterbi algorithm (Cutting et al. [1992]); the second is to convert the chain into a first order and use standard the Viterbi algorithm proposed by Rabiner [1989]. Inference in our case is done using the latter approach, as the Markov chain in our case can be easily converted to a first order one.

We convert the second order model in nodes into a first order model in difference-vectors between the nodes. We see from (4.24) that the interaction term $V_c\left(m_k, m_{k-1}, m_{k-2}, \mathcal{C}(\mathbf{d})\right)$, even though dependent on m_k, m_{k-1} and m_{k-2}, can be seen as first order chain in $\vec{u}_k$ and $\vec{u}_{k-1}$. Therefore, without any further modification the we can now use the standard Viterbi algorithm to find the displacement values. The Viterbi algorithm finds the optimal sequence,

$$M^* = \{m_1^*, m_2^*, m_3^*, \ldots, m_k^* \ldots m_{K-1}^*, m_K^*\}, \tag{4.30}$$

given the observations disparity $\mathbf{d}$ and image $\mathbf{I}$. For a given discontinuity chain $\mathcal{C}(\mathbf{d})$, we have the sequence $M = \{m_1, m_2, m_3, \ldots, m_k \ldots m_{K-1}, m_K\}$, where $m_k \in \{-1, 0, 1\}\ \forall\, k = 1, 2, \ldots, K$. The goal is to find optimal sequence M^*, also referred to as optimal *path*, which maximizes the probability in the equation (4.19). In order to do so we define for every position k

$$\delta_k(m) = \max_{m_1, m_2, m_3, \ldots, m_{k-1}} p(m_1, m_2, m_3, \ldots, m_k = m \,|\, \mathcal{C}(\mathbf{d}), \mathbf{I}) \tag{4.31}$$

where $m = \{-1, 0, 1\}$ and $\delta_k(m)$ is the best score along a single path at position k. By induction, for $m, n \in \{-1, 0, 1\}$

$$\delta_{k+1}(n) = \max_m \Big(\delta_k(m) p(m_{k+1} = n | m_k = m, m_{k-1}, \mathcal{C}(\mathbf{d}), \mathbf{I}) \Big) \tag{4.32}$$

To retrieve the sequence M^* we have to keep track of the argument of the recursive equation above for each k and n. This is done by introducing an array $\psi_k(n)$. Here it is to be noted that we already know m_1 and m_2, as they are determined using only the evidence $\mathbf{I}$ and $\mathcal{C}(\mathbf{d})$. Therefore, the complete procedure is then to determine the path $m_3^*, \ldots, m_K^*$ given the disparity and image information $p(\mathbf{m}|\mathcal{C}(\mathbf{d}), \mathbf{I})$ is as follows:

- Initialization, we start with m_3:

$$\delta_3(m) = P\big(m_1, m_2 \big| \mathcal{C}(\mathbf{d}), \mathbf{I}\big)\, S\big(m, m_2, \mathcal{C}(\mathbf{d}), \nabla|\mathbf{I}_L|\big) \quad \text{where} \tag{4.33}$$

$$S\big(m, m_2, \mathcal{C}(\mathbf{d}), |\nabla\mathbf{I}_L|\big) = \exp\Big(- U_c\big(m, \mathcal{C}(\mathbf{d}), |\nabla\mathbf{I}_L|\big) - \beta_c\, V_c\big(m, m_1, m_2, \mathcal{C}(\mathbf{d})\big)\Big),$$

$$\psi_3(m) = 0; \tag{4.34}$$

  where $m \in \{-1, 0, 1\}$.

- Recursion:

$$\delta_k(n) = \max_{m\in\{-1,0,1\}} \left(\delta_{k-1}(m)\, S\Big(n, m, \mathcal{C}(\mathbf{d}), \nabla|\mathbf{I}_L|\Big) \right) \tag{4.35}$$

$$\psi_k(n) = \arg\max_{m\in\{-1,0,1\}} \left(\delta_{k-1}(m)\, S\Big(n, m, \mathcal{C}(\mathbf{d}), \nabla|\mathbf{I}_L|\Big) \right) \tag{4.36}$$

  where $4 \leq k \leq K$ and $n \in \{-1, 0, 1\}$.

- Termination:

$$\Delta = \max_{m\in\{-1,0,1\}} \Big(\delta_K(m) \Big) \tag{4.37}$$

$$m_K^* = \arg\max_{m\in\{-1,0,1\}} \Big(\delta_K(m) \Big) \tag{4.38}$$

  – The optimal path M^* is determined using back tracking:

$$m_k^* = \psi_{k+1}(m_{k+1}^*) \tag{4.39}$$

The above procedure is carried for all the discontinuity chains. As mentioned before, these displacement labels are embedded back into image. The optimization for both the disparity and for the discontinuity chains is carried out at different scales. We use the coarse-to-fine strategy proposed by Felzenszwalb and Huttenlocher [2006] to achieve this. We will now discuss this coarse to fine procedure in some detail before presenting the results of the proposed technique.

4.3.2 Multi-grid optimization

The coarse-to-fine approach presented by Felzenszwalb and Huttenlocher [2006] is applied to the disparity-MRF part of the framework and the boundary estimation is performed at each scale without modification to the its optimization. The optimization such as BP or Mean Field for disparity-MRF takes many iterations as information needs to flow over long distances in the grid. To circumvent this problem Felzenszwalb and Huttenlocher [2006]

present a multi-grid approach, which allows long range interactions to be captured by short paths in coarse grids.

The main idea is not to change the overall problem structure, that is leaving the graph structure and the energy function unchanged, but to use the hierarchy to initialize probabilities (messages in case of BP) at successively finer levels. The method to perform this is as follows: The optimization is carried out at one level of resolution and the probabilities at this level are used to initialize the next finer level. This way, at each level the probabilities move closer to the fixed point faster and thereby converging more rapidly.

We recall that we represent the 2D image grid as $\mathcal{S}$. The hierarchy of grids is represented as $\mathcal{S}^0, \mathcal{S}^1, \ldots, \mathcal{S}^l, \ldots$ where $\mathcal{S}^0 = \mathcal{S}$, i.e., full resolution and $\mathcal{S}^l$ corresponds to a block of $w \times w$ pixels of the original grid $\mathcal{S}$, where $w = 2^l$. At each level l, the image pixels in each $w \times w$ block are assigned the same disparity. Figure 4.7 illustrates this construction for two levels. The idea is to find the disparity map $\mathbf{d}^l$ at each level for the sites $\mathcal{S}^l$ minimizing the energy,

$$E(\mathbf{d}^l) = \sum_{(x,y)\in\mathcal{S}^l} U_d^l(d_{(x,y)}, \mathbf{I}) - \beta_d \sum_{(x,y)\in\mathcal{S}^l} V_d^l(d_{(x,y)}, d_{(x+1,y)}, \mathbf{a}) + \beta_d \sum_{(x,y)\in\mathcal{S}^l} V_d^l(d_{(x,y)}, d_{(x,y+1)}, \mathbf{a}) \tag{4.40}$$

This equation is the same as the energy within the exponential in (4.5), except that it is expressed in terms of the image grid represented at coarseness level l. The functions U_d^l and V_d^l are data and interaction cost at level l. As each block of $w \times w$ is assigned a single disparity, the data cost U_d^l is calculated as sum of data costs for the pixels within that block,

$$U_d^l(d_{(x,y)}, \mathbf{I}) = \sum_{u=0}^{w-1} \sum_{v=0}^{w-1} U_d(d_{(w\,x+u, w\,y+u)}, \mathbf{I}) \tag{4.41}$$

for all $d_{(w\,x+u,w\,y+u)} \in \mathcal{L}$. The interaction at each level is still retained in (4.11), except that instead of treating each pixel, each block is considered. Here it is important to specify that since the displacement chains are constructed at each level separately, the active neighbourhood can therefore be created at each level by considering each $w \times w$ block a one large pixel.

Once the probabilities are found using either BP or Mean Field at the coarsest level of the hierarchy, they are then used to initialize at next level and so on until the original resolution is reached. A key point in the overall optimization is that, at each level it is performed on the same set of disparity labels, but with different sized blocks of pixels. Furthermore, this approach differs from other multi-scale approaches which are commonly used in computer vision, such as the Gaussian pyramid of Burt and Adelson [1983], in that they are based on reducing the resolution of image. In our case the effect of reducing the image resolution is to rapidly decrease the number of distinguishable disparities from one level to another. For example, the disparities between 0 and 16 become indistinguishable by fourth level. In contrast, the method presented here only reduces the resolution at which the disparity labels are estimated, by accumulating the data cost over larger spatial neighbourhoods.

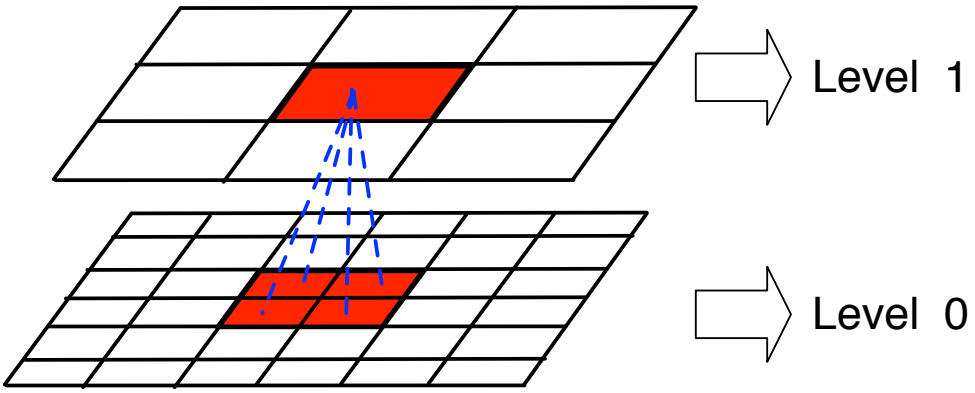


Figure 4.7: The two levels of the multi-grid method are illustrated. Each node at every level l corresponds to a 2×2 block of nodes in the level $l-1$.

The disparity and displacement estimation is carried out alternately at each level. While the disparity-MRF is modified as described above, the displacement values at the level l are found using the image at the resolution l and disparity $\mathbf{d}^l$. An important point is that the active neighbourhood $\mathbf{H}(\mathbf{a})$ is constructed at every level for the correction of disparities at discontinuities. However, the displacement calculation at each resolution is carried out independently of the other levels. In other words, unlike disparity-MRF, there is no initialization of displacement probabilities from lower resolution to higher resolution. The overall alternation procedure at each scale is explained in the following section.

4.4 Alternating Maximization procedure

The resulting alternation procedure for each scale is described below including the intermediate discontinuity chain extraction. At iteration $q = 0$ and at the coarsest scale, the displacement field values are assumed to be zero. The very first step (4.2) is performed using a Mean Field or BP approach to get a first estimate of the disparity map. Note that we drop the superscript l for each level to improve readability. The two following steps are carried out alternately at each scale l,

1. Update Displacement field $\mathbf{a}^{(q)}$ into $\mathbf{a}^{(q+1)} = (\mathbf{m}^{(q+1)}, \mathbf{e}^{(q+1)})$:

   (i) Extract disparity discontinuities as a binary map from current disparity map $\mathbf{d}^{(q)}$.

   (ii) Convert the binary map into a set of discontinuity chains $\mathcal{C}(\mathbf{d}^{(q)})$.

   (iii) At discontinuity locations, i.e., for $\mathbf{x} \in \mathcal{S}(\mathbf{d}^{(q)})$, set $\mathbf{e}_{\mathbf{x}}^{(q+1)}$ to the normal $\mathbf{w}_{\mathbf{x}}$ in $\mathcal{C}(\mathbf{d}^{(q)})$. For the remaining locations, $m_{\mathbf{x}}^{(q+1)}$ is set to zero so that $\mathbf{e}_{\mathbf{x}}^{(q+1)}$ can be set arbitrarily.

(iv) Estimate new m-values for each discontinuity chain $\mathcal{C}^t(\mathbf{d}^{(q)})$ in $\mathcal{C}(\mathbf{d}^{(q)})$ using the Viterbi algorithm for the Markov chain as defined in (4.18) and (4.19).

2. Update Disparity field $\mathbf{d}^{(q)}$ into $\mathbf{d}^{(q+1)}$:

   (i) Determine the active neighbourhood system using by $\mathbf{a}^{(q+1)}$ (see (4.7)).

   (ii) Obtain optimal disparity estimates, $\mathbf{d}^{(q+1)}$, using Mean Field or BP for the Markov random field model (4.5).

The alternation is carried out until a large percentage (in our experiments 90%) of the displacement values are equal to zero at each scale. This indicates that no more corrections are required for the disparities and the disparity discontinuities now correspond to the object boundaries.

4.5 Experimental results

In this section, we report the experimental evaluation of our algorithm. We first discuss the parameter settings used for the results presented in the this chapter.

Parameter settings :

The disparity range and therefore the value of L is fixed to different values depending on each image pair. The parameters used in the experiments are found heuristically. In the data term (4.6), the parameter λ_1 determines the number of intensity level differences permitted with a low penalty. Depending on the noise level in the image, λ_1 and also T_1 are set within the range of $\lambda_1 = 0.4$ to 1.4 and $T_1 = 5$ to 10, respectively. The parameter λ_2 in the interaction function (4.11) determines the smoothness of truncated-linear (figure 2.3(b), 21). As small value for this parameters means that large differences in the neighbouring disparities will have smaller penalties and therefore will favour smoother disparity maps. The value T_2 gives a constant penalty for disparity difference greater than itself. A small value for this parameter favours discontinuities and therefore may result in noisy disparity map, whereas a very large value would result in an effect similar to that of linear function, i.e., very smooth disparity maps. The T_2 parameter is also used for the discontinuity chain formation, where it influences the number of chains formed. This is further discussed in section 4.6. In the following results, λ_2 is set to 1 and T_1 ranges between 2.0 to 2.4. The search range for U_c, given by an integer w in (4.20) determines how far the function should search for a gradient maximum in the image. This mainly depends on how textured or noisy the image is. If the image fairly smooth and this parameter is chosen to be small it may not find any global maxima in its vicinity and therefore, may favour no correction. On the other hand, if the image is textured and a w is set to a large value then it may favour the movement of the discontinuity chain to a gradient maximum which does not belong to a object boundary, but to a texture. In our case for textured images w value is set to 2 for textured images and to 5 otherwise. The alternation carried out for 4 levels.

The parameters β_d and β_c, which determine the influence of the data and the interaction term, are set to 1. With these parameters we will now present the result obtained using the proposed approach.

We first show the result on a simple *texture* stereo-image pair, where the left reference image is shown in figure 4.8(a). This simple stereo pair consists of only two disparity values: 0 and 30. Even though for this simple example it is possible to find the exact disparity map without the boundary extraction, we use it to illustrate the evolution of the algorithm. figure 4.8(b) shows the corresponding gradient image used as evidence/data for the displacement estimation. In figures 4.8(c) and 4.8(d) we show the disparity and corresponding object boundary iteration $q = 0$ at the coarsest scale ($l = 3$). The final results of disparity and boundary for the original image (finest scale $l = 0$) are shown in figures. 4.8(e) and 4.8(f). We see how the boundary and disparity are simultaneous corrected to give a much accurate result.

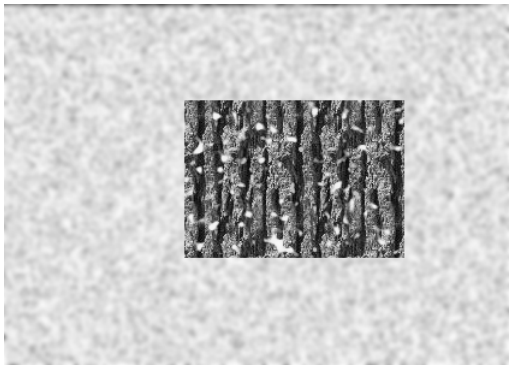
We will now present the result of the proposed approach on more realistic images. Two of the pairs correspond to the *map* (figure 4.9(a)) and the *venus* (figure 4.10(a)) images from the Middlebury database. Two others are images acquired in our laboratory; *book* (figure 4.11(a)) and *hat* (figure 4.12(a)) . It is to be noted that for all the stereo pairs only the original reference images (left) are presented in the figures.

The results for the *map* stereo pair clearly show the advantage of including the boundary estimation. In the case of highly textured images, such as the *map* image, the standard boundary detection approaches usually fail to segregate the objects. The disparity map in figure 4.9(c) is the one obtained without boundary estimation. As can be seen in figure 4.9(d) the disparity discontinuities are at improper locations but in the vicinity of the actual object boundaries. The proposed cooperative approach, using the gradient information (figure 4.9(b)), is able to obtain a corrected disparity map, and the object boundaries as shown in figures 4.9(e) and 4.9(f) respectively.

Similar results are observed for the *venus* and *book* images. With the standard Mean Field algorithm the disparity is correct at almost all regions except at the boundaries. This is illustrated in figures 4.10(c) and 4.11(c) and the boundaries corresponding to them are shown in figures 4.10(d) and 4.11(d) respectively. The corrected disparity maps obtained using our approach can be seen in figures 4.10(e) and 4.11(e). The actual object boundaries are shown in figures 4.10(f) and 4.11(f).

Finally, we present results on the *hat* image, which is particularly interesting as the object of interest has smooth boundaries. As can be seen from figure 4.12(c), the standard Mean Field algorithm that only estimates disparity information fails to obtain good results. Our method provides more satisfactory results for disparity (figure 4.12(e)) and boundary estimation (figure 4.12(f)).

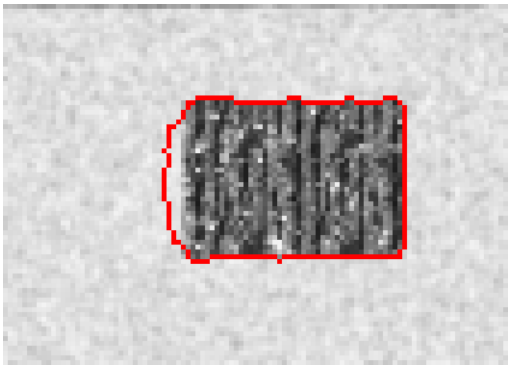
All the previous examples show results of our approach considering only Mean Field for the optimization of the disparity-MRF. The figures 4.13 and 4.14 show results of the proposed approach on *map* and *venus* using BP instead. As can be seen from the results for disparity (figures 4.13(b) and 4.14(b)) and boundary (figures 4.13(c) and 4.14(c)) the



(a) Left Reference Image

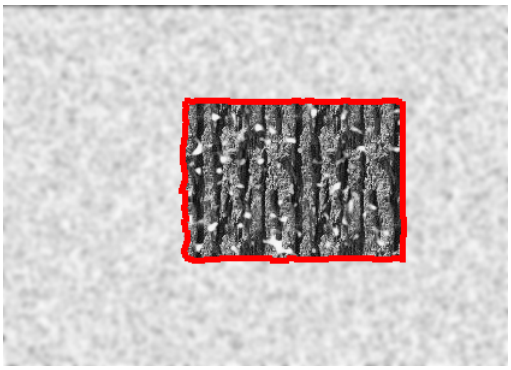
(b) Gradient Magnitude

(c) Initial Disparity at coarsest level 4



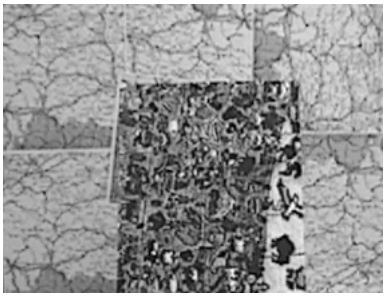
(d) Disparity discontinuities associated with 4.8(c)

(e) Disparity map using proposed approach (Mean Field)

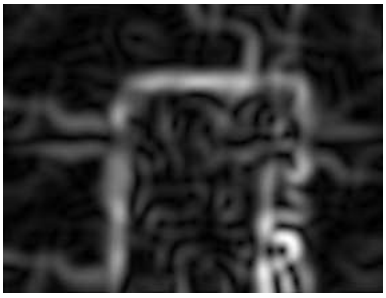


(f) Object boundary with 4.8(e)

Figure 4.8: Results on *texture* image: Shows evolution of the algorithm using the coarse-to-fine strategy suggested by Felzenszwalb and Huttenlocher [2006].

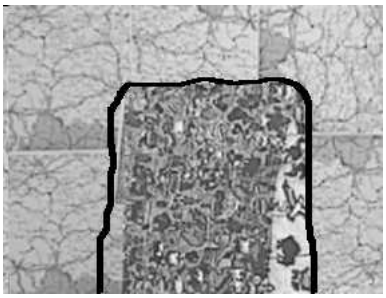


(a) Left Reference Image



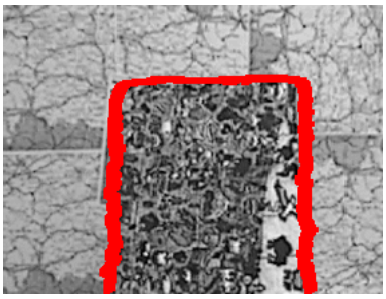
(b) Gradient Magnitude

(c) Standard Mean Field result



(d) Disparity discontinuities associated with 4.9(c)

(e) Disparity map using proposed approach (Mean Field)

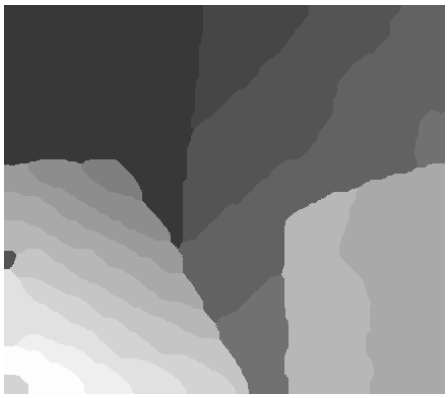


(f) Object boundary with 4.9(e)

Figure 4.9: Results on *map* image: These figures show the performance of the algorithm on a highly textured image. figure 4.9(d) show how the boundaries are improperly localized using a standard Mean Field approach. Figures 4.9(e) and 4.9(f) show the result using the proposed cooperative approach.

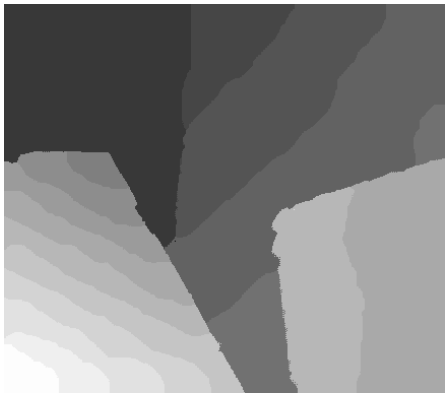
(a) Left Reference Image

(b) Gradient Magnitude

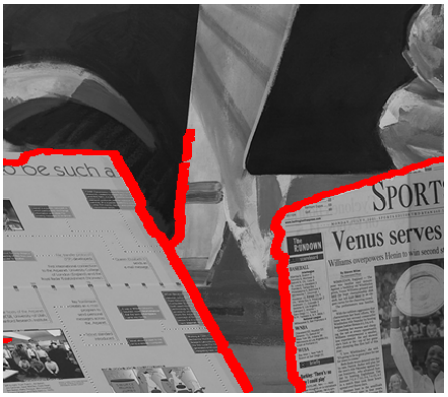


(c) Standard Mean Field result

(d) Disparity discontinuities associated with 4.10(c)



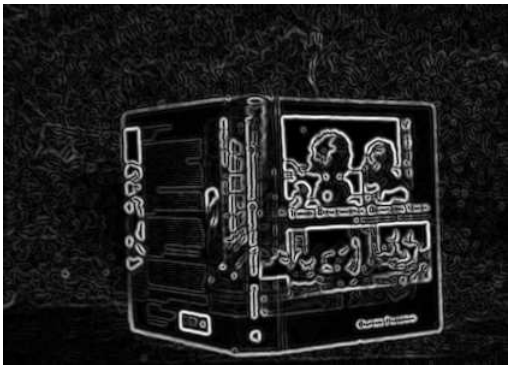
(e) Disparity map using proposed approach (Mean Field)



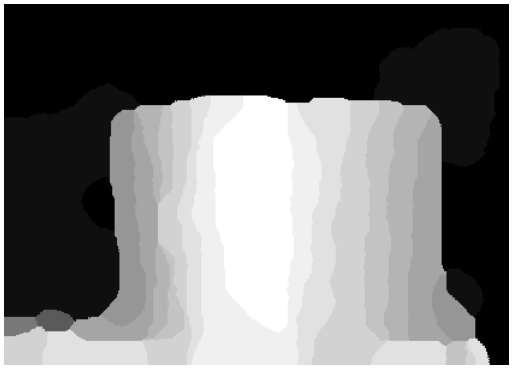
(f) Object boundary with 4.10(e)

Figure 4.10: Results on *venus* image using proposed approach:figures 4.10(e) and 4.10(f) show the result using the proposed cooperative approach.

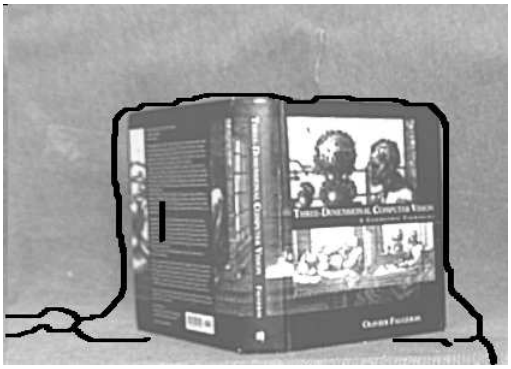
(a) Left Reference Image



(b) Gradient Magnitude

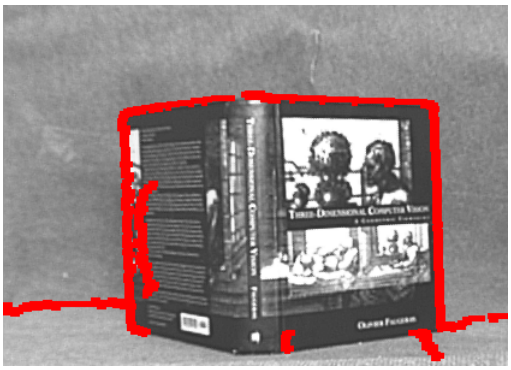


(c) Standard Mean Field result



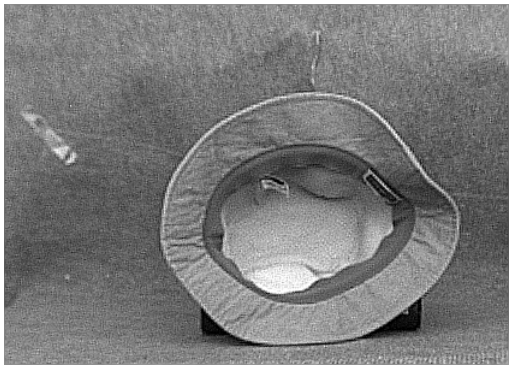
(d) Disparity discontinuities associated with 4.11(c)

(e) Disparity map using proposed approach (Mean Field)

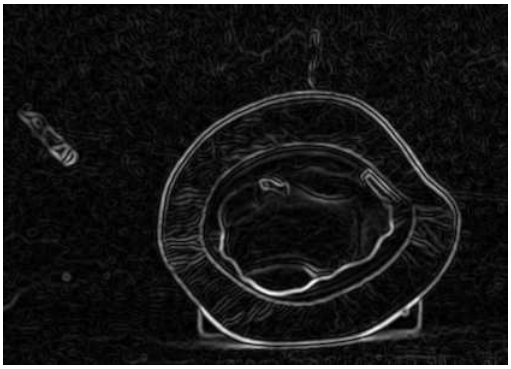


(f) Object boundary with 4.11(e)

Figure 4.11: Results on *book* image:figures 4.11(e) and 4.11(f) show the result using the proposed cooperative approach.

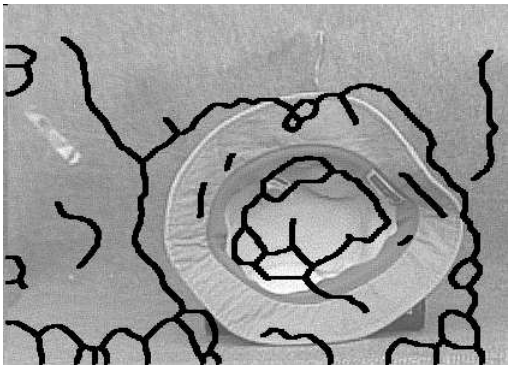


(a) Left Reference Image



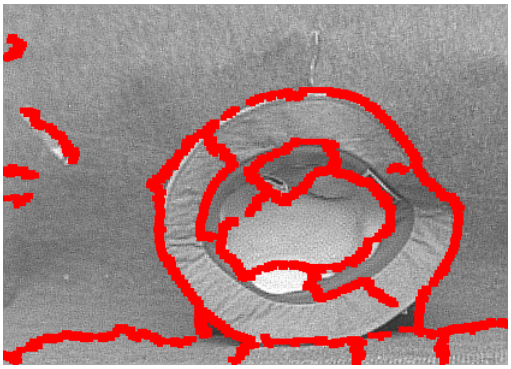
(b) Gradient Magnitude

(c) Standard Mean Field result



(d) Disparity discontinuities associated with 4.12(c)

(e) Disparity map using proposed approach (Mean Field)

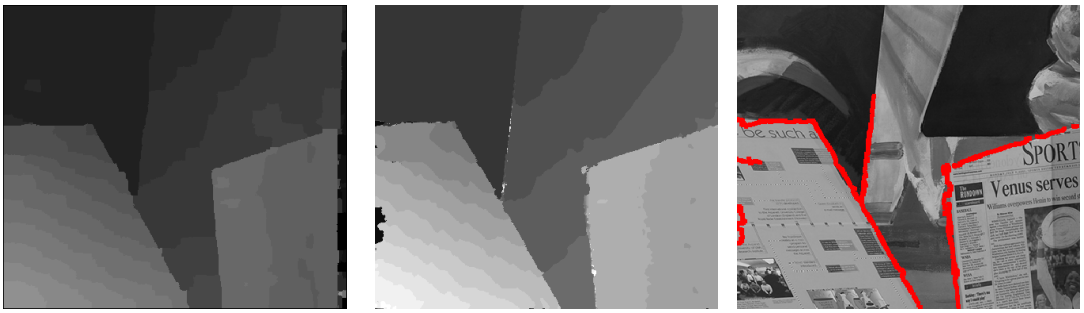


(f) Object boundary with 4.12(e)

Figure 4.12: Results on *hat* image: This example is interesting because of its smooth boundaries. We see the improved performance of proposed approach in figures 4.12(e) and 4.12(f) as compared to the results obtained by standard Mean Field in figure 4.12(c)

(a) BP results without cooperative boundary extraction (b) Disparity map using proposed approach (BP) (c) Object boundary with 4.13(b)

Figure 4.13: Results on *map* image using BP. These results are similar to the ones obtained in figures 4.9(e) and 4.9(f)



(a) BP results without cooperative boundary extraction (b) Disparity map using proposed approach (BP) (c) Object boundary with 4.14(b)

Figure 4.14: Results on *venus* image using BP. These results are similar to the ones obtained in figures 4.10(e) and 4.10(f)

results are not very different from those obtained using Mean Field. This illustrates that any other state-of-the-art MRF-based algorithm could as well be used for the optimization of the disparity-MRF. Furthermore, all these results clearly illustrate the advantage of our joint probabilistic model and confirm that such a cooperative approach is the right direction for obtaining more accurate disparity results along with better object boundary information.

4.6 Discussion

The main originality of the approach presented in this chapter is the definition of a model that explicitly considers relationships between disparity and object boundaries through conditional distributions. As a result, we observed a significant gain in disparity and boundary

(a) $T_2 = 2.5$ (b) $T_2 = 1.2$

Figure 4.15: Discontinuity chains extracted using two different values of T_2

estimations as illustrated in our experiments. The features of this algorithm are as follows:

- A probabilistic setting that focuses on defining a correct Markovian framework to model cooperation between disparity and object boundary information.
- This probabilistic setting incorporates two important information in its modelling:
  - The disparity discontinuities usually occur at object boundaries.
  - The disparities found using standard Mean Field or BP do not localize discontinuities in disparity properly (chapter 2 section 2.4.2). However, these disparities are usually in the vicinity of the "true" object boundary.
- In this context, we introduce an active neighbourhood field which allows the disparity-MRF to be correctly defined over the standard 8-neighbourhood system, similar to one proposed by Le Hégarat-Mascle et al. [2007]..
- The standard MRF neighbourhoods strictly obey the reciprocity condition (4.4). The active neighbourhood introduced in this chapter, allows interaction even between those neighbours which do not satisfy this condition, in a mathematically sound manner.
- The displacement model allows finding the direction in which the disparity correction is to be applied and also the position of the true object boundary. For this simply the image gradient and the disparity information is used as evidence.
- This displacement field can be modelled as a standard Markov chain, which allows for exact inference.
- Use of a multi-grid approach based on Felzenszwalb and Huttenlocher [2006] allows long range interaction within the MRF lattice in just few iterations. This approach permits us to retain the accuracy of disparity information by not reducing the resolution of the image, but only that of the data cost.
- We have shown that the disparity-MRF can be optimized using either Mean Field or

BP. Similarly, it is not necessary that the displacement Markov chain be optimized using Viterbi algorithm only. However, given the simple nature in which it is modelled, standard viterbi performs sufficiently well.

While the proposed approach had numerous advantageous features, there are certain limitations which have not been addressed. In particular, we did not investigate the problem of automatically estimating the parameters and they are currently fixed manually. Another important limitation lies in the extraction of discontinuity chains. The discontinuity chains are extracted taking into account the robust function defined in (4.8), where a position whose disparity difference is greater than the threshold value T_2, is retained as a point on discontinuity chain. While in most cases this chain is sufficient to demarcate the object boundary, sometimes (see figure 4.15(a)) not all the chain points are extracted. This is mainly because of threshold parameter T_2. We see in figure 4.15(b) that if this parameter is reduced to extract more chain points, it wrongly extracts chains which are within the objects. These chains within the objects occur due to the fronto-parallel assumption of the disparity-MRF model. This assumption causes a staircase affect in disparity whenever there is a sloping object surface in the scene of the stereo-pair. As the disparity values are corrected only at points of the chain and left unchanged otherwise, the final result for both disparity and boundary is only as accurate as the chain extracted.

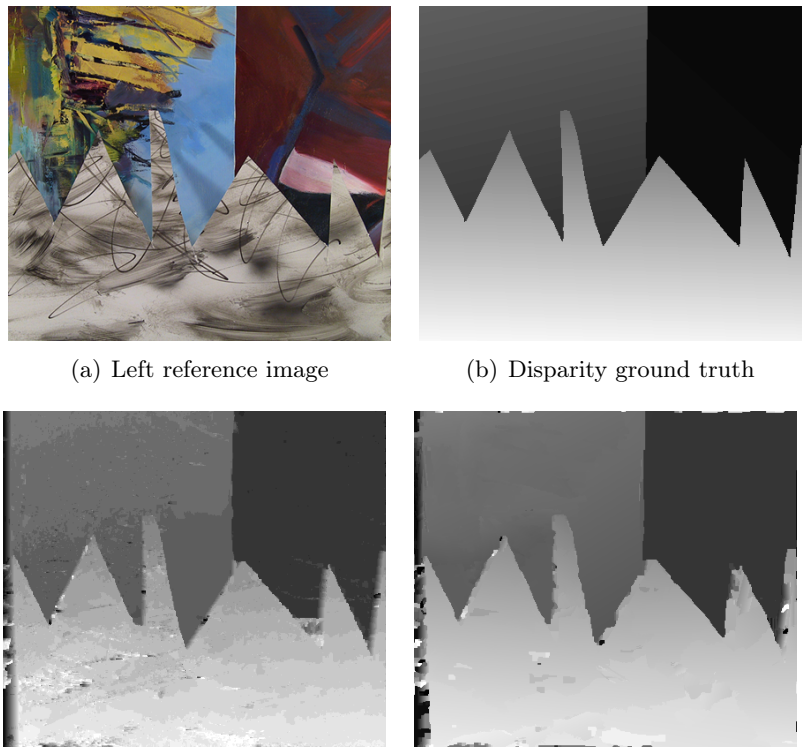
In the next chapter, we deal with the extraction of disparity without the fronto-parallel assumption. This is done mainly by taking into consideration the surface-geometric properties of the objects in the scene. We will use some of the basic aspects already presented in this as well as previous chapters, namely, coupled-MRFs, Alternation Maximization and multi-grid optimization.

Chapter 5

Estimating Disparity for Slanted and Curved Surfaces

As discussed in section 2.4.3 of chapter 2, most of the recent algorithms in stereo disparity estimation make an inherent *fronto-parallel assumption* in their modelling, thus biasing the results towards piecewise-constant "staircase solutions". As seen in section 2.4.3, a number of attempts have been made to obtain a disparity maps in accordance with surface geometric properties. For example, Devernay and Faugeras [1994] proposed to extend the classical correlation method to compute both the disparity and its derivatives and related them to the differential properties of the surface. As most of these computations are performed on the disparity, this method becomes numerically unstable while considering higher-order derivatives. Lin and Tomasi [2004] estimate the scene structure as a set of smooth surface patches, while performing segmentation and correspondence iteratively. However, unlike Devernay and Faugeras, they do not consider the geometrical properties of the surface itself. Approaches like Yang et al. [2009], Zitnick and Kang [2007], Xu and Jia [2008] and Sun et al. [2005] perform segmentation-based stereo matching (see section 2.4.2 of chapter 2 for details) within a Bayesian-MRF framework. In such a framework, these approaches are able to recover slanted surfaces due to the plane-fitting performed on the estimated disparities within segmented regions. These approaches implicitly assume fronto-parallel planes in the definition of their objective function and cannot handle curved surfaces. In order to overcome this limitation, Woodford et al. [2009] and Smith et al. [2009] present a framework incorporating higher-order priors to encode the surface properties. While Woodford et al. [2009] use a new quadratic pseudo-boolean optimization, Smith et al. [2009] suggest a non-parametric approach casting the pixels and disparity together as networks using sparse graphs which are matched then using graph cuts.

In this chapter, we propose an algorithm that recovers binocular disparities in accordance with the surface properties of the scene under consideration. For example, for a stereo



(a) Left reference image

(b) Disparity ground truth

(c) Staircase solution obtained by standard Belief Propagation

(d) Output considering scene geometric constraints

Figure 5.1: Sawtooth Image: staircase (c) versus surface consistent (d) disparity maps.

input of figure 5.1(a) with ground-truth figure 5.1(b), the result obtained from a Belief Propagation algorithm without post-processing is figure 5.1(c). The novelty our algorithm is that it attempts to provide surface consistent solutions, as illustrated in figure 5.1(d). Our method is inspired by Li and Zucker [2006b, 2010]'s work, which explicitly accounts the differential geometric contextual information, in a Markov Random Field (MRF) based disparity estimation framework. Li and Zucker use additional geometric constraint to ensure both depth and surface normals estimates are consistent with the surface, which they refer to as *geometric consistency*. They perform all the derivative computations in the depth space to ensure numerically stability. Li and Zucker therefore requires the knowledge of the internal camera parameters. One of drawbacks of this algorithm is that precomputes the local surface normals.

We propose to carry out cooperatively both disparity and normal estimations using coupled random fields that to encode consistency between disparities and surface properties. In this chapter, disparity as well as normals are modelled as Conditional Random Fields

(CRFs). The geometric contextual information is included within each of these models to favour solutions consistent with the scene surfaces – possibly slanted and/or curved. The models are built under the assumption that the scene in question is made of piecewise smooth surfaces and disparity is used as observed data. The proposed joint model results in a posterior distribution, for both the disparity and normal fields [1]. The estimated disparities and surface normals are determined according to a Maximum A Posteriori (MAP) principle. The global optimization is performed using Alternation Maximization (chapter 3, section 3.3). The idea here, we recall, is to alternately maximize two conditional probabilities; the first one pertains to disparities given the normals and the observation, and the second conditional is for normals given the disparities and the observation.

In the next section we will give a brief background on the surface differential geometry. Subsequently, we will discuss our joint model in detail section 5.2. The alternation maximization procedure along with the parameter settings for the experiments are presented in section 5.3. Finally, we show the experimental results obtained using the proposed approach and its features are discussed in section 5.4 and section 5.5 respectively.

5.1 Background

In this section, we explain the idea of geometric contextual information that is used to take into account the surface properties of the scene during disparity estimation. Details of on the differential geometry of the surface can be found in Li and Zucker [2010, 2006b] and Do Carmo [1976]. We also explain the importance of estimating the surface normals along with the disparities to obtain geometrically consistent solutions.

The idea behind geometric consistency is understood by studying the change in the *position and surface normal* as we move along this surface. That is, given two neighbouring points $\mathbf{p}$ and $\mathbf{q}$ on surface S, the idea is to analyze how the position and surface normals change as we move along a certain tangential direction say $\mathbf{v}$. To understand this, we consider a position and a normal vector field. We define a vector $\mathbf{X_p}$, in the position vector field, which is located at $\mathbf{p}$ (i.e., $\mathbf{X_p} = \mathbf{p}$), as shown in figure 5.2(a). This vector $\mathbf{X_p}$ can also be seen as a positional measurement at $\mathbf{p}$. If we move along the tangential vector $\mathbf{v}$ to the neighbouring position $\mathbf{q}$, we can compute, through first-order approximation, the position vector $\mathbf{X^*_q}$ as follows:

$$\mathbf{X^*_q} = \mathbf{X_p} + \nabla_{\mathbf{v}}\mathbf{X_p} \tag{5.1}$$

where operator $\nabla_{\mathbf{v}}$ represents the covariant derivative, which measures initial the rate of change of $\mathbf{X_p}$ as the point $\mathbf{p}$ moves along the direction $\nabla_{\mathbf{v}}$. The covariant derivative is usually expressed in terms of the directional derivatives and in case of $\nabla_{\mathbf{v}}\mathbf{X_p}$ can shown to be the same as the tangent vector $\mathbf{v}$ (see Li and Zucker [2010]). We also have direct a measurement of the position vector at position $\mathbf{q}$, which is denoted as $\mathbf{X_q}$. The discrepancy

1. This work was originally published in the papers Narasimha et al. [2009] and Narasimha et al. [2010].

between $\mathbf{X_q}$ and $\mathbf{X_q^*}$ measures the geometric consistency between the nearby candidate position vectors at $\mathbf{p}$ and $\mathbf{q}$. The first-order approximation imposes that the position vectors lie on the same planar surface, as shown in figure 5.2(a).

Similarly for surface normals, lets consider the surface normal $\mathbf{N_p}$, in a normal vector field, at the position $\mathbf{p}$ on the surface S. If we move along the tangent vector $\mathbf{v}$ to position $\mathbf{q}$, the computed surface normal $\mathbf{N_q^*}$ using first-order approximation is the following:

$$\mathbf{N_q^*} = \mathbf{N_p} + \nabla_{\mathbf{v}}\mathbf{N_p} \tag{5.2}$$

where $\nabla_{\mathbf{v}}\mathbf{N_p}$ is the covariant derivative. This derivative, which related to the shape operator $S_p(\mathbf{v}) = -\nabla_{\mathbf{v}}\mathbf{N_p}$, gives an infinitesimal description of the way the surface S is curving in the 3D space. The discrepancy between the measured normal $\mathbf{N_q}$ and $\mathbf{N_q^*}$ encodes the consistency of candidate normals at position $\mathbf{p}$ and $\mathbf{q}$. If a planar surface is assumed , then the two normals at position $\mathbf{p}$ and $\mathbf{q}$ must be the same.

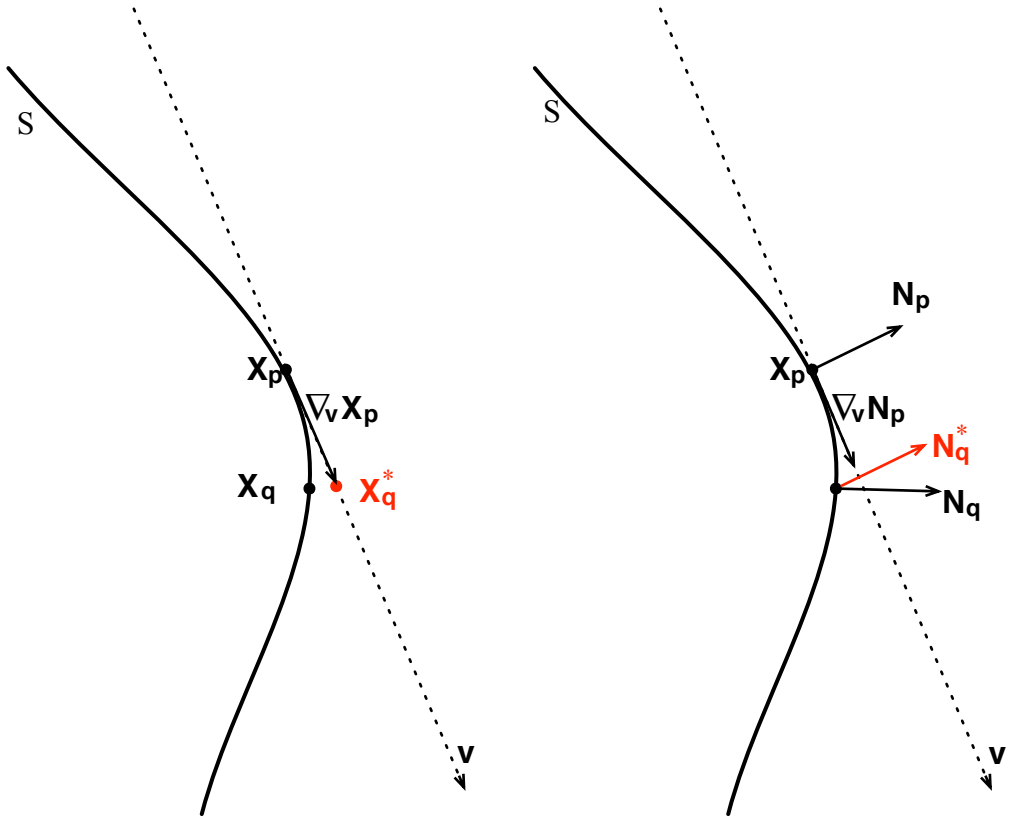
Therefore, the principle of *geometric consistency* between two neighbouring points $\mathbf{p}$ and $\mathbf{q}$ is that the information at $\mathbf{p}$ (i.e., $\mathbf{X_p}, \mathbf{N_p}, \nabla_{\mathbf{v}}\mathbf{X_p}, \nabla_{\mathbf{v}}\mathbf{N_p}$), and the computed geometric information ($\mathbf{X_q^*}$, $\mathbf{N_q^*}$) should agree with the measurements at $\mathbf{q}$ ($\mathbf{X_q}$, $\mathbf{N_q}$), if the two points are on the same surface.

In the case Li and Zucker [2006b], they include this measure in two separate models; one for slanted and the other for curved surfaces. In both the cases they transform the interaction function for the disparity-MRF in such a way as to include a general planar model for the disparity. The normal consistency is computed through first order difference while considering slanted surfaces. For curved surfaces an additional covariant derivative of the surface normal is used. Li and Zucker's model uses precomputed normals to ensure such surface geometric consistency of disparity. The pre-computation of the normals is done using a method similar to Devernay and Faugeras [1994]. First a SSD score is computed using a deformed window and then the direction set method is used to find the floating point values of the disparity and its derivatives, denoted by $\{d, \frac{\partial d}{\partial u}, \frac{\partial d}{\partial v}\}$, where (u, v) are the x and y image-pixel coordinates. These derivatives are then used to compute the normal $\mathbf{N}$ at any location (u, v) as follows:

$$\mathbf{N} = \frac{(-\frac{\partial d}{\partial u}, -\frac{\partial d}{\partial v}, 1)^T}{\sqrt{\left(\frac{\partial d}{\partial u}\right)^2 + \left(\frac{\partial d}{\partial v}\right)^2 + 1)}} \tag{5.3}$$

Li and Zucker update these normals by averaging neighbouring normals. This kind of update does not take into account the disparity changes occurring in the course of optimization and therefore relies largely on the precomputed normals.

We, on the other hand, suggest the use a separate random field to estimate the normals based on the disparity and vice-versa. Furthermore, we use a planar model, thus assuming that the scene in question is made of piecewise smooth surfaces. The argument supporting this assumption is that if we consider a small enough neighbourhood, it should fairly approximate both slanted and curved surfaces, thereby eliminating the need for two separate



(a) Disparity Consistency (b) Normal Consistency

Figure 5.2: Geometric consistency over a surface S: (a) The difference between $\mathbf{X_q^*}$ (shown in red) calculated using $\mathbf{X_p}$ and $\nabla_\mathbf{v}\mathbf{X_p}$ and the measured $\mathbf{X_q}$ allows for the positional consistency. (b) Similarly, the difference between $\mathbf{N_q^*}$ (shown in red) calculated using $\mathbf{N_p}$ and $\nabla_\mathbf{v}\mathbf{N_p}$ and the measured $\mathbf{N_q}$ allows for the surface normal consistency.

models. Since we are only interested in first-order approximation (piecewise planar-model), we avoid any numerical instability issues as encountered by Devernay and Faugeras [1994]. Moreover, disparity consistent updates of the normals allow the surface properties to be closely followed even within a planar model assumption, which was not handled in Li and Zucker's method.

This discussion therefore illustrates the need for linking the models for surface normals and disparity. It is to be noted that, in our approach, the normals are estimated in the disparity space rather than the Euclidean space. Before we describe the joint model for disparity and normal, we digress a little to explain the relationship between the normals in the disparity and Euclidean spaces. This is important to show that even if we estimate the normals in the disparity space, there is simple transformation that can be applied to convert them into surface normals in the Euclidean space.

5.1.1 Relationship between the normals in disparity and Euclidean space

In this section we derive the transformation required to convert the normals in disparity space to Euclidean space or *depth space*. We recall that we have a calibrated stereo camera set up and that the stereo images are rectified. Let (u_l, v) and (u_r, v) be the corresponding point in the left and right images respectively, and $d = x_l - x_r$ be the disparity. We denote the baseline, i.e., the distance between the two camera centres, by b and the focal length by f. The parameters f and b are known because of stereo calibration. Furthermore, we know from the stereo camera geometry that for a 3D scene point $\mathbf{P} = (X, Y, Z)$:

$$u_l = \frac{X + b/2}{Z/f}, \qquad u_r = \frac{X - b/2}{Z/f}, \qquad v = \frac{Y}{Z/f}, \quad \text{and} \quad d = \frac{fb}{Z} \tag{5.4}$$

This relation between 3D scene point $\mathbf{P}$ and the image co-ordinates can be written in the matrix form as follows:

$$\begin{bmatrix} X + b/2 \\ X - b/2 \\ Y \\ Z/f \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & b/2 \\ 1 & 0 & 0 & -b/2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/f & 0 \end{bmatrix}}_{H} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \simeq \begin{bmatrix} u_l \\ u_r \\ v \\ 1 \end{bmatrix} \tag{5.5}$$

where $\simeq$ stands for equality upto a non-zero scale and this case scale is Z/f. We denote the 4×4 matrix in the above equation as H. In order to simplify our task of deriving transformation of normals, we work on a *cyclopean* point. A cyclopean point is one which lies half way between the baseline. The cyclopean point (u_c, v_c) can therefore be expressed as

$$u_c = \frac{u_l + u_r}{2} \qquad v_c = v \tag{5.6}$$

and the disparity remains the same, $d = u_l - u_r$. The relation between the image locations and the cyclopean point can again be written in a matrix form as:

$$\begin{bmatrix} u_c \\ v_c \\ d \\ 1 \end{bmatrix} \sim \underbrace{\begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{G} \begin{bmatrix} u_l \\ u_r \\ v \\ 1 \end{bmatrix} \tag{5.7}$$

Representing the 4×4 matrix in the above equation as G and using the equations (5.5) and (5.7) we can write:

$$\begin{bmatrix} u_c \\ v_c \\ d \\ 1 \end{bmatrix} \simeq GH \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \simeq (GH)^{-1} \begin{bmatrix} u_c \\ v_c \\ d \\ 1 \end{bmatrix} \tag{5.8}$$

where

$$GH = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & b \\ 0 & 0 & 1/f & 0 \end{bmatrix} \quad \text{and} \quad (GH)^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & f \\ 0 & 0 & 1/b & 0 \end{bmatrix} \quad \text{respectively.} \tag{5.9}$$

The above equation, (5.9), provides the transformation matrix for projection and back-projection from cyclopean to Euclidean coordinate system and vice-versa. Now, let this point $\mathbf{P} = (X, Y, Z)$ lie on a plane $\boldsymbol{\pi}$ in 3D space:

$$\pi_1 X + \pi_2 Y + \pi_3 X + \pi_4 = 0 \tag{5.10}$$

where $\{\pi_1, \pi_2, \pi_3, \pi_4\}$ represent the plane coefficients. The first three coefficients correspond to the plane normal, $\mathbf{n} = (\pi_1, \pi_2, \pi_3)$ and last one $\pi_4/\|\mathbf{n}\|$ is the distance of the plane from the origin. In homogenous co-ordinates the plane normal can therefore be represented as a 4-vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)^T$. If the projective transformation of a point in space is given by $(GH)^{-1}$ then the plane transformation is $((GH)^{-1})^{-T} = (GH)^T$. Refer to Hartley and Zisserman [2000] for more detailed study on projective transformations. Therefore, any normal $\mathbf{n}$ in disparity space (represented as $\boldsymbol{\pi}$ in homogenous coordinates) can be back projected into Euclidean space using the following transformation:

$$(GH)^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/f \\ 0 & 0 & b & 0 \end{bmatrix}. \tag{5.11}$$

It is to be noted that from now on *normals* refer to the surface normals in the disparity space and the transformation in (5.11) can be easily applied to obtain the surface normals in Euclidean or depth space.

5.2 Joint disparity and normal model

We now revert to the modelling aspects of the normals and disparity. The mutual influence of the disparity and normals on one another is modelled using the coupled random fields framework. Note that we use the term coupled random fields instead of coupled-MRF. This is because both the disparities as well as the normals are modelled as a Conditional Random Field (CRF). A CRF (Lafferty et al. [2001]) is like an MRF except that the normals conditional probability can depend on arbitrary, non-independent observation sequence. Furthermore, it does so without forcing the model to account for the distribution of those dependencies. This allows us to model the surface normals by considering the entire disparity field, but not the disparity model itself, and vice-versa. This modification does not affect the Alternation Maximization procedure (chapter 3) that we use to find the MAP solutions for disparity and normals through the conditional probabilities.

We use similar notations as in chapter 4, where $\mathcal{S}$ of $p \times q$ is the set of pixels on a regular 2D-grid. The observed data are made of left and right images, and is denoted by $\mathbf{I} = (\mathbf{I}_L, \mathbf{I}_R)$, The disparity field is $\mathbf{D} = \{D_{\mathbf{x}}, \mathbf{x} \in \mathcal{S}\}$, where each of the random variables $D_{\mathbf{x}}$ take their values in a finite discrete set of L disparity labels $\mathcal{L}$ and $\mathcal{D} = \mathcal{L}^{p \times q}$. Similarly, we consider a surface normal field $\mathbf{N} = \{\mathbf{N}_{\mathbf{x}}, \mathbf{x} \in \mathcal{S}\}$. Each $\mathbf{N}_{\mathbf{x}} = (N_u, N_v, N_d)$ takes its values from the space $\mathbb{N}$ and the configuration space is denoted by $\mathcal{N} = \mathbb{N}^{p \times q}$. We use small letters $\mathbf{d}$ and $\mathbf{n}$ to denote specific realizations of the random fields $\mathbf{D}$ and $\mathbf{N}$. The MAP estimates of $\mathbf{D}$ and $\mathbf{N}$ are expressed as:

$$(\mathbf{d}^{MAP}, \mathbf{n}^{MAP}) = \arg\max_{\mathbf{d},\mathbf{n}} p(\mathbf{d}, \mathbf{n}|\mathbf{I}) \tag{5.12}$$

The Alternation Maximization procedure, at a given iteration i, can be performed as follows (for details see chapter 3, section 3.3):

$$\mathbf{d}^{(i+1)} = \arg\max_{\mathbf{d}\in\mathcal{D}} p(\mathbf{d}|\mathbf{n}^{(i)}, \mathbf{I}) \tag{5.13}$$

$$\mathbf{n}^{(i+1)} = \arg\max_{\mathbf{n}\in\mathcal{N}} p(\mathbf{n}|\mathbf{d}^{(i+1)}, \mathbf{I}) \tag{5.14}$$

In the following sections, we will define the two conditional distributions $p(\mathbf{d}|\mathbf{n}, \mathbf{I})$ and $p(\mathbf{n}|\mathbf{d}, \mathbf{I})$. The disparity model, $p(\mathbf{d}|\mathbf{n}, \mathbf{I})$, is based on the one defined by Li and Zucker [2006b]. For the surface normals, $p(\mathbf{n}|\mathbf{d}, \mathbf{I})$, we propose two possible models:

- The first model uses standard BP to estimate the normals. In order to use BP, we model the normal-CRF over a discrete normal space. This discretization of the normal space is achieved by subdividing an icosahedron to represent dense set of directions.
- The second model does not discretize the normal space. The optimization used is the Iterated Conditional Modes (ICM), which allows for estimation normals without discretization. The model itself takes its inspiration from the deterministic procedure called normal voting proposed by Page et al. [2002].

5.2.1 Disparity model given the normals

We first specify the disparity distribution conditionally to the normal field and the observed data. The model is expressed as a CRF with an energy function consisting of two terms, a data dependent term and a regularizing or interaction term. The data term is similar to the one described by Yoon and Kweon [2007] who use a weighted window matching metric for stereo matching. Our interaction term is a symmetric modified version of the one presented in Li and Zucker [2006b]. Although the interpretation is similar, we propose to include geometric information via surface normals considered as a separate random field. Expressing compatibility between the disparity and normal fields enables us to encode geometric constraints without computing disparity derivatives directly from the disparity field. Furthermore, we achieve results similar to Li and Zucker [2006b] using only first order differential information but obtained from the normal field. More specifically, we define, $p(\mathbf{d}|\mathbf{n},\mathbf{I})$ as a CRF on $\mathcal{D}$, for every $\mathbf{n} \in \mathcal{N}$:

$$p(\mathbf{d}|\mathbf{n},\mathbf{I}) \propto \Phi_D(\mathbf{d},\mathbf{I})\Psi_D(\mathbf{d},\mathbf{n}). \tag{5.15}$$

Data term

The first term in (5.15) represents the **data term**, similar to the one described by Yoon and Kweon [2007]. This term assigns a cost at each location $\mathbf{x}$ based on a weighted window matching metric that takes into account both the colour and the proximity of the pixels within the window. Furthermore, it takes into account the weights in both left and right image-windows, as considering only the reference image-window (left) the computed difference can be erroneous when the right image-window has pixels from different depths. We formulate this cost as a robust function

$$\Phi_D(\mathbf{d},\mathbf{I}) = \exp\left(-\sum_{\mathbf{x}\in\mathcal{S}} \min\Big(\phi(\mathbf{I}_L,\mathbf{I}_R,d_\mathbf{x}),2T\Big)\right), \tag{5.16}$$

where,

$$\phi(\mathbf{I}_L,\mathbf{I}_R,d_\mathbf{x}) = \frac{\sum_{\mathbf{y}\in W_\mathbf{x},\bar{\mathbf{y}}\in W_{\bar{\mathbf{x}}}} w(\mathbf{x},\mathbf{y})w(\bar{\mathbf{x}},\bar{\mathbf{y}})e\Big(\mathbf{I}_L(\mathbf{y}),\mathbf{I}_R(\bar{\mathbf{y}})\Big)}{\sum_{\mathbf{y}\in W_\mathbf{x},\bar{\mathbf{y}}\in W_{\bar{\mathbf{x}}}} w(\mathbf{x},\mathbf{y})w(\bar{\mathbf{x}},\bar{\mathbf{y}})}. \tag{5.17}$$

In order to understand the equation (5.17), consider a candidate correspondence $\bar{\mathbf{x}}$ in the right image for the point $\mathbf{x}$ in the left, i.e $\bar{\mathbf{x}} = \mathbf{x} - (d_\mathbf{x}, 0)$. To compute the cost for a candidate disparity $d_\mathbf{x}$, we first determine the pixel-wise cost as follows:

$$e\Big(\mathbf{I}_L(\mathbf{x}),\mathbf{I}_R(\bar{\mathbf{x}})\Big) = |\mathbf{I}_L(\mathbf{x}) - \mathbf{I}_R(\bar{\mathbf{x}})| \tag{5.18}$$

where $\mathbf{I}_L(\mathbf{x})$ and $\mathbf{I}_R(\bar{\mathbf{x}})$ represent the intensity values[2] at the positions $\mathbf{x}$ and $\bar{\mathbf{x}}$ of the left and the right images, respectively and $|\cdot|$ represents the absolute value. This cost computed for each pixel within the windows $W_{\mathbf{x}}$ and $W_{\bar{\mathbf{x}}}$ centred at $\mathbf{x}$ and $\bar{\mathbf{x}}$ respectively. Each pixel within the window $\mathbf{y} \in W_{\mathbf{x}}$ (respectively for $\bar{\mathbf{y}} \in W_{\bar{\mathbf{x}}}$) is weighted according its colour difference:

$$\Delta c_{\mathbf{xy}} = \sum_{c \in \{r,g,b\}} |\mathbf{I}_c(\mathbf{x}) - \mathbf{I}_c(\mathbf{y})| \tag{5.19}$$

where $\mathbf{I}_c$ is the intensity of the colour channel c and is associated with the left image ($\mathbf{I}_L$). The pixels within the window $W_{\mathbf{x}}$ are further weighted depending on the spatial proximity $\nabla g_{\mathbf{xy}}$ of $\mathbf{y}$ to $\mathbf{x}$:

$$\Delta g_{\mathbf{xy}} = \|\mathbf{x} - \mathbf{y}\|. \tag{5.20}$$

So the overall expression for the weight associated to each pixel with in the window $W_{\mathbf{x}}$ as follows:

$$w(\mathbf{x}, \mathbf{y}) = \exp\left(- \frac{\Delta c_{\mathbf{xy}}}{\gamma_c} - \frac{\Delta g_{\mathbf{xy}}}{\gamma_g} \right). \tag{5.21}$$

A similar weight $w(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is computed for $\bar{\mathbf{x}}, \bar{\mathbf{y}} \in W_{\bar{\mathbf{x}}}$. The idea behind using such a cost is that, apart from considering colour proximity it also considers the geometric proximity between the pixels.

Interaction term

The **interaction term**,$\Psi_D(\mathbf{d}, \mathbf{n})$ in (5.15), has a standard pair-wise form:

$$\Psi_D(\mathbf{d}, \mathbf{n}) = \prod_{\mathbf{x} \in \mathcal{S}} \prod_{\mathbf{y} \in \mathcal{N}_{\mathbf{x}}} \exp\left(- \frac{\psi(d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{n})}{\sigma_D} \right) \tag{5.22}$$

where $\mathbf{x} = (u_{\mathbf{x}}, v_{\mathbf{x}})$ and $\mathbf{y} = (u_{\mathbf{y}}, v_{\mathbf{y}})$ are the neighbouring pixels on the image grid. We consider a standard 8-neighbourhood system $\mathcal{N}_{\mathbf{x}}$ as in chapter 4. The term $\psi(d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{n})$ specifies how the neighbouring disparities interact but also encodes geometric constraints via consistency with the surface normal field $\mathbf{n}$. In the spirit of the co-planar model in Li and Zucker [2006b], we model the interaction term in such a way that favours neighbouring disparities lying on the same planar surface (See figure 5.3). This is done by a first order Taylor approximation of the the disparity at $\mathbf{x}$ as follows:

$$d_{\mathbf{x}} + \frac{\partial d_{\mathbf{x}}}{\partial u} \Delta u + \frac{\partial d_{\mathbf{x}}}{\partial v} \Delta v \tag{5.23}$$

where the Δu and Δv are small changes in x and y component directions of the position $\mathbf{x}$. In order to include this model in the interaction term, we first define the derivatives in (5.23) in terms of the surface normal information. For a given realization of the surface

2. For colour images I(x) represents the average of the RGB channels.

normals $\mathbf{n}$, the surface normal at position $\mathbf{x}$ can be written as $\mathbf{n_x} = (n_u, n_v, n_d)$. Thus, the disparity partial derivatives can be computed as:

$$\frac{\partial d_{\mathbf{x}}}{\partial u} = -\frac{n_u}{n_d} \;\; ; \;\; \frac{\partial d_{\mathbf{x}}}{\partial v} = -\frac{n_v}{n_d}, \tag{5.24}$$

with $\Delta u = (u_{\mathbf{y}} - u_{\mathbf{x}})$ and $\Delta v = (v_{\mathbf{y}} - v_{\mathbf{x}})$. Moreover, the interaction function takes into account the disparity and normal information from two neighbouring positions $\mathbf{x}$ and $\mathbf{y}$. Therefore $\psi(d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{n})$ is the same as $\psi(d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{n_x}, \mathbf{n_y})$ and is expressed as follows:

$$\begin{aligned} \psi(d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{n_x}, \mathbf{n_y}) &= \left(\left| d_{\mathbf{y}} - d_{\mathbf{x}} - \frac{\partial d_{\mathbf{x}}}{\partial u}(u_{\mathbf{y}} - u_{\mathbf{x}}) - \frac{\partial d_{\mathbf{x}}}{\partial v}(v_{\mathbf{y}} - v_{\mathbf{x}}) \right| \right) \\ &+ \left(\left| d_{\mathbf{x}} - d_{\mathbf{y}} - \frac{\partial d_{\mathbf{y}}}{\partial u}(u_{\mathbf{x}} - u_{\mathbf{y}}) - \frac{\partial d_{\mathbf{y}}}{\partial v}(v_{\mathbf{x}} - v_{\mathbf{y}}) \right| \right). \end{aligned} \tag{5.25}$$

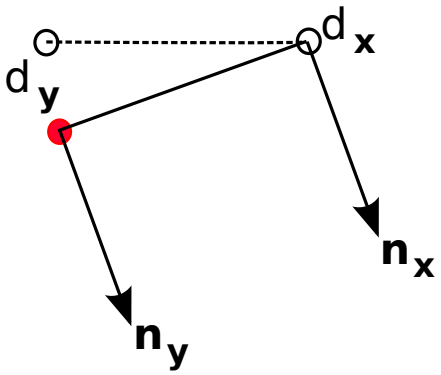


Figure 5.3: The disparity at $d_{\mathbf{x}}$ imposes $d_{\mathbf{y}}$ (shown in red) to lie on the same planar surface based on the normals $\mathbf{n_x}$ and $\mathbf{n_y}$ at positions $\mathbf{x}$ and $\mathbf{y}$ respectively. The black circle indicates the fronto-parallel value of $d_{\mathbf{y}}$.

Disparity optimization

The optimization for disparity is done using standard techniques such as Mean field approximation or BP. For details of such techniques see section 2.3 (page 25) of chapter 2. To incorporate the differential properties, these techniques have to be modified to take into consideration floating point disparities instead of integer. In order to do so, we compute an interpolated disparity map obtained using a segmentation and plane-fitting procedure: The interpolated values are derived by first segmenting the intensity image (left) and the performing plane-fitting on the disparities within each of these image segments. These interpolated disparities enable us to follow Li and Zucker [2006b] by considering so-called

floating disparity labels. Li and Zucker use the deformed window approach by Devernay and Faugeras [1994] to find both disparity and its derivatives. As the disparity derivatives in our case is obtained from the separate normal model, it is sufficient if we can find just the interpolated disparities.

Now we use a similar approach as Li and Zucker to find the floating disparities. Starting from a discrete set of L integer disparity labels $\mathcal{L} = \{d_1, \ldots, d_L\}$, we allow them to move to another set of L possibly non-integer labels. In order to prevent a bias towards planar solutions, the plane-fitting procedure is carried out at every iteration so that each new set of non-integer labels are generated taking into account the differential geometric constraints within the model. Considering at iteration t, a current continuous value d at pixel $\mathbf{x}$, we find l such that $d \in [d_l \; d_{l+1})$ and then change the disparity label d_l to the non-integer label d. Moreover, as can be seen, not all the plane-fit values are chosen, the effect of such an procedure is therefore to capture finer geometric features by adapting the initial disparity discretized grid to the image scene. Importantly, this can be done while keeping the discrete pair-wise MRF formulation. This provides an efficient alternative to the quickly intractable increase of L.

5.2.2 Discrete normal model given disparity

As previously discussed, we present two ways to model the normal-CRF. We will first begin with the discrete normal model. As explained before, the normal field $\mathbf{N} = \{\mathbf{N_x}, \mathbf{x} \in \mathcal{S}\}$ is defined such that each $\mathbf{N_x}$ takes its values in the normal space $\mathbb{N}$. We discretize this space by sampling the directions on a unit sphere. In order to obtain uniformly distributed samples we use an icosahedron. Each face of the icosahedron is subdivided 4 times and in doing so, we obtain $K = 162$ vertices. The normal vector directions are obtained by the line joining the centre and the vertices of the subdivided icosahedron as shown in figure 5.4. This space of discrete directions is denoted by $\mathbb{N}_K$ and the normal configuration space is then $\mathscr{N} = \mathbb{N}_K^{p \times q}$. The normal model is now defined in manner similar to the disparity model:

$$p(\mathbf{n}|\mathbf{d}, \mathbf{I}) \propto \Phi_N(\mathbf{n}, \mathbf{d}, \mathbf{I}) \; \Psi_N(\mathbf{d}, \mathbf{n}) \; , \tag{5.26}$$

where the **interaction term** $\Psi_N(\mathbf{n}, \mathbf{d})$ is the same as in the disparity model (5.15), that is $\Psi_D(\mathbf{d}, \mathbf{n})$. The rationale is that $\Psi_D(\mathbf{d}, \mathbf{n})$, as defined by (5.22) and (5.25) (in section 5.2.1), corresponds to consistency conditions which are joint between $\mathbf{d}$ and $\mathbf{n}$. In (5.25), the dependence in $\mathbf{n}$ actually reduces to a dependence on $\mathbf{n_x}$ and $\mathbf{n_y}$, as the normal model is conditioned on the disparity $\mathbf{d}$.

Data term

The discrete normal model differs then mainly in the form of the data term $\Phi_N(\mathbf{n}, \mathbf{d}, \mathbf{I})$. The idea is to focus on data information that can directly and specifically impact the normal field values. The particularity is that for normal estimation, data information cannot be

Figure 5.4: The red lines represent the directions for the normals from the origin, obtained by subdivided icosahedron.

expressed in terms of $\mathbf{I}$ only, but depends also on the current disparity field $\mathbf{d}$. Assuming a current $\mathbf{d}$, we compute an *observed* normal value $\mathbf{p_x}$, for each pixel $\mathbf{x} \in \mathcal{S}$, using a plane-fitting procedure. A set of planes is found by first segmenting the image (in our case the left image $\mathbf{I}_L$) into small regions and then approximating the surface corresponding to $\mathbf{d}$ in each region by a plane. For a given region, the normal to its plane provides then the value of $\mathbf{p_x}$ for all $\mathbf{x}$ in the region. We therefore define a **data term** that favours small distances between the current normals $\mathbf{n}$ and the *observed* normals denoted by $\mathbf{p} = \{\mathbf{p_x}, \mathbf{x} \in \mathcal{S}\}$:

$$\Phi_N(\mathbf{n}, \mathbf{d}, \mathbf{I}) = \exp\Big(- \sum_{\mathbf{x} \in \mathcal{S}} ||\mathbf{n_x} - \mathbf{p_x}||^2 \Big) \, . \tag{5.27}$$

It is to be noted that in the above equation the dependence on $\mathbf{I}$ and $\mathbf{d}$ is through $\mathbf{p_x}$. The segmentation and plane fitting process is essential for two reasons:

1) When the regions are small enough, the resulting set of planes provides a reasonable approximation of the possibly curved disparity surface as a piecewise linear surface.

2) The regions are found according to the colour/grey-level properties of the various objects in the scene so that the resulting normals are likely to reflect a number of discontinuities.

Furthermore, this data term (5.27) prevents smoothing of the discontinuities (because of 2)), thereby balancing the regularizing term (5.22) which is performed across regions.

With this discrete model and the defined data and interaction terms, a standard BP can be used to find the best normals at every pixel. However, this model poses a large problem due to discretization of the normal field. A dense discretization requires 162 or more directions to be uniformly defined over a unit sphere. Even if we use a multi-grid approach, the BP has to perform optimization over at least 162 labels, which is computationally inefficient. Furthermore, even though a plane-fitting and segmentation procedure has its

advantages, it greatly biases the final solution to a planar solution. These serious drawbacks provide a motivation to find a better model without the discretization of the normal space and where no planefitting-segmentation procedure is required.

5.2.3 Normal model without discretization

To overcome the drawbacks of the discrete normal model we introduce a disparity conditional normal model which is a Conditional Random Field with a simple Gaussian distribution:

$$p(\mathbf{n}|\mathbf{d},\mathbf{I}) \propto \prod_{\mathbf{x}\in\mathcal{S}} \prod_{\mathbf{y}\in\mathcal{N}_\mathbf{x}} \exp\left(- \frac{\|\mathbf{n_x} - \vec{E}_\mathbf{xy}(\mathbf{d},\mathbf{I},\mathbf{n_y})\|^2}{2\sigma^2} \right) \tag{5.28}$$

where $\mathcal{N}_\mathbf{x}$ is the 8-neighbourhood set. The above equation represents a pairwise relationship between the normal at $\mathbf{x}$ and its neighbours $\mathbf{y} \in \mathcal{N}_\mathbf{x}$. Instead of just computing an Euclidean distance between the two normals at positions $\mathbf{x}$ and $\mathbf{y}$, we compute the distance between $\mathbf{n_x}$ and the vector $\vec{E}_\mathbf{xy}$, which is the influence of the neighbouring normal $\mathbf{n_y}$ taking into account disparity $\mathbf{d}$ and the image $\mathbf{I}$ information. The vector $\vec{E}_\mathbf{xy}$ is determined using a method inspired by Page et al. [2002]. We express $\vec{E}_\mathbf{xy}$ as:

$$\vec{E}_\mathbf{xy} = w_\mathbf{xy}(\mathbf{d},\mathbf{I})\, \vec{N}_\mathbf{xy}, \tag{5.29}$$

where,

$$\vec{N}_\mathbf{xy} = \mathbf{n_y} - 2\cos\left(\theta_\mathbf{xy}(\mathcal{S},\mathbf{d})\right)(\vec{xy}/\|\vec{xy}\|) \tag{5.30}$$

The first term in the above equation is the current normal estimate at site $\mathbf{y}$. The $\theta_\mathbf{xy}$ in the second term, is the angle between the normal at $\mathbf{y}$ and the vector $\vec{xy}$, from x to y, where $x = (\mathbf{x}, d_\mathbf{x})$ and $y = (\mathbf{y}, d_\mathbf{y})$:

$$\cos\left(\theta_\mathbf{xy}(\mathcal{S},\mathbf{d})\right) = \frac{\mathbf{n_y}^T\, \vec{xy}}{\|\vec{xy}\|}. \tag{5.31}$$

The equation (5.30) motivated by assuming that the surface is locally of constant curvature. We assume that the surface passing through x and y is a unique sphere of constant curvature. As shown in figure 5.5, the sphere centre is on the intersection of the bisector plane and the line through y with direction $\mathbf{n_y}$. Reflecting this line at the bisector results in $\vec{N}_\mathbf{xy}$ at the point x, which can then be compared to the actual normal $\mathbf{n_x}$ using the equation (5.28). Moreover, the equation (5.31) is equal to zero that if the two points $(\mathbf{x}, d_\mathbf{x})$ and $(\mathbf{y}, d_\mathbf{y})$ are on a plane consistent with $\mathbf{n_y}$. So the second term in (5.30) can also be seen as the error between $\vec{xy}$ and the plane described by $\mathbf{n_y}$. The weight $w_\mathbf{xy}(\mathbf{d},\mathbf{I})$ is described as

$$w_\mathbf{xy}(\mathbf{d},\mathbf{I}) = \exp\left(- \frac{|d_\mathbf{x} - d_\mathbf{y}| + |\nabla\mathbf{I}(\mathbf{x})|}{\sigma_N} \right) \tag{5.32}$$

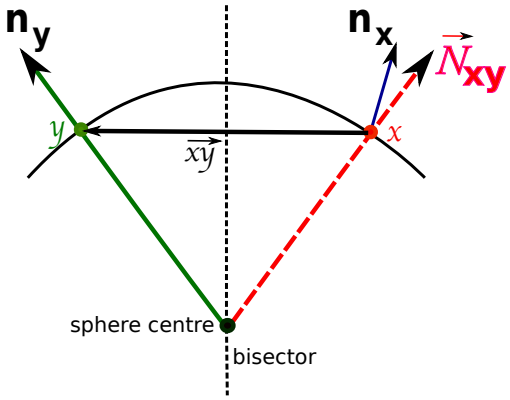


Figure 5.5: The normal $\mathbf{n_y}$ is reflected at the bisector of $\vec{xy}$ to obtain $\vec{N}_{\mathbf{xy}}$ which is compared with $\mathbf{n_x}$ through the equation (5.28).

where $|\nabla \mathbf{I}(\mathbf{x})|$ represents the gradient magnitude at $\mathbf{x}$ of the reference image[3]. The image gradient is used to weight the influence of the neighbours so as to prevent normals from being smoothed across boundaries.

While Page et al. describe a deterministic voting procedure that uses Eigen decomposition to determine the normals, we maximize $p(\mathbf{n}|\mathbf{d}, \mathbf{I})$ using an approximate MAP estimate of $\mathbf{n}$. This is done by using an Iterated Conditional Modes (ICM) procedure (Besag [1986]), in which $\mathbf{n}$ is set iteratively $\forall \mathbf{x} \in \mathcal{S}$ as follows,

$$\mathbf{n}_{\mathbf{x}}^{MAP} \approx \frac{1}{\text{Card}(\mathcal{N}_{\mathbf{x}})} \sum_{\mathbf{y} \in \mathcal{N}_{\mathbf{x}}} \vec{E}_{\mathbf{xy}}(\mathbf{d}, \mathbf{I}, \mathbf{n_y}) \tag{5.33}$$

where $\text{Card}(\mathcal{N}_{\mathbf{x}})$ is the cardinality of the set $\mathcal{N}_{\mathbf{x}}$. The ICM algorithm proceeds first by choosing an initial configuration for normals $\mathbf{n}$. Then, it iterates over each site $\mathbf{x}$ and calculates the value that maximizes (5.28) given the current values for all the normals $\mathbf{n_y}$ in its neighbourhood $\mathcal{N}_{\mathbf{x}}$. At the end of an iteration, the new values for each normal become the current values, and the next iteration begins. The algorithm is guaranteed to converge, and may be terminated according to a chosen criterion of convergence. The use of an ICM algorithm for the maximization of (5.28) here eliminates the need for discretizing the normal space. Furthermore, as it can be seen from equations (5.33) and (5.29) that the normals are not determined by a simple mean, but their estimation incorporates disparity information and the piecewise smooth assumption. This model, therefore, does not require an auxiliary step to find segments and plane-fits, thereby preventing the bias towards planar solutions, which was seen in discrete normal model.

3. For colour images $\mathbf{I}(\mathbf{x})$ represents the average of the RGB channels.

5.3 Overall optimization procedure

The resulting alternation procedure is the following: at iteration $t = 0$, all the normal field values are assumed to be $\{0, 0, 1\}$ and the first step (5.13) is performed to get a first estimate of the disparity map. That is we maximize $p(\mathbf{d}|\mathbf{n}^{(0)}, \mathbf{I})$ through standard BP or Mean Field. Then, denoting by $\mathbf{n}^{(t)}$ and $\mathbf{d}^{(t)}$ the current estimates of the normal and disparity fields, the two steps below are carried out alternately,

1) Update the normal field $\mathbf{n}^{(t)}$ into $\mathbf{n}^{(t+1)}$ by:
   - If discrete normal model described in section 5.2.2 is used perform the following steps :
     (i) segmenting the left image into small regions;
     (ii) computing for each of the obtained regions, the observed normals $\mathbf{p} = \{\mathbf{p_x}, \mathbf{x} \in \mathcal{S}\}$ from a plane fitting procedure using $\mathbf{d}^{(t)}$ and $\mathbf{I}$ on each of the obtained region; and
     (iii) updating the normal field using BP for the model defined in (5.26).
   - Else update normal field $\mathbf{n}^{(t)}$ into $\mathbf{n}^{(t+1)}$ by applying ICM on (5.28) described in section 5.2.3.
2) Update the disparity field $\mathbf{d}^{(t)}$ into $\mathbf{d}^{(t+1)}$ by:
   (i) computing the first order disparity derivatives using $\mathbf{n}^{(t+1)}$ and
   (ii) updating disparity estimates into $\mathbf{d}^{(t+1)}$ with BP or Mean Field applied to the conditional disparity model (5.15).

The described alternation steps are performed at multiple scales using the multi-grid approach described in chapter 4, section 4.3.2. We recall that this method allows for long range interaction to take place in few iterations. For both disparity as well as normals, the probabilities (messages in the case of BP) at the coarsest level of the grid hierarchy are used to initialize the next level and so on until the original grid. Furthermore, in both the cases (disparity and normals) the multi-grid optimization does not reduce the image resolution, but only the resolution of the computed cost. This is done by accumulating these costs over larger spatial neighbourhoods.

5.4 Experimental results

This section is divide into two parts; the first part compares the results of discrete normal model presented in section 5.2.2 and the model presented in section 5.2.3. We highlight the deficiencies of the discrete model and show the superiority of the non-discrete approach in the first part. The second part of the results section shows results pertinent only to the non-discrete model. We present results on both synthetic and real images. Before we discuss the two parts in detail, we briefly present the parameter settings used in the models.

Parameter settings :

The disparity range and therefore the value of L is fixed to different values depending on each image pair. The Alternation Maximization is carried out for a prescribed number of iterations (in our case we obtained good results with 5 iterations) at 4 different scales, ranging from coarse to fine. The other parameters are found heuristically as follows:

Parameters for disparity model: For the data term in (5.17): The window size W determines the support of neighbouring intensities for the given disparity values. Larger window can also be used as each of the pixels inside the window are weighted by colour and proximity (Yoon and Kweon [2007]). Even though the data term is calculated once overall the image and disparities, a larger window may take longer time to process. Therefore, in our case W is set to 5×5. The parameters γ_c and γ_g determine the influence of the neighbouring pixels within the window. The parameters γ_c and γ_g are set to 10 and 21 respectively. The T parameter of the robust function is fixed to the average pixel cost computed over all pixels and disparity labels. For the interaction term (5.22) the parameter σ_D determines the smoothness of the disparity map. As a robust formulation for interaction is not used a very large value of σ_D may lead to too much smoothing of the disparities across discontinuities. The value of σ_D is therefore set between 1.0 to 2.5. We consider a first order neighbourhood where each pixel has 8 neighbours for the interaction term (5.22). The Mean Field processes is performed until the total average energy change is less than 0.01 which corresponds to about $4 - 5$ iterations in practice.

Parameters for BP-based normal model: For normal initialization the image segmentation is performed using the Mean Shift algorithm of Comaniciu and Meer [2002] for both grey level and colour images, with range and distance sigma set to 6 and 5 respectively and a minimum region size of 80 pixels. It is to be noted that the segmentation is carried out once per scale. The plane fitting is performed using RANSAC (Fischler and Bolles [1987]) with 500 iterations and the maximum distance set to 0.3. The normal BP process is also performed for 5 iterations.

Parameters for ICM-based normal model: In (5.32) of the normal model (5.28), the σ_N parameter determines the influence of the neighbouring normals based on the disparity difference and magnitude of the image intensity gradient. If σ_N value is small, only those neighbouring normals which have similar disparities and image intensities will have an influence. This parameter mainly influences the smoothness of output normals, where a large value of σ_N allows the influence of most of the neighbours. The smoothness of the normal output in turn determines that of the disparity. For our experiments we found that setting σ_N to 1.2 provides good results. The ICM for normals was carried out for 20 iterations.

5.4.1 Comparing the performance of the two normal models

In order to compare the two algorithms for normal estimation we use two example images. The first example is the *corridor* and the second one, the *head* images.

We first begin our discussion with the *corridor* image (courtesy University of Bonn) (figure 5.6(a)). This stereo pair consists of images of the size 256 with the disparity range $[0, 11]$. While the ground-truth disparity is available (figure 5.6(b)), the normal "ground-truth" is generated using the `surfnorm` function in MATLAB. The `surfnorm` function computes surface normals for the surface defined by any X, Y, and Z, in our case these correspond the pixel coordinates for X and Y, and Z corresponds to the ground-truth disparity values. This function determines the surface normals based on a cross product of the tangent vectors. The tangent vectors are determined using the first difference between the neighbouring positions. In other words, tangent vectors are the gradients computed in the x and y-directions. As these gradients are not defined at discontinuities, we slightly smooth the ground-truth disparities with a Gaussian filter (with a standard deviation of 1.0). We use the normals, shown in figure 5.6(e), computed using this function as groundtruth for surface normals.

In figures. 5.6(c) and 5.6(d) we show the disparity maps obtained using Mean Field optimization and each of the two methods of normal estimation within alternation, namely ICM and BP respectively. We see that the disparity map obtained alternating with ICM-based normal estimation is much smoother compared to the one obtained using BP-based normal estimation. We further show the results of the normals obtained using ICM and BP as an arrow diagram in the the two figures 5.6(f) and 5.6(g). We see that the normals obtained using the ICM procedure are noisier compared to that of BP. However, it is difficult to compare the arrow diagrams as we do not completely see the difference in two results. We therefore convert the two normals into colour coded normal images.

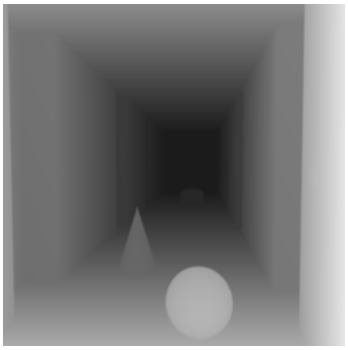
We use an HSV colour-code [4] to represent the normals. The colour coding is done as follows: Let $\mathbf{n_x} = (n_u, n_v, n_d)$ for $\mathbf{x} = (u, v)$ and disparity at the location $d_\mathbf{x} = d$. Now the colour code is obtained by mapping the azimuth θ and elevation ϕ of each normal to hue H and saturation S, respectively, and setting the value V to 1. The azimuth θ and elevation ϕ are computed in the standard way:

$$\theta = \pi + \arctan \frac{n_v}{n_u} \qquad \phi = \arccos n_d \tag{5.34}$$

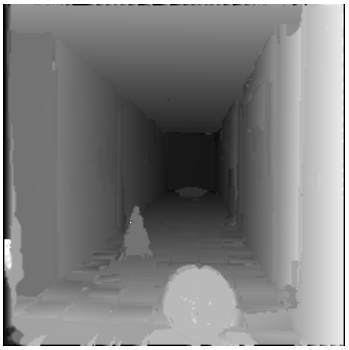
The range colours obtained using such a technique can be visualized by normals on a sphere. figure 5.7(a) shows spherical disparity surface. The normals obtained using `surfnorm` function in MATLAB is shown in figure 5.7(b) as an arrow diagram. The different directions associated with the sphere can be seen better on the flattened arrow map in figure 5.8(a). The figure 5.8(b) shows the range of colours obtained using the equation (5.34). It can be seen that the fronto-parallel normal is represented as white (middle of the sphere).

4. Note that the same colour code is used for all the other normal-maps

(a) Original Image of size 256 × 256

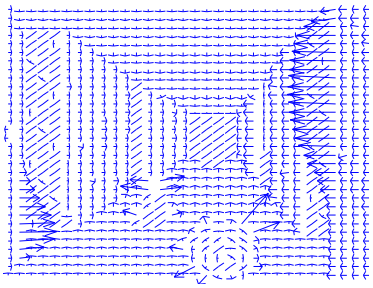


(b) Disparity Groundtruth

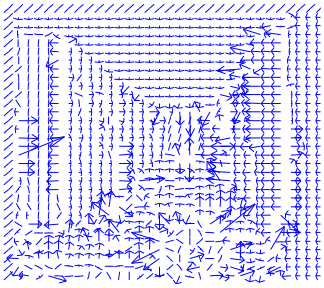


(c) Disparity estimated using Normals in figure 5.6(f) (below)

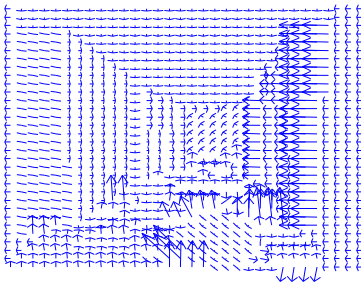
(d) Disparity estimated using Normals in figure 5.6(g) (below)



(e) Normals Groundtruth

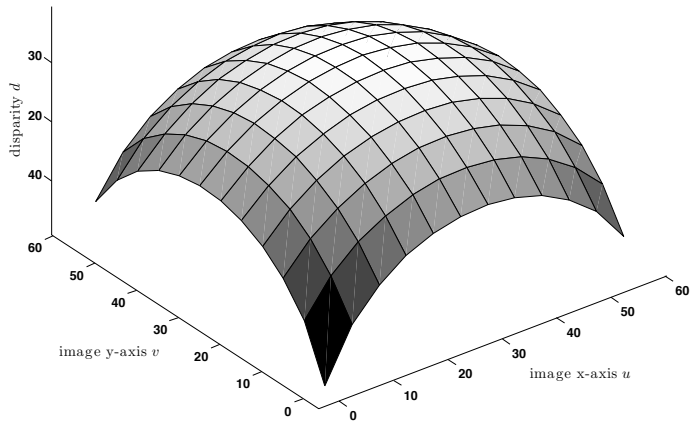


(f) Normal Estimation using ICM

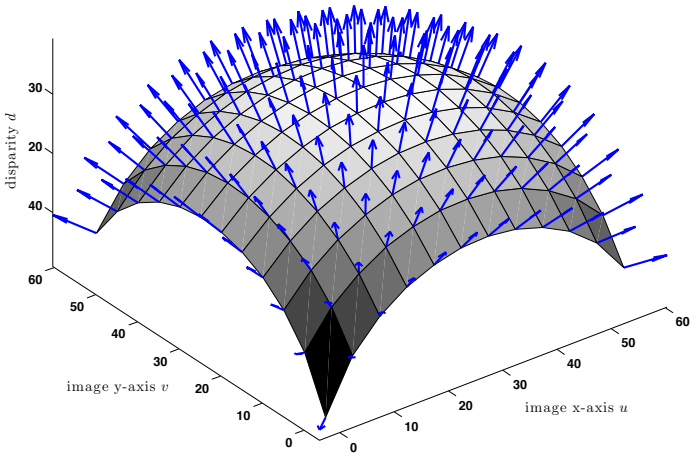


(g) Normal Estimation using BP

Figure 5.6: Disparity and Normals obtained for the *corridor* image using the ICM and BP for Normal estimation

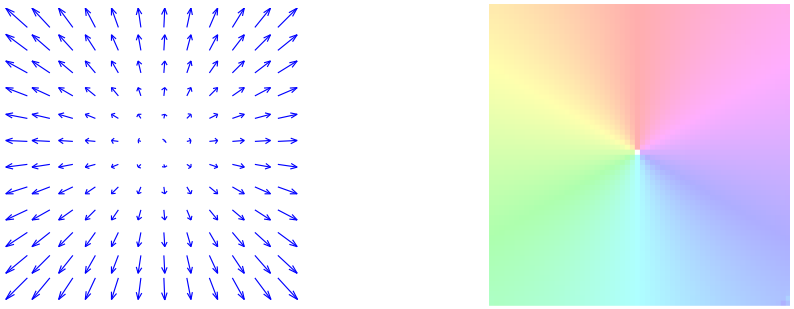


(a) Disparity Surface



(b) Normals as arrow map

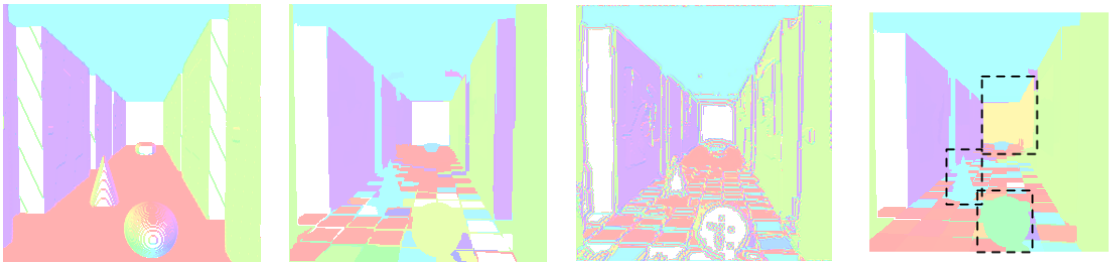
Figure 5.7: Example showing the normals for a spherical surface



(a) Flattened arrow map of figure 5.7(b)

(b) Colour Coded normals for figure 5.7(b)

Figure 5.8: Example showing the normals as colour for the spherical surface



(a) Normals Groundtruth

(b) Normals Initialization through plane-fitting

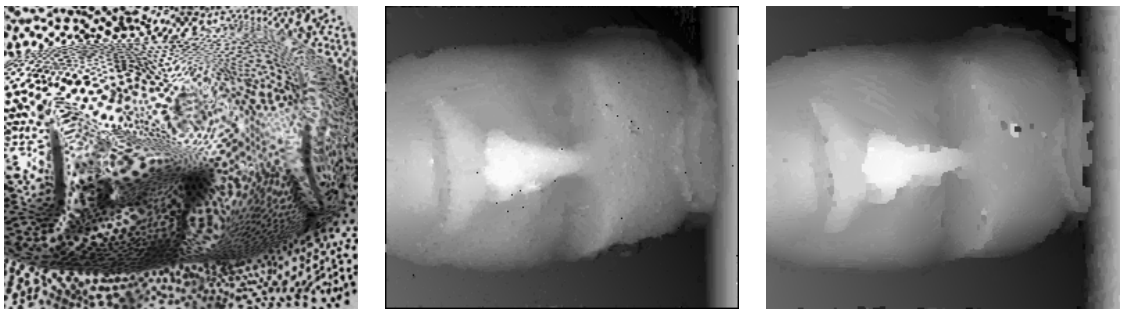
(c) ICM Normals

(d) BP Normals

Figure 5.9: Comparing the normals obtained for the *corridor* image using the two approaches to groundtruth. The normals are colour coded into HSV colours using (5.34). The highlighted regions in 5.9(d) shows how the normals obtained using BP follow the plane-fit solution in 5.9(b).

Now, using such a colour code for the normals in figures 5.6(e), 5.6(g) and 5.6(f), we compare the results obtained by BP and ICM procedures. We see that even though the normals obtained using the BP procedure are not noisy they have large errors, especially at the highlighted regions of the figure 5.9(d), as compared to the ground-truth in figure 5.6(e). Comparing the results of the normals obtained using BP figure 5.9(d) to its initialization obtained using segmentation/plane-fit figure 5.9(b), we see that the bias towards the plane fit solution is quite large. But this is natural because as the image is mainly made of planar surfaces the final solution tends to stay close to the plane-fit normals. We see that the normals obtained using ICM based approach (figure 5.9(c)) better approximates the groundtruth (figure 5.9(a)) as compared to the BP based solution.

We now compare results obtained on much smoother surface like the *head* image shown in



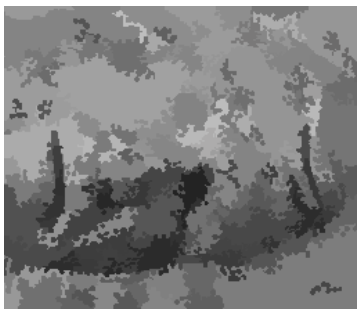
(a) Original Image of size 219×255 (b) Disparity estimated using Normals in figure 5.11(c) (c) Disparity estimated using Normals in figure 5.11(d)

Figure 5.10: Disparity estimated using Normals from BP and ICM

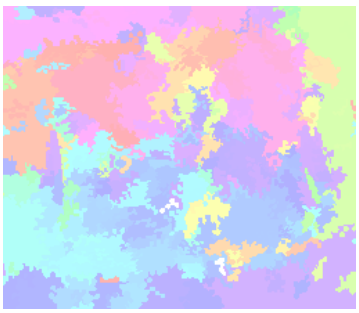
figure 5.10(a) (courtesy of University of Manchester and University of Sheffield). This image is of size 219×255 and has a disparity range of $[-30, 10]$. The figure 5.10 shows the results obtained using the normals obtained from ICM and BP optimizations. In the absence of groundtruth, disparity results obtained using the two solutions seems satisfactory. However, comparing the results of the normals using ICM and BP (figure 5.11(c) and 5.11(d)), we see that the results obtained using the BP approach do not conform to the surface variations. This discrepancy is mainly due to the discretization of the normal space. In absence of dense discretization the final solution either tends to the closest normal in discrete-normal space. As a result, if the surface is close to being fronto-parallel most solutions will tend to fronto-parallel solutions, $\mathbf{n_x} = (0, 0, 1)$ indicated by white in the figure 5.11(d). But increasing the discretization of the normal space makes the procedure computationally inefficient. Already, with $K = 162$ levels of discretization the BP procedure takes 60 seconds for a 32×32 image, as compared to ICM procedure which takes 1 second. Furthermore, this estimation procedure also depends on the size of the regions obtained during segmentation. This is because a certain minimum region size is required in order for the RANSAC procedure (used for plane-fitting in our approach) to provide a good plane representation of the region, and thus the initialization for the normal optimization. Due to these limitations of the discrete BP approach, we now concentrate on the ICM-based continuous normal estimation procedure.

5.4.2 Further results using ICM-based normal estimation

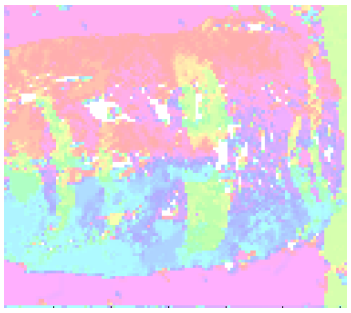
We now show some more results obtained through alternation of Mean Field-based disparity and ICM-based normal estimation. We show the results on *wood*, *cloth1* and *cloth2* images from the Middlebury database corresponding to figures 5.12, 5.13, and 5.14 respectively, with a disparity range of $[0, 60]$ pixels in all cases. In each of these figures, we show the left original image, the ground-truth disparity, estimated disparity using Mean



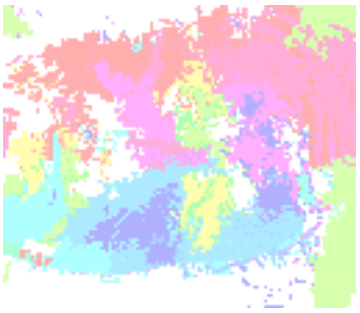
(a) Image Segmentation result



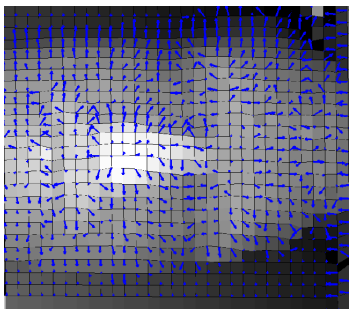
(b) Normals Initialization through plane-fitting colour coded



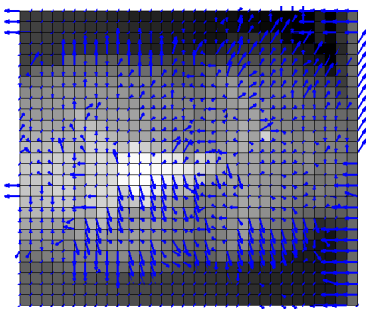
(c) Normal Estimation using ICM



(d) Normal Estimation using BP



(e) Normal Estimation using ICM



(f) Normal Estimation using BP

Figure 5.11: . The figure 5.11(b) shows the initial normal map obtained by plane-fitting disparities within the segments of figure 5.11(a). Disparity and Normals obtained for the *head* image using the BP and ICM for Normal estimation are shown in figures 5.11(d) and 5.11(c)

Field and estimated normals using ICM.

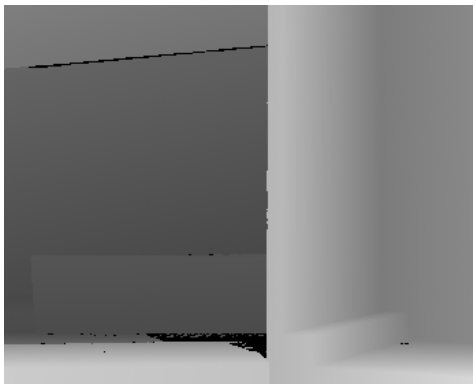
In order to obtain a quantitative comparison of the result, we compute the bad pixels error maps for *corridor*, *wood1*, *cloth1* and *cloth3* disparities. We use the Middlebury evaluation software by Scharstein and Szeliski [2002] to compute the this error map. This error is thresholded, that is all disparity differences which are less that 1.0 pixels are set to 0 and the rest set to 1. The error maps corresponding to each of these images are shown in figures 5.15(a), 5.15(b), 5.15(c) and 5.15(d) respectively. In figures 5.16 and 5.16, we show the error plots comparing our approach to standard BP along with sub-pixel interpolation (Scharstein and Szeliski [2002]). These plots obtained by varying the error threshold from 0.25 to 1.5. We see that our algorithm consistently performs well in all of the cases. It is to be noted that in each case the percentage error is calculated by ignoring the first L columns, where L stands for disparity range. From all the bad pixel maps we see that most errors occur at the disparity discontinuities which have not been explicitly handled in the algorithms.

5.5 Discussion

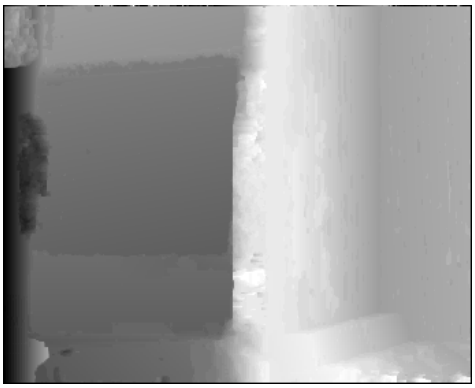
To summarize, we proposed a new joint probabilistic model with the following features:

- The proposed approach moves beyond the fronto-parallel assumption and estimates surface consistent disparity solutions.
- It embeds the estimation of surface properties in the model rather than refining the results using post-processing like Yang et al. [2009], Klaus et al. [2006].
- Unlike Devernay and Faugeras [1994], it does not require the direct computation of high-order disparity derivatives.
- The proposed approach does not precompute the normals as in Li and Zucker [2006b], instead uses a separate random field to estimate the normals based on the disparity and vice-versa.
- Two CRFs one for disparity and other for normals are defined and the geometric contextual information pertaining to the scene is included in each of these models.
- The alternating procedure results in mutual improvement of both disparities and normals.
- The consideration of two conditional models allows for more dependence or independence according to the information to be incorporated, thereby increasing flexibility.
- We propose two different models for normals; one which discretizes the normals space and optimizes the normal field using BP and; the other which uses an ICM based optimization and resides in the continuous space.
- We demonstrate that the BP normal estimation has several limitations:

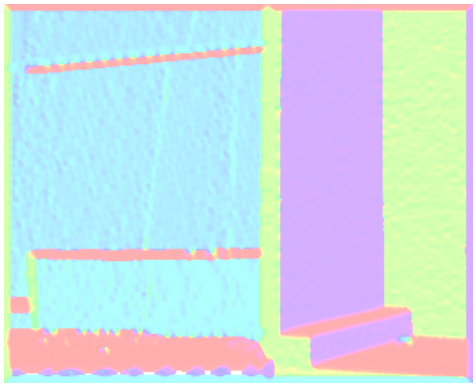
(a) Original Image of size 277×343



(b) Disparity Ground-truth



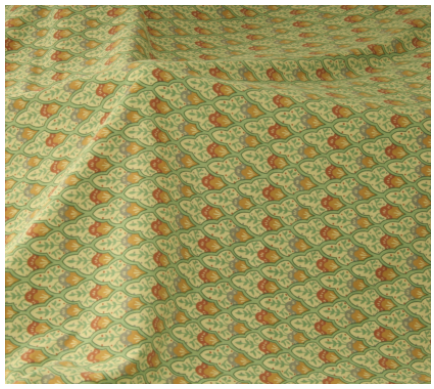
(c) Disparity estimated using Normals in figure 5.12(e)



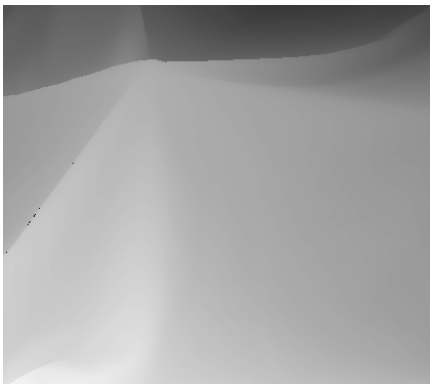
(d) Ground-truth Normals

(e) Normal Estimation using ICM

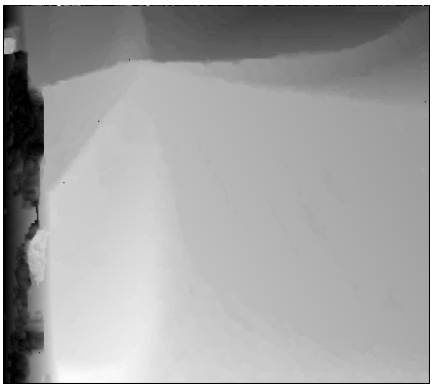
Figure 5.12: Disparity and Normals obtained for the *wood* image using the ICM for Normal estimation



(a) Original Image of size 370×417

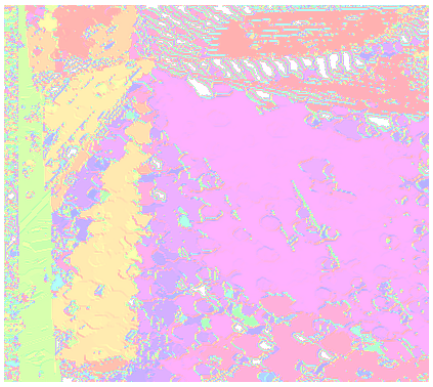


(b) Disparity Ground-truth



(c) Disparity estimated using Normals in figure 5.13(e)

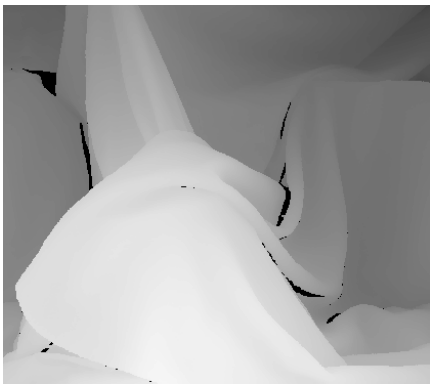
(d) Ground-truth Normals



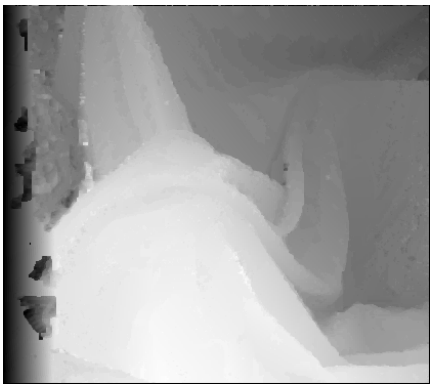
(e) Normal Estimation using ICM

Figure 5.13: Disparity and Normals obtained for the *wood* image using the ICM for Normal estimation

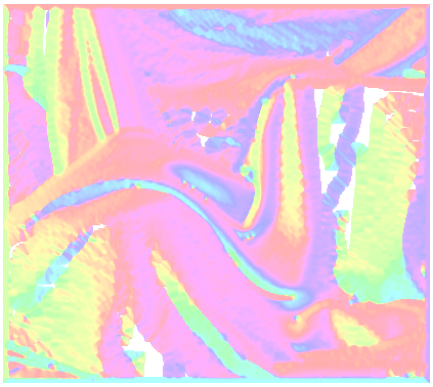
(a) Original Image of size 370×417



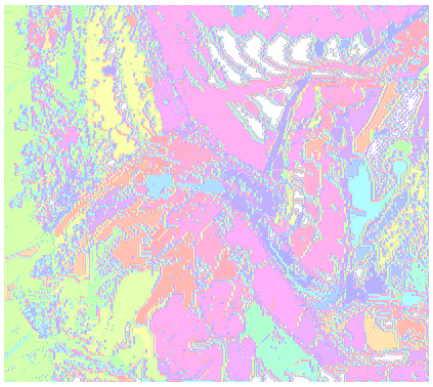
(b) Disparity Ground-truth with disparity range



(c) Disparity estimated using Normals in figure 5.14(e)

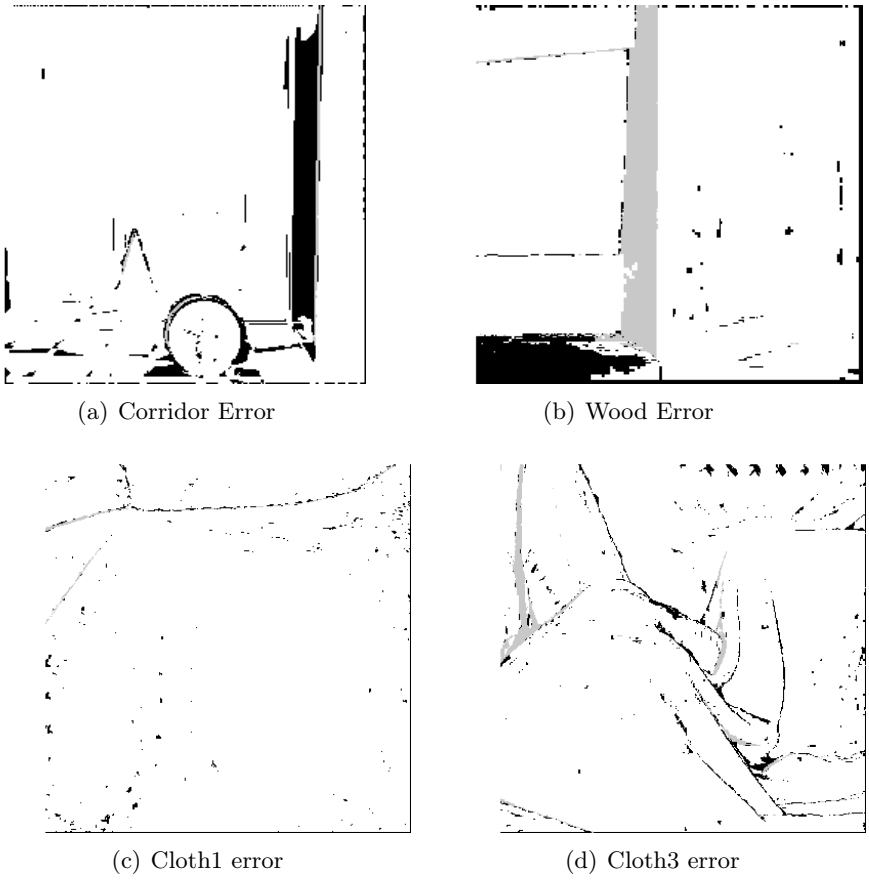


(d) Ground-truth Normals



(e) Normal Estimation using ICM

Figure 5.14: Disparity and Normals obtained for the *wood* image using the ICM for Normal estimation



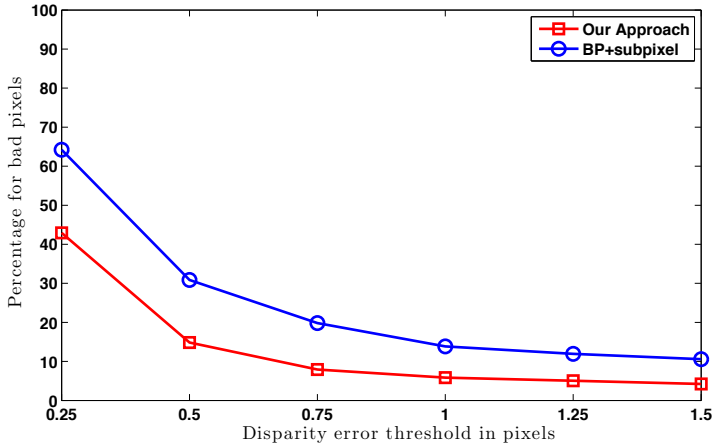
(a) Corridor Error

(b) Wood Error

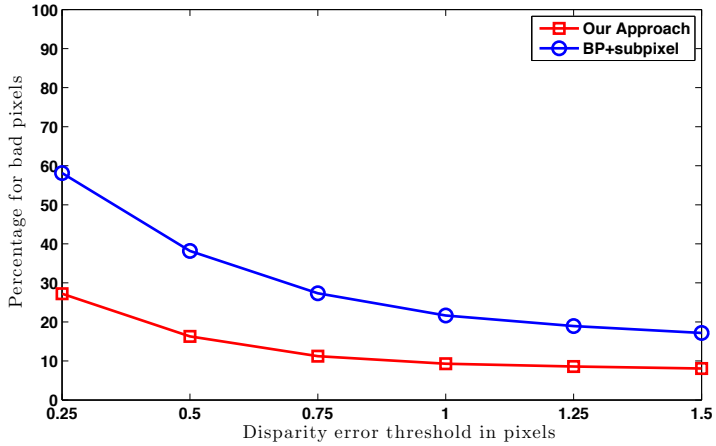
(c) Cloth1 error

(d) Cloth3 error

Figure 5.15: Bad pixels error for Disparity with a threshold of 1.0

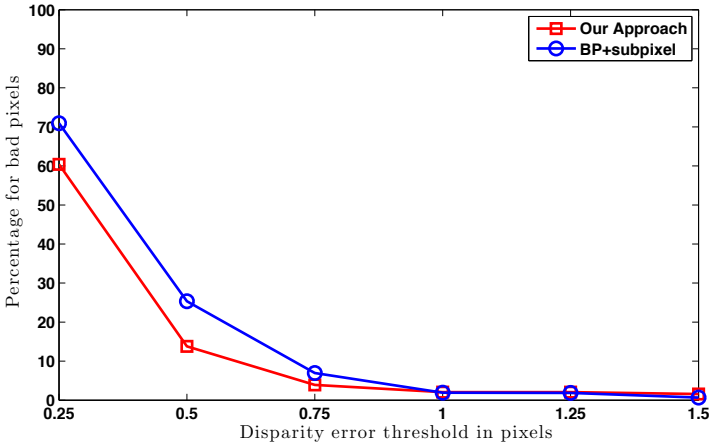


(a) Corridor Error Plot

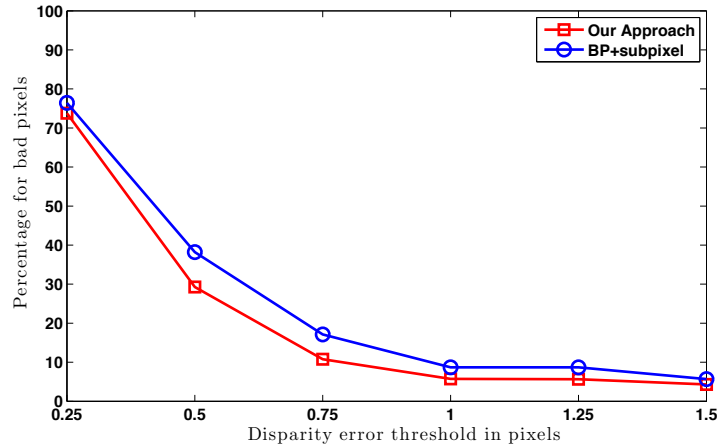


(b) Wood1 error plot

Figure 5.16: Bad pixels error plots for disparity maps of *corridor* and *Wood1* for error threshold ranging from 0.25 to 1.5



(a) Cloth1 error plot



(b) Cloth3 error plot

Figure 5.17: Bad pixels error plots for disparity maps of *Cloth1* and *Cloth3* for error threshold ranging from 0.25 to 1.5

  - it requires dense discretization of the normal space, which makes the optimization computationally inefficient.
  - requires an initialization for normals, which in our case is provided by using image segmentation and plane-fitting of disparities. This biases the final solution to the plane-fit normals.
  - the procedure is sensitive to the region size as the the plane-fitting procedure to obtain the initial normals requires enough number of points for a proper fit.
- We show that the ICM-based Normal estimation overcomes these limitations and gives better disparity as well as normal solutions.

We also see that in case of the disparity model, special modifications have to be made to the disparity optimization to allow for non-integer disparities. Although there are algorithms that allow direct estimation of sub-pixel disparities, like (Faugeras and Keriven [1998], Bhusnurmath and Taylor [2008], and Fleet et al. [1991]), they are all restrictive in their modelling of the energy functional. While Faugeras and Keriven [1998], and Bhusnurmath and Taylor [2008] require the energy functional to be convex (poor performance at object boundaries), Fleet et al.'s model is restrictive in terms of the maximum disparity range and is sensitive to image characteristics, such as textureless regions. Another recently introduced algorithm by Woodford et al. [2009] though allows for higher-order cliques resorts to image segmentation and disparity plane-fitting to generate real valued disparity maps. This being the case, an important yet difficult direction of research would be look for or develop a continuous optimization technique, which does not put too many restrictions on the formulation of the energy functional. Pock et al. [2008] suggest a technique to lift the non-convex energy functional to a higher dimension where it becomes convex and can therefore be optimized using convex programming. It would be interesting to see how our models can fit into such a framework.

Chapter 6

Conclusion

The problem of stereo matching can be summarized as - identifying the corresponding points in the left and the right images which are the projections of the same scene point. While the epipolar constraint reduces the search space, there are a number of ambiguities in matching two images (textureless regions, occlusions, discontinuities) which make the problem ill-posed. In this thesis, we focussed on Bayesian random Field techniques to handle ill-posedness of the stereo-matching problem. The motivation for using this technique was that instead of making a hard choice for a possible match, it relaxed the decision by the introduction of probabilities. Apart from being able to deal with uncertainties in stereo matching, it allows to incorporate explicit smoothness assumption with in the model. While the smoothness of the disparity output is important there are further constraints that should be included, for example, disparities should not be smoothed across object boundaries or the disparities should be consistent with geometric properties of the surface or regions with similar colour should have similar disparities. Our goal throughout this thesis was to incorporate such constraints using monocular cues and differential geometric information about the surface.

To this end, in this thesis we considered two important problems associated with stereo matching. The first was that of localizing disparity discontinuities. The second aimed to recover the binocular disparities in accordance with the surface properties of the scene under consideration. We presented a possible solution for each of these problems; In order to deal with disparity discontinuities, we proposed to cooperatively estimating disparities and object boundaries. This was motivated by the fact that the disparity discontinuities occur near object boundaries. The second method dealt with recovering surface consistent disparity and surface normal estimates by estimating the two simultaneously.

In the next section we provide a brief summary of the two models presented in this thesis. In section 6.2, we list the shared features of the two models and finally, conclude by mentioning some future directions of research in section 6.3.

6.1 Specific features of each of the proposed approaches

Features of the disparity boundary estimation

We carry out cooperatively both disparity and object boundary estimations by setting the two tasks in a unified Markovian framework. We define an original joint probabilistic model that allows us to estimate disparities through a coupled MRF model. Boundary estimation cooperates with disparity estimation to gradually and jointly improve accuracy of both the estimates. The feedback from boundary estimation to disparity estimation is made through the auxiliary field referred to as the displacement field. This field suggests the corrections that need to be applied at disparity discontinuities in order that they align with the object boundaries. The joint model is an MRF when considering disparities which reduces to a Markov chain when focusing on the displacement field. The features specific to this model are as follows:

- The coupled-MRF framework was proposed involving two MRFs one for the disparity and the other for the displacement field.
- The influence of the boundary estimation was encoded in the displacement field as directions in which disparity corrections needed to be applied.
- The core of this idea of applying corrections to the disparities at the discontinuities comes from the assumptions that the disparity continuities occur in the vicinity of the actual boundaries and that the depth discontinuities are in fact object boundaries.
- These corrections were incorporated in the disparity-MRF using the idea of active neighbourhood field, which is able to deal with non-standard neighbourhood systems.
- The displacement field was reduced to a second-order Markov chain which is active only at disparity discontinuities. This allows us to find the position of the true object boundary based on the corrections applied at the disparity discontinuities.
- The approximate inference of disparity was done using standard algorithms such as BP or Mean Field and exact inference of the displacement field using the Viterbi algorithm.
- The overall algorithm allowed for the simultaneous extraction of the object boundary and the disparities through use of a simple monocular cue, in this case the image gradient.

Features of the disparity normal estimation

The goal of this second algorithm is to recover binocular disparities in accordance with the surface properties of the scene under consideration. To do so, we estimate the disparity as well as the normals in the disparity space, by setting the two tasks in a unified framework. We defined a novel joint probabilistic model through two random fields to favour both intra field (within neighbouring disparities and neighbouring normals) and inter field (between

disparities and normals) consistency. Geometric contextual information is introduced in the models for both normals and disparities and then optimized using an appropriate alternating maximization procedure. The overall framework has the following features:

- A disparity and surface normal model, with the two variables modelled as CRFs, was proposed. These CRFs are coupled to incorporate the influence of each variable on the other.
- The two models were built under the assumption that the scene in question is made of piecewise smooth surfaces.
- The disparity-CRF was defined such that the interaction term involved first-order disparity derivatives, thereby enforcing the nearby disparities to lie on the same plane. These derivatives were extracted from the normal model.
- The optimization of the disparity-CRF was then carried out using standard Mean Field algorithm.
- Two models were presented for normals, one discrete and the other continuous.
- The discrete normal model involved discretization of normal space and was optimized using standard BP. However, this model require dense discretization of the normal space and therefore proved inefficient during optimization.
- The continuous model on the other hand provided a better alternative for the normal-CRF model. The model allowed for the extraction of the normals using the ICM algorithm.

6.2 Shared features of the two proposed approaches

The two approaches presented in this thesis share a number of common features, such as:

- We propose models which, in a probabilistic setting, allowed for conditional distributions that could model explicitly relationships between two variables.
- The two conditional distributions improved the flexibility of overall model in that they could be made dependent or independent according to the incorporated information.
- The use of the Alternation Maximization technique for optimization of the two fields results in mutual improvement of both variables involved.
- The use of multi-grid approach in optimization allows for long range interaction within a lattice, without reducing the resolution of the image, but only that of the costs.
- The proposed approaches have the further advantage of making a clear distinction between the probabilistic model and the subsequent optimization procedure. Separate and off-the-shelf optimization techniques can be used to infer variables associated with each of the random fields.

6.3 Further directions of research

The two models suggested in this thesis deal with the two important issues of disparity discontinuity localization and the extraction of disparity surfaces. The natural direction of research is to combine these two models so that we simultaneously correct disparities at the discontinuities, extract the disparity surface using surface geometric constraints and obtain object boundaries. This would involve dealing with three random fields, each of which influences the other in a different way. Given that surface geometric constraints enforce smoother disparities and the boundary model instigates discontinuities, the probabilistic interaction of the normal and boundary model with disparity, and vice-versa is not straight forward to model.

As for the probabilistic setting of the two proposed techniques, we focused on defining a valid unified framework to model cooperations, and used the MAP principle for inference. This model can be further investigated by recasting our approaches into an Expectation Maximization (EM) (Dempster et al. [1977]) like framework. For example, one way would be to develop a fully Bayesian model as in Scherrer et al. [2008], where the two interacting variables are set in an EM framework. The idea proposed by Scherrer et al. [2008] is as follows: suppose we have two sets of random variables $\mathbf{A}$ and $\mathbf{B}$ both of which are modelled as MRFs, with Θ representing the random variable for the parameters, and let $\mathbf{y}$ represent the observation information. Now the posterior distribution we are interested in maximizing is $p(\mathbf{a}, \mathbf{b}, \theta|\mathbf{y})$, where $\mathbf{a}, \mathbf{b}$ and θ are realizations of random variables $\mathbf{A}$, $\mathbf{B}$ and Θ respectively. In case of Scherrer et al., the E-step of the EM procedure is not performed exactly but using the alternation maximization procedure. So, the E-step involves alternation between the expectations of conditional probabilities, $p(\mathbf{a}|\mathbf{b}, \mathbf{y}, \theta)$ and $p(\mathbf{b}|\mathbf{a}, \mathbf{y}, \theta)$. The M-step finds the estimate of θ by maximizing $p(\theta|\mathbf{a}, \mathbf{b}, \mathbf{y})$. This kind of model provides alternatives in which, rather than using the the realizations of fields $\mathbf{A}$ and $\mathbf{B}$ for estimating one another, their full distributions could be used. As outlined above, such an EM-like framework also provides a good theoretically based parameter estimation procedure.

References

Alvarez, L., Deriche, R., Weickert, J., and Sànchez, J. (2000). Dense disparity map estimation respecting image discontinuities: A pde and scale-space based approach. In *International Workshop on Machine Vision Applications*. 30

Baker, H. and Binford, T. (1981). Depth from edge and intensity based stereo. In *International Joint Conference on Artificial Intelligence*, pages 631–636. vii, 11, 35

Balakrishnan, G., Sainarayanan, G., Nagarajan, R., and Yaacob, S. (2007). Wearable real-time stereo vision for the visually impaired. *Engineering Letters*, 14(2):6–14. vi, 4

Barnard, S. T. (1989). Stochastic stereo matching over scale. *International Journal of Computer Vision*, 3(1):17–32. 19

Bedini, L., Corso, G. M. D., and Tonazzini, A. (2001). Preconditioned edge-preserving image deblurring and denoising. *Pattern Recognition Letters*, 22(10):1083–1101. 47, 50, 51

Belhumeur, P. and Mumford, D. (1992). A Bayesian treatment of the stereo correspondence problem using half-occluded regions. In *Computer Vision and Pattern Recognition*, pages 506–512. 32, 33, 35

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, B*, 36:192–236. 22

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of Royal Statistical Society, B*, 48(3):259–302. 25, 46, 99

Bhusnurmath, A. and Taylor, C. J. (2008). Solving stereo matching problems using interior point methods. In *International Symposium on 3D Data Processing, Visualization and Transmission*, pages 321–329. 31, 39, 115

Birchfield, S. and Tomasi, C. (1999a). Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, 35(3):269–293. 20, 30

Birchfield, S. and Tomasi, C. (1999b). Multiway cut for stereo and motion with slanted surfaces. In *International Conference on Computer Vision*, pages 489–495. 38

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer. 22

Black, M. J. and Rangarajan, A. (1996). On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19(1):57–91. 21, 47

Blake, A. and Zisserman, A. (1987). *Visual Reconstruction*. MIT Press. 47

Bleyer, M. and Gelautz, M. (2004). A layered stereo algorithm using image segmentation and global visibility constraints. In *International Conference on Image Processing*, pages 2997–3000. 36

Bobick, A. F. and Intille, S. S. (1999). Large occlusion stereo. *International Journal of Computer Vision*, 33(3):181–200. 17, 32, 35

Bolles, R. C. and Woodfill, J. I. (1993). Spatiotemporal consistency checking of passive range data. In *International Symposium on Robotics Research*. 32, 37

Bouguet, J.-Y. (2008). Camera calibration toolbox for Matlab. 5

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via Graph Cuts. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 23(11):1222–1239. 19, 20, 25, 29, 47, 55

Burt, P. and Adelson, E. (1983). The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540. 72

Burt, P. and Julesz, B. (1980). Disparity-gradient limit for binocular fusion. *Science*, 208(4444):615–617. 18

Chandler, D. and Percus, J. K. (1988). Introduction to Modern Statistical Mechanics. *Physics Today*, 41. 25

Chang, J. Y., Lee, K. M., and Lee, S. U. (2007). Stereo matching using iterative reliable disparity map expansion in the color-spatial-disparity space. *Pattern Recognition*, 40(12):3705–3713. 36

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 24(5):603–619. 36, 37, 101

Cumming, B. G. and DeAngelis, G. C. (2001). The physiology of stereopsis. *Annual Review of Neuroscience*, 24(1):203–238. 4

Cutting, D. R., Kupiec, J., Pedersen, J. O., and Sibun, P. (1992). A practical part-of-speech tagger. In *Applied Natural Language Processing*, pages 133–140. 70

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society, Series B*, 39(1):1–38. xiii, 120

Devernay, F. and Faugeras, O. (1994). Computing differential properties of 3-D shapes from stereoscopic images without 3-D models. In *Computer Vision and Pattern Recognition*, pages 208–213. 38, 39, 85, 88, 90, 96, 108

Do Carmo, M. (1976). *Differential Geometry of Curves and Surfaces*. Prentice-Hall. 87

Faugeras, O. (1993). *Three-dimensional computer vision: a geometric viewpoint*. MIT Press, Cambridge, MA, USA. 6, 8

Faugeras, O. D. and Keriven, R. (1998). Complete dense stereovision using level set methods. In *European Conference on Computer Vision*, pages 379–393. 31, 115

Felzenszwalb, P. and Huttenlocher, D. (2006). Efficient Belief Propagation for early vision. *International Journal of Computer Vision*, 70(1):41–54. 20, 28, 29, 37, 47, 60, 71, 76, 82

Fischler, M. A. and Bolles, R. C. (1987). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In *Readings in computer vision: issues, problems, principles, and paradigms*, pages 726–740. Morgan Kaufmann Publishers Inc. 37, 101

Fleet, D. J., Jepson, A. D., and Jenkin, M. R. M. (1991). Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210. 31, 115

Forbes, F. and Fort, G. (2007). Combining Monte Carlo and Mean-Field-Like Methods for Inference in Hidden Markov Random Fields. *IEEE Transactions on Image Processing*, 16:824–837. 25

Fusiello, A., Trucco, E., and Verri, A. (2000). A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12(1):16–22. 8

Gamble, E. and Poggio, T. (1987). Visual integration and detection of discontinuities: The key role of intensity edges. Technical report, Massachusetts Institute of Technology. 20, 45, 46, 50, 51

Geiger, D. and Girosi, F. (1991). Parallel and deterministic algorithms from MRFs: Surface reconstruction. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 13(5):401–412. 45, 48, 51, 57

Geiger, D., Ladendorf, B., and Yuille, A. L. (1995). Occlusions and binocular stereo. *International Journal of Computer Vision*, 14(3):211–226. 32, 33, 35

Geiger, D. and Yuille, A. (1991). A common framework for image segmentation. *International Journal of Computer Vision*, 6(3):227–243. 25, 45, 51

Geman, D., Geman, S., Graffigne, C., and Dong, P. (1990). Boundary detection by constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):609–628. 56

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741. 22, 25, 44, 48

Goldberg, S., Maimone, M., and Matthies, L. (2002). Stereo vision and rover navigation software for planetary exploration. In *Aerospace Conference Proceedings, 2002. IEEE*, volume 5, pages 2025–2036. v, 4

Greig, D. M., Porteous, B. T., and Seheult, A. H. (1989). Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51(2):271–279. 29

Günsel, B., Jain, A. K., and Panayirci, E. (1996). Reconstruction and boundary detection of range and intensity images using multiscale MRF representations. *Computer Vision and Image Understanding*, 63(2):353–366. 46, 50

Hannah, M. J. (1974). *Computer matching of areas in stereo images.* PhD thesis, Stanford University. vii, 11

Hartley, R. and Zisserman, A. (2000). *Multiple view geometry in computer vision.* Cambridge University Press, New York, NY, USA. 6, 8, 91

Heitz, F. and Bouthemy, P. (1993). Multimodal estimation of discontinuous optical flow using Markov Random Fields. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 15(12):1217–1232. 46, 48, 50, 51, 56, 57

Helmholtz, H. v. (1867/1925). *Handbook of Physiological Optics*, volume III. Optical Society of America, New York. Translated by J. P. C. Southall. 4

Hirschmüller, H. (2008). Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 30(2):328–341. 32, 34

Hong, L. and Chen, G. (2004). Segment-based stereo matching using Graph Cuts. In *Computer Vision and Pattern Recognition*, pages 74–81. 36

Hua, G. and Wu, Y. (2006). Sequential Mean Field variational analysis of structured deformable shapes. *Computer Vision Image Understanding*, 101(2):87–99. 25

Ishikawa, H. (2003). Exact optimization for Markov Random Fields with convex priors. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 25(10):1333–1336. 29, 34

Jaakkola, T. (2000). Tutorial on variational approximation methods. *MIT Artificial Intelligence Laboratory 545 Technology Square Cambridge, MA 02139*. 25, 26

Julesz, B. (1971). *Foundations of Cyclopean Perception*. University of Chicago Press. 4

Julez, B. (1959). A method of coding tv signals based on edge detection. *Bell System Tech.*, 38(4):1001–1020. 4

Kanade, T. (1994). Development of a video-rate stereo machine. In *1994 ARPA Image Understanding Workshop*. 16

Kanade, T. and Okutomi, M. (1994). A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 16(9):920–932. 17

Kirkpatrick, S., Gelatt, C. D., Jr., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220:671–680. 25, 44

Klaus, A., Sormann, M., and Karner, K. F. (2006). Segment-based stereo matching using Belief Propagation and a self-adapting dissimilarity measure. In *International Conference on Pattern Recognition*, pages 15–18. 20, 36, 108

Kolmogorov, V. (2006). Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 28(10):1568–1583. 28

Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., and Rother, C. (2006). Probabilistic fusion of stereo with color and contrast for bilayer segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1480–1492. 30

Kolmogorov, V. and Zabih, R. (2001). Computing visual correspondence with occlusions via Graph Cuts. In *International Conference on Computer Vision*, pages 508–515. 20, 25, 29, 34

Kolmogorov, V. and Zabih, R. (2002a). Multi-camera scene reconstruction via Graph Cuts. In *European Conference on Computer Vision*, pages 82–96. 29

Kolmogorov, V. and Zabih, R. (2002b). What energy functions can be minimized via Graph Cuts? In *European Conference on Computer Vision*, pages 65–81. 29, 39

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *IEEE International Conference on Machine Learning*, pages 282–289. 39, 92

Le Hégarat-Mascle, S., Kallel, A., and Descombes, X. (2007). Ant Colony Optimization for image regularization based on a nonstationary Markov modeling. *IEEE Transactions on Image Processing*, 16(3):865–78. 64, 82

Li, G. and Zucker, S. (2006a). Differential geometric consistency extends stereo to curved surfaces. In *European Conference on Computer Vision*, pages 44–57. 38, 39, 40

Li, G. and Zucker, S. (2006b). Surface geometric constraints for stereo in Belief Propagation. In *Computer Vision and Pattern Recognition*, pages 2355–2362. 38, 39, 86, 87, 88, 90, 92, 93, 94, 95, 96, 108

Li, G. and Zucker, S. W. (2010). Differential geometric inference in surface stereo. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 32(1):72–86. 86, 87

Li, S. Z. (2001). *Markov Random Field modeling in image analysis*. Springer-Verlag New York, Inc. 22

Lin, M. and Tomasi, C. (2004). Surfaces with occlusions from layered stereo. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 26(8). 38, 85

Loop, C. T. and Zhang, Z. (1999). Computing rectifying homographies for stereo vision. In *Computer Vision and Pattern Recognition*, pages 1125–1131. 8

Marr, D. and Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, 194(4262):283–287. 20, 31

Marr, D. and Poggio, T. (1979). A Computational Theory of Human Stereo Vision. *Royal Society of London Proceedings Series B*, 204:301–328. vii, 11

Marroquin, J. L. (1984). Surface reconstruction preserving discontinuities. Technical report, Massachusetts Institute of Technology. 44, 45, 50, 51, 57

Matthies, L., Kanade, T., and Szeliski, R. (1989). Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3(3):209–238. 16

Medrano, C., Herrero, J. E., Martínez, J., and Orrite, C. (2009). Mean Field approach for tracking similar objects. *Computer Vision Image Understanding*, 113(8):907–920. 25

Miled, W. and Pesquet, J. C. (2006). Disparity map estimation using a total variation bound. In *Canadian Conference on Computer and Robot Vision*, page 48. 31

Min, D. and Sohn, K. (2008). Cost aggregation and occlusion handling with WLS in stereo matching. *IEEE Transactions on Image Processing*, 17(8):1431–1442. 34

Murino, V. and Trucco, A. (1998). Edge/Region-based segmentation and reconstruction of underwater acoustic images by Markov Random Fields. In *Computer Vision and Pattern Recognition*, pages 408–413. 46, 50, 51

Narasimha, R., Arnaud, E., Forbes, F., and Horaud, R. P. (2008). Cooperative disparity and object boundary estimation. In *IEEE International Conference on Image Processing*. viii, 13, 56

Narasimha, R., Arnaud, E., Forbes, F., and Horaud, R. P. (2009). A joint framework for disparity and surface normal estimation. Research Report RR-7090, INRIA Rhone-Alpes. ix, 13, 87

Narasimha, R., Arnaud, E., Forbes, F., and Horaud, R. P. (2010). Disparity and normal estimation through alternating maximization. In *IEEE International Conference on Image Processing*. ix, 13, 87

Nasrabadi, N. M., Clifford, S. P., and Liu, Y. (1989). Integration of stereo vision and optical flow by using an energy-minimization approach. *Journal Optical Society of America (A)*, 6(6):900–907. 49

Ohta, Y. and Kanade, T. (1985). Stereo by intra- and inter-scanline search using Dynamic Programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(1):139–154. 30, 35, 36

Page, D. L., Sun, Y., Koschan, A., Paik, J. K., and Abidi, M. A. (2002). Normal vector voting: Crease detection and curvature estimation on large, noisy meshes. *Graphical Models*, 64(3-4). 92, 98

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288. 27

Pock, T., Schoenemann, T., Graber, G., Bischof, H., and Cremers, D. (2008). A convex formulation of continuous multi-label problems. In *European Conference on Computer Vision*, pages 792–805. 31, 115

Poggio, T., Torre, V., and Koch, C. (1985). Computational vision and regularization theory. *Nature*, 317:314–319. 20

Pollard, S., Mayhew, J., and Frisby, J. (1985). PMF: A stereo correspondence algorithm using a disparity gradient limit. *Perception*, 14:449–470. vii, 11, 17

Prazdny, K. (1985). Detection of binocular disparities. *Biological Cybernetics*, 52:93–99. 18

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285. 70

Ramachandran, V. S. and Nelson, J. I. (1976). Global grouping overrides point-to-point disparities. *Perception*, 5(2):125 Ð 128. 4

Rother, C., Kolmogorov, V., Lempitsky, V. S., and Szummer, M. (2007). Optimizing binary MRFs via extended roof duality. In *Computer Vision and Pattern Recognition*. 39

Ryan, T., R., G., and B., H. (1980). Prediction of correlation errors in stereo images. *Optical Engineering*, 19(3):312–322. 16

Sanger, T. (1988). Stereo disparity computation using gabor filters. *Biological Cybernetics*, 59:405–418. 31

Scharstein, D. and Szeliski, R. (1998). Stereo matching with nonlinear diffusion. *International Journal of Computer Vision*, 28(2):155–174. 19

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42. 108

Scherrer, B., Forbes, F., Garbay, C., and Dojat, M. (2008). Fully Bayesian joint model for MR brain scan tissue and structure segmentation. In *Medical Image Computing and Computer-Assisted Intervention, Part II*, pages 1066–1074. 120

Smith, B., Zhang, L., and Jin, H. (2009). Stereo matching with nonparametric smoothness priors in feature space. In *Computer Vision and Pattern Recognition*, pages 485–492. 39, 85

Strecha, C., Fransens, R., and Van Gool, L. (2006). Combined depth and outlier estimation in multi-view stereo. In *Computer Vision and Pattern Recognition*, pages 2394–2401. 20, 25, 26, 34, 47

Sudhir, G., Banerjee, S., Biswas, K. K., and Bahl, R. (1995). Cooperative integration of stereopsis and optic flow computation. *Journal of Optical Society of America (A)*, 12(12):2564–2572. 49, 50, 51

Sun, J., Li, Y., and Kang, S. B. (2005). Symmetric stereo matching for occlusion handling. In *Computer Vision and Pattern Recognition*, pages 399–406. 20, 32, 34, 36, 37, 49, 50, 51, 85

Sun, J., Zheng, N., and Shum, H. (2003). Stereo matching using Belief Propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787–800. 20, 25, 28, 36, 47, 50, 51, 55

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., and Rother, C. (2008). A comparative study of energy minimization methods for Markov Random Fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 30(6):1068–1080. 28, 29, 31

Tao, H., Sawhney, H., and Kumar, R. (2001). A global matching framework for stereo computation. In *International Conference on Computer Vision*, pages 532–539. 36

Tappen, M. F. and Freeman, W. T. (2003). Comparison of Graph Cuts with Belief Propagation for stereo, using identical MRF parameters. In *International Conference on Computer Vision*, page 900. 29

Tonazzini, A., Bedini, L., and Salerno, E. (2006). A Markov model for blind image separation by a mean-field EM algorithm. *IEEE Transactions on Image Processing*, 15(2):473–482. 25

Torr, P. H. S. and Criminisi, A. (2004). Dense stereo using pivoted Dynamic Programming. *Image and Vision Computing*, 22(10):795–806. 30

Wainwright, M. P. and Jordan, M. I. (2005). A variational principle for graphical models. In *New Directions in Statistical Signal Processing: From Systems to Brain*. MIT Press. 25, 26

Wang, Z.-F. and Zheng, Z.-G. (2008). A region based stereo matching algorithm using cooperative optimization. In *Computer Vision and Pattern Recognition*. 37

Weiss, Y. (2001). Comparing the Mean Field method and Belief Propagation for approximate inference in MRFs. In Opper, M. and Saad, D., editors, *Advanced Mean Field methods: theory and practice*. 28

Wheatstone, C. (1838). Contributions to the Physiology of Vision. Part the First. On Some Remarkable, and Hitherto Unobserved, Phenomena of Binocular Vision. *Philosophical Transactions of the Royal Society of London*, 128:371–394. 1, 2, 3

Wheatstone, C. (1852). The Bakerian Lecture: Contributions to the Physiology of Vision. Part the Second. On Some Remarkable, and Hitherto Unobserved, Phenomena of Binocular Vision (continued). *Philosophical Transactions of the Royal Society of London*, 142:1–17. 1

Woodford, O., Torr, P., Reid, I., and Fitzgibbon, A. (2009). Global stereo reconstruction under second-order smoothness priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12). 39, 85, 115

Wu, J. and Chung, A. (2007). A segmentation model using compound Markov Random Fields based on a boundary model. *IEEE Transactions on Image Processing*, 16(1):241–252. 48, 50, 51, 56, 57

Xia, Y., Feng, D., and Zhao, R. (2006). Adaptive segmentation of textured images by using the coupled-Markov Random Field model. *Image Processing, IEEE Transactions on*, 15(11):3559 –3566. 48, 50

Xu, L. and Jia, J. (2008). Stereo matching: An outlier confidence approach. In *European Conference on Computer Vision*, pages 775–787. 20, 32, 34, 37, 85

Xue, J., Zheng, N., Geng, J., and Zhong, X. (2008). Tracking multiple visual targets via particle-based Belief Propagation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 38(1):196–209. 47, 50, 51

Yang, Q., Wang, L., Yang, R., Stewénius, H., and Nistér, D. (2009). Stereo matching with color-weighted correlation, hierarchical Belief Propagation, and occlusion handling. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 31(3):492–504. 20, 32, 34, 37, 85, 108

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2003). Understanding Belief Propagation and its generalizations. In *Exploring artificial intelligence in the new millennium*, pages 239–269. Morgan Kaufmann Publishers Inc. 25, 28

Yoon, K.-J. and Kweon, I.-S. (2007). Stereo matching with the distinctive similarity measure. In *International Conference on Computer Vision*, pages 1–7. 17, 37, 93, 101

Yuille, A. L., Geiger, D., and Bülthoff, H. (1990). Stereo integration, Mean Field theory and psychophysics. In *European Conference on Computer Vision*, pages 73–82. 20, 25, 26, 45

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transctions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334. 6

Zitnick, C. and Kanade, T. (2000). A cooperative algorithm for stereo matching and occlusion detection. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 22(7):675–684. 31, 34

Zitnick, C. L. and Kang, S. B. (2007). Stereo for image-based rendering using image oversegmentation. *International Journal of Computer Vision*, 75(1):49–65. 37, 85

Résumé

La profondeur des objets dans la scène 3-D peut être récupérée à partir d'une paire d'images stéréo en trouvant des correspondances entre les deux points de vue. Cette tâche consiste à identifier les points dans les images gauche et droite, qui sont les projections du même point de la scène. La différence entre les emplacements des deux points correspondants est la disparité, qui est inversement proportionnelle à la profondeur 3D. Dans cette thèse, nous nous concentrons sur les techniques Bayésiennes qui contraignent les estimations des disparités en appliquant des hypothèses de lissage explicites. Cependant, il ya des contraintes supplémentaires qui doivent être incluses, par exemple, les disparités ne doivent pas être lissées au travers des bords des objets, les disparités doivent être compatibles avec les propriétées géométriques de la surface. L'objectif de cette thèse est d'intégrer ces contraintes en utilisant des informations monoculaires et des informations géométrique différentielle sur la surface. Dans ce but, cette thèse considère deux problèmes importants associés à stéréo : le premier est la localisation des discontinuités des disparités et le second vise à récupérer les disparités binoculaires en conformité avec les propriétés de surface de la scène. Afin de faire face aux discontinuités des disparités, nous nous proposons d'estimer conjointement les disparités et les frontières des objets. Cette démarche est motivée par le fait que les discontinuités des disparités se trouvent à proximité des frontières des objets. La seconde méthode consiste à contraindre les disparités pour qu'elles soient compatibles avec la surface et les normales à la surface en estimant les deux en même temps.

Mots clés: *Stéréo-vision, Champs de Markov, Estimation de la disparité, l'inférence Bayésienne*

Abstract

The depth of objects in 3-D scene can be recovered from a stereo image-pair by finding correspondences between the two views. This stereo matching task involves identifying the corresponding points in the left and the right images, which are the projections of the same scene point. The difference between the locations of the two corresponding points is the disparity, which is inversely related to the 3-D depth. In this thesis, we focus on Bayesian techniques that constrain the disparity estimates. In particular, these constraints involve explicit smoothness assumptions. However, there are further constraints that should be included, for example, the disparities should not be smoothed across object boundaries, the disparities should be consistent with geometric properties of the surface, and regions with similar colour should have similar disparities. The goal of this thesis is to incorporate such constraints using monocular cues and differential geometric information about the surface.To this end, this thesis considers two important problems associated with stereo matching; the first is localizing disparity discontinuities and second aims at recovering binocular disparities in accordance with the surface properties of the scene under consideration. We present a possible solution for each these problems. In order to deal with disparity discontinuities, we propose to cooperatively estimating disparities and object boundaries. This is motivated by the fact that the disparity discontinuities occur near object boundaries. The second one deals with recovering surface consistent disparities and surface normals by estimating the two simultaneously.

Keywords: *Stereo Vision, Markov Random Field, Disparity Estimation, Bayesian Inference*