University of Crete

Department of Computer Science

Visual Detection of Independent 3D Motion by a Moving Observer

Ph.D Thesis

Antonis A. Argyros

Heraklion, October 1996

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

Οπτική Ανίχνευση Ανεξάρτητης Τρισδιάστατης Κίνησης από Κινούμενο Παρατηρητή

Διατριβή που υποβλήθηκε από τον

Αντώνη Α. Αργυρό

ως μερική απαίτηση για την απόκτηση του

ΔΙΔΑΚΤΟΡΙΚΟΥ ΔΙΠΛΩΜΑΤΟΣ

Οκτώβριος 1996

Συγγραφέας:

Αντώνης Α. Αργυρός

Τμήμα Επιστήμης Υπολογιστών

Επταμελής Εξεταστική Επιτροπή:

Στέλιος Ορφανουδάκης, Καθηγητής, Επόπτης

Γιάννης Παπαδάκης, Καθηγητής, Μέλος

Γιάννης Πίτας, Καθηγητής, Μέλος

Στέφανος Κόλλιας, Αναπληρωτής Καθηγητής, Μέλος

Χρήστος Νικολάου, Αναπληρωτής Καθηγητής, Μέλος

Γιώργος Τζιρίτας, Αναπληρωτής Καθηγητής, Μέλος

Πάνος Τραχανιάς, Επίκουρος Καθηγητής, Μέλος

Δεκτή:

Πάνος Κωνσταντόπουλος, Αναπληρωτής Καθηγητής

Πρόεδρος Επιτροπής Μεταπτυχιακών Σπουδών

Ευχαριστίες

Η διατριβή αυτή ολοκληρώθηκε κάτω από την καθοδήγηση και εποπτεία του σύμβουλου καθηγητή μου, κ. Στέλιου Ορφανουδάκη, με τον οποίο είχα μία εξαιρετική συνεργασία σε όλη την διάρκεια των σπουδών μου. Θα ήθελα να τον ευχαριστήσω θερμά για την βοήθειά του στο επιστημονικό επίπεδο, την υποστήριξη και την εμπιστοσύνη του προς το πρόσωπό μου, καθώς και για την πάντα άψογη συμπεριφορά του στο επίπεδο των προσωπικών, ανθρώπινων σχέσεων.

Αισθάνομαι επίσης την ανάγκη να ευχαριστήσω θερμά τον καθηγητή του Πανεπιστημίου του Maryland κ. Γιάννη Αλοίμονο και την ερευνήτρια του ίδιου πανεπιστημίου κ. Cornelia Fermüller με τους οποίους είχα εξαιρετικά ενδιαφέρουσες συζητήσεις για την Ενεργό και Τελεολογική Όραση (active and purposive vision) καθώς επίσης και πιό συγκεκριμένα, για το πρόβλημα της ανίχνευσης ανεξάρτητης τρισδιάστατης κίνησης. Σίγουρα, η επαφή μαζί τους μου έδωσε πολλά και ιδιαίτερα χρήσιμα μαθήματα.

Θα ήθελα επίσης να ευχαριστήσω τα μέλη της εξεταστικής επιτροπής της διατριβής μου, τους κυρίους Γιάννη Παπαδάκη (καθηγητή του Μαθηματικού Τμήματος και Κοσμήτορα της Σχολής Θετικών Επιστημών του Πανεπιστημίου Κρήτης), Γιάννη Πίτα (καθηγητή του Τμήματος Πληροφορικής του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης), Στέφανο Κόλλια (αναπληρωτή καθηγητή του Τμήματος Ηλεκτρολόγων Μηχανικών του Μετσόβειου Πολυτεχνείου), Χρήστο Νικολάου (αναπληρωτή καθηγητή του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης), Γιώργο Τζιρίτα (αναπληρωτή καθηγητή του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης) και Πάνο Τραχανιά (επίκουρο καθηγητή του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης). Θα ήθελα να σταθώ ιδιαίτερα στον Πάνο Τραχανιά ο οποίος κατά την διάρκεια των τελευταίων και κρισιμότερων χρόνων υπήρξε πάντα εξαιρετικά πρόθυμος να ακούσει, να

συζητήσει και να με βοηθήσει σε όλες τις προσπάθειές μου. Θα ήθελα επίσης θερμά να ευχαριστήσω τον κ. Γιώργο Τζιρίτα στον οποίο οφείλω την αρχική μου επαφή με τα στατιστικά εργαλεία που εκτενώς χρησιμοποίησα στην εργασία μου. Ιδιαίτερα τον ευχαριστώ και για τα χρήσιμα σχόλιά του στο τελικό κείμενο της διατριβής.

Ιδιαίτερη μνεία αξίζει στον συνάδελφο μου, Μανόλη Λουράκη με τον οποίο μοιραστήκαμε ατέλειωτες ώρες συζητήσεων, πειραμάτων, ενθουσιασμού και απογοητεύσεων. Θεωρώ τον εαυτό μου πραγματικά τυχερό που σε ένα τόσο εξαιρετικό συνεργάτη βρήκα ένα τόσο σπάνιο φίλο. Του εύχομαι ολόψυχα ό,τι επιθυμεί.

Η εργασία αυτή ολοκληρώθηκε στο Τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης και στο Ινστιτούτο Πληροφορικής του Ιδρύματος Τεχνολογίας και Έρευνας. Θα ήθελα να ευχαριστήσω θερμά όλους τους φίλους και συνεργάτες σε αυτά τα ιδρύματα που ο καθένας με τον τρόπο του συνέβαλε στην πραγματοποίηση αυτής της εργασίας. Αποφεύγω να τους αναφέρω ονομαστικά γιατί μια τέτοια προσπάθεια θα ήταν εκ των πραγμάτων ελλειπής και επομένως και άδικη. Θα ήθελα όμως να σταθώ στους ανθρώπους των γραμματειών των δύο ιδρυμάτων, την Ρένα Καλαϊτζάκη και Μαρία Σταυρακάκη (Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης) και Γιάννη Καλαϊτζάκη, Μαρία Πρεβελιανάκη και Λιάνα Κεφαλάκη (Ινστιτούτο Πληροφορικής), όχι μόνο γιατί μου πρόσφεραν την βοήθειά τους όποτε την χρειάστηκα αλλά και γιατί το έκαναν πάντα με ιδιαίτερη προθυμία. Θέλω επίσης να ευχαριστήσω τα ίδια τα ιδρύματα για την αρωγή τους σε όλη την διάρκεια των σπουδών μου. Ιδιαίτερα θα ήθελα να ευχαριστήσω το Ινστιτούτο Πληροφορικής για την οικονομική υποστήριξη και την υλικοτεχνική υποδομή που μου προσέφερε. Πρέπει να τονίσω ότι η πραγματοποίηση αυτής της δουλειάς απαίτησε πέρα από τη θεωρητική μελέτη, πειραματισμό με εξειδικευμένες διατάξεις (robot, σύστημα ενεργούς όρασης) τις οποίες διέθεσε το Ιδρυμα Τεχνολογίας και Ερευνας.

Θερμά θα ήθελα να ευχαριστήσω και τους προσωπικούς μου φίλους και γνωστούς οι οποίοι συνέβαλαν ώστε η παραμονή μου στο Ηράκλειο Κρήτης όλα αυτά τα χρόνια να

είναι πραγματικά αξέχαστη. Εύχομαι και ελπίζω να μην χαθούμε.

Την ευγνωμοσύνη μου θα ήθελα επίσης να εκφράσω προς τους γονείς μου Ανάργυρο και Μαρία αλλά και προς τον αδερφό μου Γιώργο. Και οι τρεις τους στήριξαν πάντοτε τις επιλογές μου, τόνωναν τον ενθουσιασμό μου στις ευχάριστες στιγμές και μετρίαζαν την απογοήτευσή μου στις δυσάρεστες. Τους χρωστάω, πραγματικά, πάρα πολλά. Τέλος, θα ήθελα να ευχαριστήσω από την καρδιά μου τη σύντροφό μου Λένα, που με πραγματική αφοσίωση και με κάθε δυνατό τρόπο, στήριξε όλες μου τις προσπάθειες. Στη Λένα, με την οποία μοιράζομαι καθημερινά τους κόπους, τις επιτυχίες και τις αποτυχίες αυτών των προσπαθειών, θέλω να αφιερώσω αυτή την εργασία.

Στη Λένα

Preface

We are so familiar with seeing that it takes a leap of imagination to realize that there are problems to be solved

Richard L. Gregory

Imagine that you are relaxing in an arm-chair, without moving your body, your head, or your eyes. In the meantime, something moves in front of your visual field. Of course, you are able to see the change and react to the event; you do perceive the motion before even you understand what is really moving. How does your brain process this external event and enable you to focus your attention on the moving object? The first, quick answer somebody would give is ‘‘I just see it moving’’. But this is not really an answer. It just stresses the fact that motion perception is something so effortless for humans, that it is taken for granted in every day living. On a second thought, somebody would say ‘‘Well, maybe there exists a capability of detecting changes in the visual field: In the areas where motion occurs, the image changes over time, whereas everywhere else the image remains unchanged’’. Is this the right answer? It may be part of the truth, but it is definitely, not all of it. Imagine for example that you are in your car, driving on a highway. In this case, you continuously move relative to the surroundings. If there is another car within your visual field, you can tell whether it is moving or not. But since you also move, everything in your visual field moves and changes. Clearly, detecting changes between consecutive views in time does not suffice for the detection of independent motion. Finally, consider what happens when your eyes track a moving object. The image is cast on the same part of the retina just as if the eye does not move relative to the object, but still movement is perceived. After additional thought, it becomes clear that the mechanisms enabling biological systems to perceive motion are not trivial at all. What appears effortless for humans and other animals, is ultimately extremely hard to explain and reproduce.

This dissertation studies, from a computational point of view, the ability of a moving

observer to detect independently moving objects. It provides an overview of the research done along this direction and proposes a new framework for tackling the problem. Biological vision systems are always a source of inspiration. However, the aim of this work is not to imitate biological vision, but to provide an appropriate computational framework that will allow artificial robotic creatures to possess certain similar capabilities. To quote the title of a paper written by J. Hochberg ‘‘Machines should not see as people do, but must know how people see’’ [79]. It is our belief that the proposed methodology is able to overcome many of the shortcomings of methods that have previously been proposed and permits the detection of independent motion in a broad variety of situations and environments.

Οπτική Ανίχνευση Ανεξάρτητης Τρισδιάστατης Κίνησης από Κινούμενο Παρατηρητή

Αντώνης Α. Αργυρός

Διδακτορική Διατριβή

Τμήμα Επιστήμης Υπολογιστών

Πανεπιστήμιο Κρήτης

Περίληψη

Η ικανότητα αντίληψης της κίνησης είναι μία από τις πρωτογενείς και πιο βασικές οπτικές ικανότητες των περισσοτέρων από τους βιολογικούς οργανισμούς που είναι εφοδιασμένοι με οπτικούς αισθητήρες. Η κύρια αιτία γι' αυτό είναι η σημασία που έχει η αντίληψη της κίνησης στην επιβίωση. Με βάση την αντίληψη της κίνησης, ένας κινούμενος παρατηρητής μπορεί να εξάγει πληροφορία σχετικά με τον τρόπο με τον οποίο κινείται μέσα στον τρισδιάστατο χώρο. Τέτοιου είδους πληροφορία είναι εξαιρετικά χρήσιμη σε θέματα πλοήγησης. Ακόμη, η κίνηση σαν οπτικό ερέθισμα, μπορεί να καθοδηγήσει μηχανισμούς εστίασης της προσοχής. Τόσο για τους οργανισμούς που βρίσκονται σε χαμηλότερη βαθμίδα της εξέλιξης όσο και για τον άνθρωπο, κινούμενα αντικείμενα είναι πολύ πιθανόν να αντιπροσωπεύουν είτε επικίνδυνους εχθρούς είτε πιθανή τροφή. Σε κάθε περίπτωση, απαιτείται γρήγορη και κατάλληλη αντίδραση. Για το λόγο αυτό, οι περισσότεροι βιολογικοί οργανισμοί με οπτικούς αισθητήρες έχουν εξαιρετικά αποτελεσματικούς μηχανισμούς αντίληψης της κίνησης. Είναι χαρακτηριστικό ότι μόνο

τα μάτια των πιο εξελιγμένων ζώων μπορούν να μεταδώσουν οπτικά ερεθίσματα στον εγκέφαλο κατά την απουσία κίνησης στο εξωτερικό περιβάλλον [74].

Η κατασκευή μηχανικών συστημάτων με οπτικές ικανότητες αποτελεί όνειρο αιώνων και φιλόδοξο ερευνητικό στόχο δεκαετιών. Ειδικότερα, εξαιτίας της πληθώρας των πρακτικών εφαρμογών που μπορούν να έχουν μηχανές με δυνατότητες αυτόνομης οπτικής πλοήγησης στο χώρο, πάρα πολλές ερευνητικές προσπάθειες έχουν επικεντρωθεί σε θέματα κατανόησης της οπτικής αντίληψης της κίνησης. Οι περισσότερες από αυτές τις προσπάθειες, έχουν επηρρεαστεί από την ανακατασκευαστική θεωρία για την όραση, σύμφωνα με την οποία στόχος της όρασης είναι να κατασκευάσει εσωτερικά μια πλήρη αναπαράσταση του τρισδιάτατου κόσμου. Με βάση αυτή την προσέγγιση, κάθε πρόβλημα που αναφέρεται στο περιβάλλον μπορεί να λυθεί με αναφορές σε αυτή την αναπαράσταση. Παρά τα πολλά κομψά θεωρητικά αποτελέσματα, αυτός ο τρόπος σκέψης δεν οδήγησε στην κατασκευή συστημάτων που να μπορούν να λειτουργούν με ευστάθεια και αξιοπιστία σε ρεαλιστικά περιβάλλοντα. Σαν εναλλακτικές προσεγγίσεις, οι πρόσφατες θεωρίες της ενεργούς και τελεολογικής όρασης[1] στηρίζονται στο γεγονός ότι η πλήρης αναπαράσταση του περιβάλλοντος αφενός είναι αδύνατη, αφετέρου δεν είναι απαραίτητη. Σε μία δεδομένη στιγμή, ένα μηχανικό σύστημα πρέπει να εξάγει από το περιβάλλον μόνο εκείνα τα χαρακτηριστικά που μπορούν να εξυπηρετήσουν την πραγμάτωση των στόχων του.

Αυτή η διατριβή μελετά ένα ειδικό αλλά εξαιρετικά σημαντικό πρόβλημα αντίληψης της κίνησης: Το πρόβλημα της οπτικής ανίχνευσης ανεξάρτητης τρισδιάστατης κίνησης. Δεδομένου ενός παρατηρητή που κινείται στις τρεις διαστάσεις, το πρόβλημα της ανίχνευσης της ανεξάρτητης κίνησης μπορεί να οριστεί σαν το πρόβλημα της ανίχνευσης αντικειμένων που κινούνται ανεξάρτητα από τον παρατηρητή στον τρισδιάστατο χώρο. Οι περισσότερες από τις υπάρχουσες μεθόδους για την επίλυση αυτού του προβλήματος βασίζονται σε δισδιάστατα μοντέλα κίνησης και σε ιδιαίτερα περιοριστικές υποθέσεις για το περιβάλλον, τον τρόπο κίνησης του παρατηρητή, ή και τα δύο. Ακολουθώντας

[1]Με τον όρο *τελεολογική όραση* αποδίδεται ο αγγλικός όρος *purposive vision*

την ανακατασκευαστική θεωρία, βασίζονται στον υπολογισμό της οπτικής ροής. Ο υπολογισμός οπτικής ροής είναι ένα ασθενώς ορισμένο πρόβλημα, αφού είναι ισοδύναμο με το πρόβλημα της αντιστοίχισης (correspondence problem). Επιπρόσθετα, ο υπολογισμός οπτικής ροής βασίζεται στην υπόθεση ενός ομαλού πεδίου ταχυτήτων, που εξ ορισμού δεν μπορεί να γίνει στην περίπτωση ύπαρξης ανεξάρτητα κινούμενων αντικειμένων.

Σε αυτή την εργασία, η προσέγγιση στην ανίχνευση ανεξάρτητης κίνησης βασίζεται στις αρχές της τελεολογικής όρασης. Προτείνονται τέσσερις νέες μέθοδοι ανίχνευσης ανεξάρτητης κίνησης κάθε μία από τις οποίες έχει το δικό της, ανεξάρτητο ενδιαφέρον αφού καταφέρνει να εξάγει πληροφορία διαφορετικής μορφής από τις υπόλοιπες, με διαφορετικό υπολογιστικό κόστος. Ένα σύστημα που μπορεί να χρησιμοποιήσει οποιαδήποτε από αυτές αποκτά ευελιξία ως προς μια ποικιλία παραμέτρων. Έτσι, το συνολικό πρόβλημα της ανίχνευσης ανεξάρτητης κίνησης μπορεί να αντιμετωπισθεί πιό αποτελεσματικά μέσω της συνεργασίας των τεσσάρων προτεινόμενων μεθόδων, απ' ότι με κάθε μία από αυτές ξεχωριστά. Κοινό χαρακτηριστικό και για τις τέσσερις μεθόδους αποτελεί το γεγονός ότι βασίζονται σε ρεαλιστικά τρισδιάστατα μοντέλα κίνησης και ότι προσπαθούν να εξάγουν από το περιβάλλον μόνο εκείνα τα χαρακτηριστικά τα οποία είναι απαραίτητα για την αντιμετώπιση του προβλήματος της ανίχνευσης ανεξάρτητης κίνησης. Κοινό χαρακτηριστικό τους είναι επίσης ότι δεν βασίζονται στον υπολογισμό οπτικής ροής, αλλά στον υπολογισμό του πεδίου κάθετης ροής, το οποίο αν και περιέχει λιγότερη πληροφορία σε σχέση με την οπτική ροή, μπορεί αποδεδειγμένα να υπολογιστεί με ακρίβεια από ακολουθίες εικόνων.

Η πρώτη από τις προτεινόμενες μεθόδους προσεγγίζει το πρόβλημα ανίχνευσης ανεξάρτητης κίνησης σαν ένα πρόβλημα εύρωστης εκτίμησης παραμέτρων με βάση οπτικά δεδομένα προερχόμενα από στερεοσκοπικό παρατηρητή που εκτελεί στερεά κίνηση. Οι μετρήσεις που παίρνονται από την στερεοσκοπική διάταξη συνδυάζονται με τις μετρήσεις από την κίνηση με βάση ένα γραμμικό μοντέλο. Οι παράμετροι αυτού του μοντέλου σχετίζονται με τις παραμέτρους κίνησης του παρατηρητή και τις παραμέτρους της

στερεοσκοπικής του διάταξης. Η χρήση μεθόδων εύρωστης εκτίμησης οδηγούν στην τμηματοποίηση της σκηνής με βάση τις τρισδιάστατες παραμέτρους κίνησης των σημείων της.

Με βάση τη δεύτερη μέθοδο, η επεξεργασία της στερεοσκοπικής πληροφορίας οδηγεί στην διαστρωμάτωση μιας σκηνής με βάση το βάθος της: όλα τα σημεία της εικόνας που αντιστοιχούν σε ένα στρώμα έχουν κοινή, κατά προσέγγιση, απόσταση από τον παρατηρητή. Οι πραγματικές τιμές του βάθους δεν υπολογίζονται αφού αυτή η πληροφορία δεν απαιτείται για την επίλυση του προβλήματος. Ωστόσο αποδεικνύεται πως μπορεί να εξαχθεί ποιοτική πληροφορία για το βάθος μιας σκηνής σαν έμμεσο αποτέλεσμα της μεθόδου. Τέτοια πληροφορία είναι μία σχέση διάταξης μεταξύ των στρωμάτων κοινού βάθους. Σε κάθε στρώμα κοινού βάθους εκτελείται ανίχνευση της ανεξάρτητης κίνησης με χρήση μεθόδων εύρωστης εκτίμησης των παραμέτρων της κίνησης. Ο κινούμενος παρατηρητής μπορεί να επιλέξει να χρησιμοποιήσει την τμηματοποίηση ενός μόνο στρώματος βάθους ως προς τις τρισδιάστατες παραμέτρους κίνησης, εστιάζοντας πρακτικά την προσοχή του σε ένα συγκεκριμένο τμήμα μιας σκηνής και πληρώνοντας μικρότερο υπολογιστικό κόστος απ' ότι εφαρμόζοντας την πρώτη μέθοδο. Ωστόσο, η μέθοδος συμπληρώνεται και από την δυνατότητα συνδυασμού της πληροφορίας κίνησης από όλα τα στρώματα, με στόχο την εξαγωγή μιας συνολικής εικόνας για τις τρισδιάστατες παραμέτρους κίνησης σε μία σκηνή.

Οι δύο πρώτες προτεινόμενες μέθοδοι παρέχουν μια λεπτομερή καταγραφή της ανεξάρτητης κίνησης σε μία σκηνή. Ωστόσο, σε πολλές περιπτώσεις, είναι επιθυμητή η ταχεία εξαγωγή κάποιων συμπερασμάτων σχετικά με την ανεξάρτητη κίνηση σε μία σκηνή, ακόμα κι αν αυτή είναι πολύ ποιοτική. Αυτό επιτυγχάνεται με την τρίτη μέθοδο, η οποία βασίζεται στην έμμεση σύγκριση συναρτήσεων του βάθους, που εξάγονται από τη στερεοσκοπική διάταξη και την κίνηση. Ασυνέπειες στις συγκρινόμενες ποσότητες αποτελούν ένδειξη περισσοτέρων από μία τρισδιάστατων κινήσεων στην περιοχή σύγκρισης. Η μέθοδος που χρησιμοποιείται είναι περισσότερο ποιοτική από τις δύο

προηγούμενες μιας και αποφεύγει παντελώς την εκτίμηση των παραμέτρων κίνησης του παρατηρητή. Επίσης, σε αντίθεση με τις δύο προηγούμενες, γίνεται χρήση κάποιων υποθέσεων σε σχέση με την κίνηση του παρατηρητή, οι οποίες όμως έχουν καθαρά ποιοτικό χαρακτήρα. Το τελικό αποτέλεσμα της μεθόδου είναι ένας χάρτης των ασυνεχειών της τρισδιάστατης κίνησης σε μία σκηνή.

Τέλος, η τέταρτη μέθοδος είναι ακόμα πιο ποιοτική και γρήγορη, μιας και δεν αποκρίνεται σε αυτή καθαυτή την ύπαρξη ανεξάρτητης κίνησης αλλά σε μεταβολές της τρισδιάστατης κίνησης του παρατηρητή ή των ανεξάρτητα κινούμενων αντικειμένων. Με τον ίδιο τρόπο που η ανίχνευση μεταβολών, όταν εφαρμόζεται στις εικόνες, δίνει πληροφορία για τις αλλαγές στη θέση των αντικειμένων, η ανίχνευση μεταβολών εφαρμοζόμενη σε μετρήσιμες αναπαραστάσεις της κίνησης παρέχει πληροφορία για τις μεταβολές της ταχύτητας.

Η σύγκριση των προτεινόμενων μεθόδων για ανίχνευση ανεξάρτητης κίνησης και αυτών που ήδη υπάρχουν στη βιβλιογραφία δείχνει ότι οι πρώτες κάνουν πολύ λιγότερο περιοριστικές υποθέσεις για το περιβάλλον, την κίνηση ή το σώμα του παρατηρητή.

Πέρα από το ανεξάρτητο ενδιαφέρον κάθε μίας από τις προτεινόμενες μεθόδους ανίχνευσης ανεξάρτητης κίνησης, ιδιαίτερο ενδιαφέρον παρουσιάζει η ολοκλήρωσή τους σε ένα ενιαίο πλαίσιο. Στόχος της ολοκλήρωσης είναι ο συνδυασμός των θετικών χαρακτηριστικών της κάθε μεθόδου, ως προς μία ποικιλία διαφορετικών κριτηρίων. Στο ολοκληρωμένο περιβάλλον, η ενεργοποίηση των πιό πληροφοριακών (και πιό απαιτητικών υπολογιστικά) μεθόδων εξαρτάται από την πληροφορία που εξάγεται από τις λιγότερο πληροφοριακές (και λιγότερο απαιτητικές υπολογιστικά) μεθόδους. Η ολοκλήρωση αυτή μπορεί να χαρακτηριστεί σαν οριζόντια, υπό την έννοια ότι επιτυγχάνεται μέσω της συνεργασίας αυτόνομων διεργασιών που ανταλλάσσουν τις διαφορετικές απόψεις που διαθέτουν για το πρόβλημα, αντί μέσω μιας ακολουθίας διαδικασιών κάθε μία από τις οποίες επεξεργάζεται το αποτέλεσμα της άλλης (κάθετη ολοκλήρωση). Ένα τέτοιο ολοκληρωμένο

σύστημα έχει πολλαπλά πλεονεκτήματα που σχετίζονται με το πληροφοριακό περιεχόμενο των αποτελεσμάτων του, την ευστάθεια, την ευρωστία, την υπολογιστική απόδοση και την επεκτασιμότητά του.

Τα πειραματικά αποτελέσματα που παρέχονται για κάθε μία από τις προτεινόμενες μεθόδους δείχνουν την αποτελεσματικότητά τους σε διάφορες χαρακτηριστικές περιπτώσεις σκηνών με ανεξάρτητα κινούμενα αντικείμενα. Πιο συγκεκριμένα, οι προτεινόμενες μέθοδοι μπορούν να ανιχνεύουν ανεξάρτητη τρισδιάστατη κίνηση σε σκηνές με μεγάλη διακύμανση του βάθους, χωρίς να απαιτούν γνώση των παραμέτρων κίνησης του παρατηρητή, γνώση των παραμέτρων της στερεοσκοπικής του διάταξης ή καλιμπράρισμα των καμερών. Ο πειραματισμός έγινε σε προσομοιομένα δεδομένα, καθώς επίσης και σε πραγματικές ακολουθίες στερεοσκοπικών εικόνων που συγκεντρώθηκαν κάνοντας χρήση μιας ρομποτικής πλατφόρμας.

Visual Detection of Independent 3D Motion by a Moving Observer

Antonis A. Argyros

Doctoral Dissertation

Department of Computer Science

University of Crete

Abstract

Motion perception is one of the primitive and very basic visual capabilities of most of the biological organisms possessing visual sensors. This is due to the importance of motion perception in survival. Based on the perception of motion, a moving observer can acquire useful information about his own motion in the 3D space, which is extremely useful for navigational purposes. Moreover, motion information is a perceptual cue that can drive visual attention. For all animals, starting from the organisms in the lowest evolutionary scale and up to man, moving objects are likely to be dangerous enemies to be avoided or prey to be caught and, therefore, rapid and appropriate reaction is of utmost importance. Thus, most biological organisms with optical sensors have extremely effective mechanisms for motion perception. It is interesting to note that only the eyes of the highest animals can signal to the brain in the absence of movement [74].

The construction of artificial systems possessing visual capabilities has been a dream for centuries and a challenging research goal for decades. Motion perception in particular, has

attracted a lot of research efforts due to its fundamental importance for many visually assisted tasks. Most of the efforts have been influenced by the reconstructionist vision paradigm, according to which, the goal of vision is to provide an internal full representation of the external 3D world. According to this approach to vision, any world related problem can be solved through references in this global representation. Despite the multitude of elegant theoretic results, this methodological approach did not give rise to systems working robustly in realistic environments. As an alternative, the new theories of active and purposive vision suggest that a global representation of the environment, besides being impossible to achieve, is also not needed. At any given instance in time, the aspects of the environment that should be recovered are those that can sufficiently serve the current goals of the system.

This dissertation studies the problem of visual detection of independent 3D motion. Given an observer pursuing unrestricted 3D motion, the problem of independent motion detection can be defined as the problem of detecting objects that move independently of the observer in 3D space. Most of the existing techniques for solving this problem rely on two dimensional models of motion and on restrictive assumptions about the environment, the observers' motion, or both. Following the reconstructionist approach, they are based on the computation of optical flow, which amounts to solving the ill-posed correspondence problem. In addition, the computation of optical flow is based on the assumption of a smooth motion field, which by definition does not hold in the case of a scene with objects that move independently of the observer.

In the proposed framework, independent motion detection is approached based on the principles of purposive vision. Four new methods for independent motion detection are proposed, each of which has its own, independent interest because it deals with different aspects of the problem with a different associated computational cost. A system that can select from this set of methods gains flexibility in a number of different parameters. Thus the overall problem of independent motion detection can be more effectively solved through the cooperation of the four proposed methods rather than by each of them separately. A common characteristic of the four methods is that they rely on realistic three dimensional models of motion. Another common

characteristic is that they do not rely on the computation of optical flow, but on the computation of normal flow, which although less informative than optical flow, can be accurately computed from sequences of images.

In the first of these methods, the problem is formulated as robust estimation applied to the visual input acquired by a binocular, rigidly moving observer. Measurements based on the stereoscopic configuration of the observer and his motion are combined in a linear model. The parameters of this model are related to the 3D motion parameters of the observer and the parameters of his stereo configuration. The use of robust regression methods for this model leads to a segmentation of the scene points according to their 3D motion parameters.

In the second method, the processing of the stereo information is used to perform depth layering, i.e. segmentation of the full set of scene points into subsets of points corresponding to a common depth. The actual depths are not computed, since they are not needed for the solution of the problem. It is shown, however, that as a side effect, qualitative information about depth can be extracted, in the form of an ordering of the depth layers. Within each depth layer, motion segmentation is performed. Motion segmentation is achieved through the robust estimation of motion parameters in each depth layer. The moving observer may choose to use the 3D motion segmentation of only one depth layer. Practically, this is equivalent to focusing attention on a part of the scene in view. As a result, the observer may acquire the desired information at a lower computational cost compared to the cost of the first method. However, the method is completed with a mechanism that enables the combination of information across depth layers to provide an integrated view of 3D motion in the whole of the scene.

The first two methods provide a detailed segmentation of the 3D scene with respect to its 3D motion characteristics. However, in many cases, it is important to provide information about independent motion detection very quickly, even if this information is of qualitative nature. This is achieved by the third method which performs an indirect comparison of depth functions computed from stereo and motion. Inconsistencies in the compared quantities signal

the presence of more than one rigid motions. This method is more qualitative compared to the two previous methods, because solving for the motion parameters of the observer is completely avoided. The final outcome of the method is a map of 3D motion discontinuities.

Finally, the fourth method is even faster and more qualitative than the previous ones, since it does not provide information on independent motion itself, but on changes of the 3D motion parameters of the observer, or of the independently moving objects. Therefore, it can be used as a method for the detection of maneuvering objects. In the same way that change detection applied to the image intensity function provides information about constancy in the locations of objects, change detection applied to computable visual motion representations renders information about constancy of the 3D motion parameters.

In addition to the independent interest in each of the proposed methods, their integration in one unified framework is also important. The goal of integration is the combination of the best of the properties of each method with respect to a number of different criteria. The activation of the more informative (and computationally more intensive) methods depends on the information provided by the less informative (and computationally less intensive) ones as well as on the goals of the system. This type of integration can be characterized as horizontal, in the sense that the problem is actually solved through the cooperation of autonomous modules that exchange their different view on the problem at hand, rather than through a long chain of modules each of which processes the output of the previous module (vertical integration).

Experiments have been carried out using each of the proposed methods. The results of these experiments demonstrate their effectiveness in detecting independently moving objects in a variety of scenes. More specifically, the methods are capable of detecting independent 3D motion in scenes with large depth variations, without requiring prior information regarding the motion parameters of the observer or regarding the parameters of his stereo configuration. Experimentation was conducted using both simulated data, and real sequences of stereoscopic images that were acquired using a robotic platform.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Those who wish to succeed must ask the right preliminary questions.

Aristotle

The understanding of the vision process has been a challenging interdisciplinary area of research for centuries. The motivation behind efforts to understand the mechanisms of visual perception is twofold. On the one hand, there is the long sought goal of humans to understand their body and mind and to answer important questions about their own nature. On the other hand, there is the also long sought practical goal of knowledge exploitation in everyday life.

Philosophers and scientists have studied diverse aspects of visual perception over the centuries and proposed theories explaining the underlying mechanisms [76, 99, 43, 72, 112]. Biologists, neuroscientists, psychologists and ethologists [89, 200, 41, 84, 107, 205, 68, 51, 146, 183] have studied biological organisms and their visual capabilities. The eyes of biological organisms have lenses and a network of very small receptive units that constitute the light-sensitive retina [74]. The lens focuses an image of the world onto the retina, which then sends

messages along the optic nerve fibers to the brain. Sets of specialized neurons in the brain can then represent the imaged scene. Given that more than half of the neurons in our brain are devoted to solving vision problems [7], it is not surprising that the understanding of our visual system is not an easy problem. Equivalently, designing machines that can ‘‘see’’ is a very difficult task.

The advent of digital computers allowed the perspective of a computational theory for visual perception. Grounded in information theory, a computational theory of vision seeks to describe what information is extracted from the image(s) and how this information is extracted, processed and used [13]. In its earliest years (mid-1950's), computer vision was concerned primarily with recognition tasks in two dimensional images of static scenes such as documents, or high-altitude views of the earth surface. Research in robot vision has started in the mid 1960's. Information about the third dimension, which is lost due to the projection of light on the 2D image plane, cannot be ignored in robot vision, because a robot must operate in environments where depth variations are large compared to the absolute distance of the robot from the scene points. Research on recovering the three dimensional structure started in the early 1970's, by considering only single images of a static scene. By the mid-1970's research started to deal with time sequences of images of a possibly time-varying scene, obtained by a moving visual sensor.

Forty years of research have produced theoretical solutions to many computer vision problems; however many of these solutions are based, explicitly or implicitly, on unrealistic assumptions about the class of allowable scenes, and consequently, they often perform unsatisfactorily when applied to real-world situations. Although the results of machine vision research during the last decades are by no means negligible, there are fundamental difficulties in developing computational theories that support robust and real-time vision systems [13]. Most of the previous research efforts were based on the assumption that the purpose of vision is to provide an accurate symbolic representation of the external 3-D world. The general interpretation of the world, could then be further used to solve any world-related problem. This *reconstructionist paradigm* [112] considered vision as a goal by itself.

During the last decade, a new vision paradigm has attracted the interest of the computational vision research community. In the new paradigm, called *active vision* [9, 23, 7], vision is more readily understood in the context of the visual behaviors in which the system is engaged [6]. Active vision considers vision as part of a complex system that interacts in specific ways with the world. Consequently, it tries to explore those aspects of the world that are important to the system at a given point in time. The interest in active vision is largely motivated by the fact that all biological vision systems are highly active and purposive. Other motivations come from interdisciplinary influences from the fields of psychophysics and neurophysiology on one hand [63, 205], and progress in robotics on the other. In particular, an active vision platform could not be built without the presence of the compact CCD arrays, motors and control systems of today, as well as current microprocessor technology [165].

A successful vision system should support two general goals: *navigation* and *recognition* in complex, dynamic environments [7]. A large proportion of research in computer vision addresses, implicitly or explicitly, these two goals. For both goals, the notion of *visual attention* [129] is of central importance. Attention can be understood as the selective sensing in space, time and resolution, whether it is achieved by modifying physical camera parameters or the way the data is processed after leaving the camera [165]. Attention may be either *task-driven* or *data-driven* [23]. In the first case, attention is used to focus information gathering and processing in areas that are more likely to support the achievement of the system's goals. For example, in the case of visual search for a specific object, focus of attention mechanisms may be used to direct visual processing in areas of the visual field which are more likely to contain the desired object. In the second case, attention is used in order to cope with unexpected events. For example, a sudden motion across the visual field may represent danger, so attention should be attracted for a more detailed investigation of this event. In either case, the role of attention is crucial for a vision system, since it drastically reduces the computational effort that should be spent in order to accomplish its tasks. Consequently, advances in the understanding of visual attention are considered very important towards the development of effective, real-time vision

systems.

One of the visual cues that play an extremely important role in driving attention is motion. Most of the primitive, survival tasks of biological organisms are based on the perception of motion: It is very likely for a moving object to be either prey to be caught, or enemy to be avoided [74]. Even in the case where the observer moves in a static environment, the perception of motion provides important information regarding his own motion in the 3D space [64]. Consequently, the perception of motion is essential for many behaviors that an autonomous biological or man-made system should exhibit in real world environments.

Recognizing the importance of the visual perception of motion, this dissertation studies one of its very basic subproblems, that of independent 3D motion detection. Due to the *egomotion* (or *self-motion*) of an observer in the 3D space the whole visual field appears to be moving in a specific manner, which depends on the observer's 3D motion parameters and the structure of the scene in view [64]. In case that certain objects move independently, the 3D velocity of the observer relative to all points in his environment is not the same. The problem of independent 3D motion detection is defined as the problem of locating such objects, if they exist in a scene. The existence of biological organisms with independent motion detection capabilities motivates efforts towards constructing artificial systems with similar capabilities [65]. The ability to detect independent motion is a prerequisite for the autonomous navigation of a robot in dynamic environments. Furthermore, in other application areas that are not directly related to robotics, such as animation and virtual reality, independent motion detection mechanisms form powerful tools for automating tasks which are otherwise too complicated to perform.

In the rest of this chapter, we present the traditional, reconstructionist approach taken in computer vision during the last decades. We will reason on problems related to the application of this approach to real systems and we will present the new paradigm of purposive, active vision. The role of attention mechanisms in active vision systems will be investigated and special emphasis will be given to motion perception as a cue to attention.

1.1 Marr's vision paradigm

Marr [112] was the first to articulate a comprehensive computational theory of vision, which has been an enormous contribution in many different respects. Specifically, Marr proposed that the general goal of computer vision is to produce an accurate, 3D representation of a scene and a quantitative description of the characteristics of the objects present in it. Then, all vision related problems could be solved through symbolic processing applied to the information provided by such a representation. Marr's theory reflects the theories of the neurologists of his time [89, 189] according to which the various parts of the brain that are responsible for vision, analyze all available information, at different levels of abstraction and complexity.

Since the goal of computer vision was to provide a representation of the world, the studies of the various AI fields became independent and isolated. Vision researchers tackled the problem of building a 3D representation of the environment as a problem of building a number of operationally independent subsystems and started studying each subsystem in isolation from the other subsystems and from the other AI disciplines [64]. A common assumption and hope was that, after studying each subsystem in sufficient detail, all of them could be optimally integrated in one practical system. Unfortunately, the correct blending of results has not been successful to date. One basic problem of the reconstructionist theory of vision is that the extraction of a complete representation of the external world has immense computational requirements. For example, the human brain devotes half of its neurons to visual processing [7]. Moreover, even if a complete representation was available, the extraction of information out of it would require enormous computational resources.

1.2 Purposive and qualitative active vision

Recognizing the difficulties of the traditional approach to handle the problem of visual perception, many researchers started in the mid 80's to look for alternative approaches. At that time,

researchers in the fields of psychophysics and neurophysiology [16] and also of computational studies of behavior [37], begun to realize that the sensory capabilities of an organism are adapted to its environment, its goals and the physiology of its body. Living organisms use vision in order to navigate safely in their environment, to recognize their prey and enemies, and to support goals of vital importance. There are various visual systems with completely different visual capabilities. For example, certain frog species are able to see something only if it moves [74]. There are numerous other examples of biological visual systems, whose properties deviate significantly from those of the human visual system. A common characteristic of all these systems is the efficiency with which they support the functions of the organisms possessing them. In general, the visual capabilities of an organism are more effectively studied in the context of the behaviors it is expected to exhibit. Vision and intelligence cannot be disembodied: they have no meaning if they have no goals to achieve and a body to interact with the environment [38].

Based on the above, a new theory of vision was formulated, which was named *purposive* and *active vision* [9, 23, 7]. The terms *animate vision* [26] and *exploratory vision* [24] have also been used. One of the fundamental aspects of the purposive theory of vision is that the decomposition of a vision system into subsystems is not based on the type of their input (e.g. structure from motion, structure from shading, structure from texture, planning, learning etc), rather on the basis of the behaviors that the system should exhibit. Each behavior is implemented by a set of processes which cooperate for achieving the goals of the system. Each process is dedicated to understand certain aspects of the world that are immediately related to the goal to be achieved and, therefore, it uses a partial representation of the world. The set of processes that are active at a certain moment in time is determined by the goals of the system. Essentially, according to this theory, the goal of vision is to transform the visual stimuli in sequences of actions (behaviors). This is contrasted with Marr's theory, where the goal of vision is to transform the visual stimuli into a global representation of the environment, which may serve the solution of any vision related problem. A key contribution of the theory of active vision is that it provides researchers with a new interesting question, which leads to a new way of thinking about vision: *Is a general*

representation of the environment needed after all?

The purposiveness of visual processes enables the formulation and the solution of simpler problems. Consider for example the problem of visual detection and avoidance of obstacles [71, 87, 142, 125, 28]. According to Marr's theory, a complete and accurate representation of the world is built, and the problem of obstacle avoidance can be solved through symbolic processing based on that representation. On the contrary, purposive vision approaches the problem by defining and handling a set of immediate questions that should be answered for the obstacle avoidance behavior to be achieved. Is the observer in collision trajectory with some objects? How much time (in terms of the observer's reaction time) remains until collision? Such questions have a relative small number of answers and can be treated in a qualitative manner [7]. If it is possible to directly answer such questions, then the combination of the answers may lead to a robust solution of the overall problem and the complete representation of the environment is not needed any more (e.g. [125]). The goal of algorithms that answer the above questions is to solve many, very specific problems under general (loose) assumptions rather than trying to solve a general problem, which can only be done under very restrictive assumptions.

A fundamental issue in active vision is that the observer is actively involved in the image acquisition process by controlling some of the related parameters that can be classified in the following four categories [132]:

- **Optical parameters**, which include focal length, lens parameters etc. Such parameters determine the way in which a scene is projected on the photo-sensitive surface.
- **Sensor parameters**, that determine the way in which the photo-sensitive surface is sampled. Such parameters include the shape of the photo-sensitive surface, the distribution and sensitivity of receptors, etc.
- **Mechanical parameters**, which control the position and velocity of the camera in 3D space. Through their control, the observer is able to see different aspects of the environment, to fixate on certain objects, to track them, etc.

- **Algorithmic parameters.** Their control determines the detail in which computations proceed, as well as the coordination of the individual subprocesses running in parallel. The large volume of computations to be performed suggests their parallel execution and the support of various models of parallel processing [113]. The coordination among them is of utmost importance since all processes compete for the limited computational resources of a system which should have a reaction time capable of keeping up with rapid changes in the environment.

Through the control of the above parameters, the observer becomes able to [132]:

- **Exploit natural search.** Instead of using algorithmic search in an image, the selection of appropriate parameters may enable the acquisition of images that simplify further processing. Consider for example the case of partial occlusion. A certain object may not be completely visible in a specific frame. By moving the visual system, the observer may acquire more frames containing views of the object from different visual angles, which provide supporting evidence for the object's identity.

- **Exploit controlled movement.** Since the mechanical parts are under the system's control, it can perform known movements that constrain the relation of images taken at successive time instances.

- **Exploit object-centered reference coordinate systems.** The ability to fixate on certain points of the world, enables the system to use coordinate systems centered on the fixated points, rather than on the camera nodal point (which is the case in egocentric reference systems). Object-centered reference coordinate systems enable the definition of problems and the achievement of behaviors relative to the fixation point (e.g. grab the object which is at the fixation point, or look behind the current fixation point). The ability of an active vision system to focus its attention to objects of interest, facilitates the use of *deictic* representations [199], which isolate the objects that are directly related to the system goals.

Such representations are much less complex than those targeting a complete description of the environment, thus facilitating the applicability of learning algorithms [200].

- **Exploit simple mathematical models.** The ability of a system to fixate on certain points of the world can bring the points of interest near the optical axis, enabling the use of simple projection models (e.g. orthographic projection), as opposed to the use of the more complex perspective projection model.

- **Exploit sensors with nonuniform resolution.** Recently, technological advances have made it possible to build foveated cameras [170, 196, 195]. It has been demonstrated that, in certain cases, the use of nonuniform resolution cameras simplifies some of the required computations [196]. The characteristics of such sensors [56] require a visual system to be active, so that objects of interest can be brought into the center of the visual field and images at maximum available detail, with the proper control of the optical and mechanical parameters.

A general model of vision, which conforms with the general principles of active and purposive vision, is the one proposed by Fermüller and Aloimonos [67] and shown schematically in Fig. 1.1. The basic entities in this model are the visual competences, the purposive representations, the actions, learning and memory.

The system has always a set of goals to achieve. The visual competences process the visual input extracting the information that is directly related to the system goals. This results in purposive representations of the environment, which actually encode the intelligence of the system. These representations trigger specific actions of the system, which can be either internal or external. Internal actions change the state of the system, while external actions enable the interaction of the system with its environment. Actions are stored in an associative manner [98], that is they are accessible based on the contents of the representations. A series of actions constitutes a specific behavior of the system. Representations and actions, as well as other important information are stored in memory. Learning procedures enable the system to learn

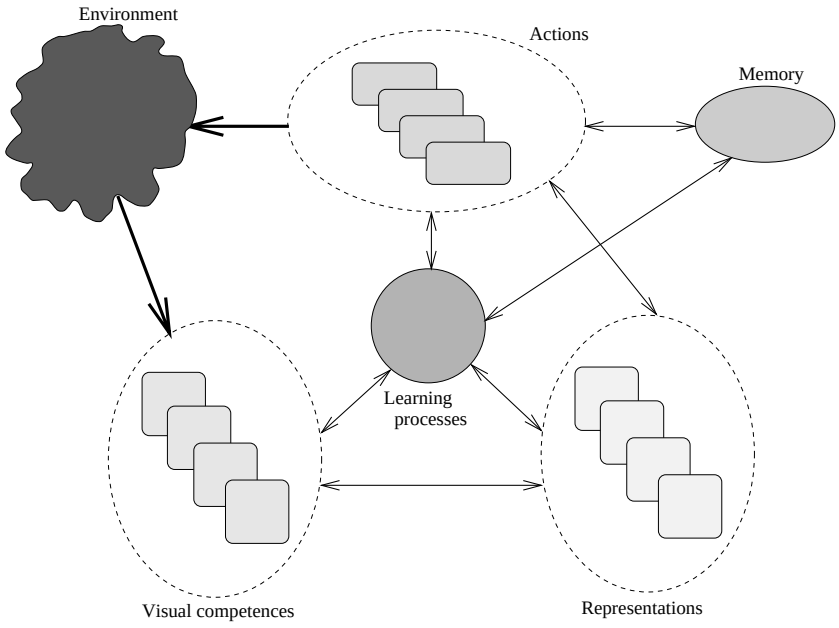


Figure 1.1: An abstract model of a purposive active vision system.

proper values for its parameters, new actions, new visual capabilities, new representations and proper representation/action combinations that facilitate the achievement of specific goals. In this way, a system may adapt to new situations and environments, without the need for detailed pre-programming of all possible situations that the system may encounter. The role of learning is of great importance in the case of systems with selective perception [114].

1.3 The role of attention in active vision

A visual signal contains information about shape, motion, color, depth, texture etc. According to the theory of active and purposive vision, this information is much more than the system needs or can simultaneously cope with. Thus, a fundamental problem for a vision system is to determine what information from the image should be used and what representation of it needs to be built so that the interaction of the system with its world can be most effective [8]. In other words, the system needs to recover partial information about the scene. The partial information to be recovered depends on the tasks that the system must carry out, i.e. on its

Reconstructive approach	Active and purposive approach
Goal: Built representation	Goal: Define visual capabilities
Use of all subsystems	Use of processes needed
Passive image acquisition	Active image acquisition
Complete visual processing	Selective visual processing
Processing at maximum detail	Processing at sufficient detail
Unlimited resources	Resource limitations

Table 1.1: Comparison of reconstructionist and purposive vision paradigms.

purpose. This is precisely the role of attention mechanisms in such a system. In general, focus of attention in vision can be understood as a selection mechanism for economizing sensory information gathering, processing and storing, with respect to a given task [24]. Thus, focus of visual attention, active vision and purposive vision are tightly coupled concepts. A system can only be purposive if it is active [24] and, in order to be purposive it must be able to focus its attention [25]. Finally, for a system to be able to focus its attention, it must be active [52].

In the presence of focus of attention mechanisms, a vision system performs a spatially selective processing of a scene, at the required resolution, taking into account only the information needed for the effective solution of the problem at hand. On the contrary, in the absence of any focus of attention mechanisms, a vision system must process the whole visual field, at maximal resolution, taking into account all possible visual cues and all internal world models. Table 1.1 (taken from [27]) summarizes the differences between the reconstructionist and the purposive vision paradigms. In the latter, attention is used to filter the information acquired and processed by a vision system.

1.3.1 Types of visual attention

According to the taxonomy proposed in [165], visual attention may be *task driven*, *context driven* or *feature driven*. Task driven control of visual attention may be independent of visual processing. For example, during navigation, visual attention may shift between a distant point and a point just in front of the vision system. If nothing unexpected happens, this strategy does not depend on visual processing.

Context driven strategies are those based upon the availability of partial evidence supporting a given task. For example, in a face recognition system, once the mouth of a face is located, the nose may be searched in specific regions of the image. The TEA-1 system, presented by Rimey and Brown [141], supports certain context related attentional strategies. TEA-1 operates on the 2D world of dinner-table images. Another effort towards context-based attention has been reported by Wixson [201], who tried to incorporate knowledge about world structure to efficiently search for objects.

Feature-related control of visual attention, on the other hand, is based on the result of low-level visual processing. Feature based focus of attention mechanisms may be classified into two classes, according to the purpose they serve [24]: Focus on an expected stimulus and focus on an unexpected object/event. The former steers the system under normal conditions in the sense of ''watch what you are doing'' type of guidance. The latter is invoked as a protection mechanism of the ''watch out for what is happening'' type. If an unexpected event is encountered, the system is interrupted and must decide whether to pay attention to the unexpected event or continue the previous activity.

1.3.2 Inputs to attention

Independently of the purpose that attention mechanisms serve, one important issue is the identification of visual primitives and features that drive attention. Potential candidates are

separable features, which can be attended to selectively, and processed separately and in parallel [91]. Moreover, such visual features must be informative at low resolution, since, prior to any shift of attention, the target object or event is typically in the periphery of the visual field.

Visual cues that drive attention can be distinguished into *static* and *dynamic* features [176]. For example, static features can be color, texture, corners, line endings, symmetry etc. Dynamic features may represent motion, illumination change, and other changes in the field of view [157]. A lot of research effort has been devoted to determining features that can drive attention. These include color, brightness, texture, motion, depth, line orientation, curvature, line endings, symmetry etc. [176, 91, 46, 186, 153, 145, 140]. Motion and color deserve special mention, since they are descriptive at low resolution.

Besides determining what are the visual primitives that may trigger attention, an important question is how these features can be combined to permit a decision on where attention should be focused. This is the problem of target selection. Ahuja and Abbott [4] see the problem of target selection as one of minimizing an objective function in a system for surface estimation. Treisman [176] proposes a framework for visual search in which feature maps are related to a master map of locations, through which focused attention serially conjoins the properties of different objects in the scene. This model, inspired by her findings of relevant psychological studies, is consistent with the idea of a pooled signal conveying information about the presence, but not the absence, of a distinctive feature. Swain and Ballard [164] proposed a model of object indexing based on color cues. In the work of Rimey and Brown [141], Bayes nets are used to incorporate evidence provided by visual cues into the system. Clark and Ferrier [49] propose a saliency based attention control. As visual data are acquired, a saliency value is computed for each possible position and the attention is shifted towards the configuration of maximum saliency. Visual features are combined to produce a saliency map by forming a weighted combination of feature values. Swain and Stricker [165] discuss an interesting analogy between an attentional system and a computer operating system. The operating system serves to allocate scarce resources. Likewise, the attention mechanism in vision allocates the scarce

resources of the fovea to multiple visual tasks. These tasks may have conflicting requirements and, therefore, may direct the fovea to different points in visual space. Shifts in visual attention can be thought of as arising from a scheduling system which gives control of the position of the fovea to different visual tasks. Finally, Hinton [78] proposes a connectionist approach based on the notion of population coding, as the means to combining simpler features to obtain more abstract objects or constructs.

1.3.3 Attention mechanisms

The photosensitive surfaces of the eyes of many biological systems are characterized by a nonuniform spatial distribution of photoreceptors which gives rise to nonuniform spatial resolution. More precisely, they present maximum resolution near the fovea, at the center of the optical axis, while the resolution decreases towards the periphery [165]. One feature of this design is the simultaneous support of a large field of view and high acuity in the fovea. This is an economical solution both in terms of sensory elements and in terms of the number of neurons required for processing of the visual input. Studies have shown that, if the resolution of the human eye were everywhere equal to its resolution near the optical axis, humans would have a brain weighting approximately ten thousands tons [147, 132]. With the small fovea at a premium in a large visual field, it is not surprising that the human visual system has special fast mechanisms (saccades) for moving the fovea to different spatial targets. These fovea-directing mechanisms actually implement elaborate focus of attention mechanisms. The control mechanisms that are used to change or maintain attention comprise the larger problem of *gaze control* [52]. The problem of gaze control may be broken down functionally into the subproblems of *gaze holding* on a visual target and *gaze shifting* between targets. Gaze shifts, called *saccades*, transfer fixation rapidly from one visual target to another. Holding gaze (or *tracking* [85, 195, 130, 12, 40, 97, 53, 5, 134, 119, 139]) involves maintaining fixation on a moving visual target. Figure 1.2 (taken from Coombs [52]) presents a top view of binocular gaze geometry. The selection of the fixation point requires the control of the gaze pan and tilt

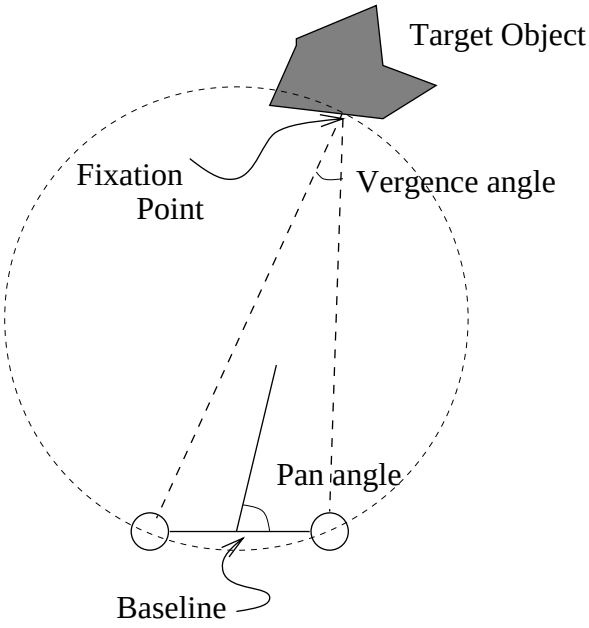


Figure 1.2: Top view of binocular gaze geometry. The gaze vector consists of the gaze pan and tilt angles, and the vergence angle (from [52]).

angles (baseline control) and the vergence angle. Biological organisms continuously change their optical and mechanical parameters by exploiting their gaze control mechanisms.

In recent years, technological advances in the fields of optical sensors (CCD cameras), microprocessors, microcontrollers and motors has enabled the construction of artificial, binocular head-eye systems [39, 131], that possess most of the degrees of freedom of biological systems. Such systems, along with the design of special hardware for real time image processing, can be employed to experimentally test and verify active vision methodologies. Figure 1.3 shows the four independent degrees of freedom of a typical head-eye vision system: Vergence control for each camera, pan and tilt control. The control of active vision heads is a topic of current research interest. Coombs [52] and Coombs et al. [40] present mechanisms for real time gaze holding. The basic assumption in this work is that the target object is initially fixated. That is, attention may not be attracted by a moving object, but may be maintained on a moving object, if this is initially fixated by the head-eye system. In [39], Brown presents several relevant efforts at the University of Rochester and, in [131], Pahlavan and Eklundh describe issues related to the

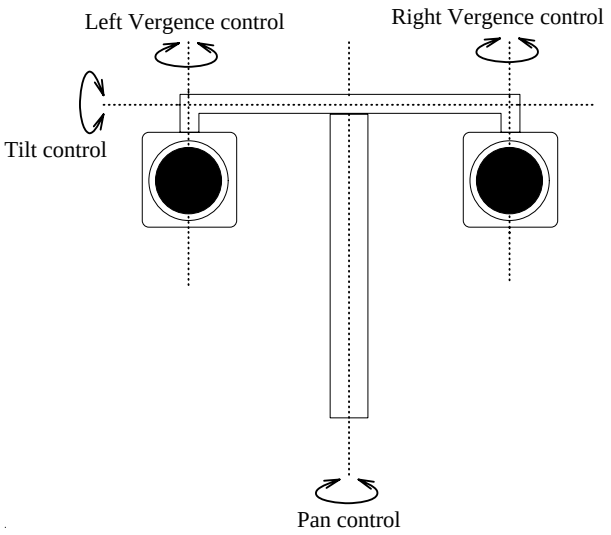


Figure 1.3: The geometry of a typical active vision head.

design and control of the KTH head eye system. Clark and Ferrier [49] deal with issues regarding the control of the Harvard head system, and Weiman presents in [196] simple-target centering algorithms in log-polar image representations.

1.4 Independent motion detection

Detection of motion is essential to survival. From animals at the lower end of the evolutionary scale to man, moving objects are likely to be either dangerous enemies or potential food and, therefore, appropriate rapid action is absolutely essential. Indeed, it is only the eyes of the animals higher in the evolutionary scale which can signal to the brain in the absence of movement. The peripheral vision system is sensitive only to movement [74]. This can be easily verified by waving an object at the periphery of the visual field. In this case, movement is seen, although it is impossible to identify the object. When the movement stops, the object becomes invisible. The very extreme edge of the retina is even more primitive. When stimulated by movement we perceive nothing, but a gaze shifting mechanism is initiated, which rotates the eye to bring the moving object into central vision [74]. In a world in which changes of state

are more important than the states themselves, the perception of motion provides a rich input to attention, informing the system about dynamic changes in the environment.

In terms of our previous discussion about attention, motion is a dynamic cue driving feature based attention. Furthermore, in terms of Bajcsy's decomposition [24], the motion cue may drive attention to both the unexpected stimulus and the expected event. The former is the case in which attention is shifted to a moving object, while the second is the case in which attention is maintained on a moving object (e.g. through tracking).

1.5 Main theme of the dissertation

The problem of visual detection of moving objects comprises the main theme of this dissertation. For the purpose of this dissertation, we will assume an observer moving relative to its static 3D environment. In this environment, there may be some objects that may be moving independently of the observer. The problem of independent motion detection amounts to segmenting the scene viewed by the observer into its static and moving components. This problem takes a special form if the observer does not move relative to the environment. In this case, all points of the static environment are projected at the same locations of the image plane, while, points belonging to moving objects are projected at different 2D coordinates as a function of time. Thus, in the case of a still observer, the problem of detecting moving objects can be treated as a problem of *change detection* [86, 157]. The situation is much more complicated when the observer is moving relative to the environment. In such a case, the points of both the environment and the moving objects project in different 2D locations on the image plane. If independent knowledge on the observer's motion parameters does not exist, change detection cannot anymore handle the problem of detection of moving objects. This dissertation studies this second case. This case is also the most interesting, because biological and man made visual systems are mobile. Even if the body of an observer is still, the eyes are continuously moving [118].

This dissertation provides an overview of the research done on the problem of independent motion detection and contributes to the field with the proposal of a new framework for tackling this problem. This framework is formed by the proper integration of four novel methods for independent motion detection which follow the purposive vision paradigm. The advantage that the proposed methods offer over existing ones is that they make fewer explicit or implicit assumptions about the external world and the observer. Biological vision systems are always a source of inspiration. However, the aim of this work is not to imitate biological vision, but to provide an appropriate computational framework that will endow artificial robotic creatures with similar capabilities.

1.6 Organization of the dissertation

The rest of this dissertation is organized as follows. Chapter 2 reviews the reconstructionist approach to motion perception with emphasis on the specific problem of independent motion detection. In the same chapter, the new purposive approach is also presented and related attempts at independent 3D motion detection are described. The chapter concludes with a brief description of the proposed methods and the contribution of the dissertation.

Based on past and current work on independent 3D motion detection, chapter 3 is devoted to the selection of an adequate motion representation for this problem. The geometry of a binocular visual system is described. Additionally, we study how the measurements that can be taken from a temporal sequence of 2D images relate to the 3D motion and structure of the environment. The analysis reveals that normal flow encapsulates much of the information regarding 3D motion and structure and, at the same time, can be accurately computed from a sequence of images. This is contrasted with the computation of optical flow, which is more informative, but turns out to be a poor selection for independent motion detection.

Having defined the basic motion representation, the following four chapters (chapters 4, 5,

6 and 7) propose a repertoire of four methods for independent motion detection, each of which produces information of added value to the results of the others. The methods are presented in decreasing order of complexity and computational requirements. As we move from expensive towards cheaper methods, the obtained results are of increasing qualitative nature. Each of these chapters also contains information regarding implementation issues and performance characteristics of the corresponding method.

In chapter 4, the problem of independent 3D motion detection is formulated as robust estimation applied to the visual input acquired by a binocular, rigidly moving observer. Measurements due to the stereoscopic configuration of the observer and due to his motion are combined in a linear model. The parameters of this model are expressions that relate the 3D motion parameters of the observer and the parameters of his stereo configuration. The use of robust regression as a means to estimate the parameters of this model leads to a segmentation of the scene based on the 3D motion parameters of its points.

For the method described in chapter 5, measurements obtained by the stereo configuration at a certain moment in time provide qualitative information about the structure of the scene in view. Based on this information, the scene is decomposed into layers of approximately constant depth. For the points of a certain layer, motion segmentation is performed, based on robust regression techniques. After having segmented each layer, the results from the various layers are merged into the independent motion segmentation of the entire 3D scene. However, the observer may use just the results of motion segmentation in a single depth layer, thus focusing his attention at a specific depth. By considering the subset of the scene points that belong to a specific depth layer, important performance gains may be obtained in comparison to the first of the proposed methods.

Chapter 6 presents a computational approach to independent motion detection that is even more qualitative and has even lower computational requirements. More specifically, this method makes an indirect comparison between qualitative structure information that is

computed from the stereo configuration and qualitative structure information computed from motion information. This information is computed in image patches. If all the points of a patch have the same 3D motion parameters, then the compared representations are the same within a constant scale and shift factor. Inconsistencies in the compared representations signal independent motion. It is noted that this comparison is able to determine whether in an image patch there is one or more rigid motions. Therefore, it can effectively be used to delineate 3D motion discontinuities.

In chapter 7, a different aspect of independent motion is recovered. A comparison of motion representations acquired by a monocular moving observer is performed to detect changes in 3D motion, rather than to detect independent motion itself. Besides being able to detect maneuvers of rigidly moving objects, the method is also capable of detecting non-rigidly moving objects and of deciding whether the observer has constant 3D motion parameters.

In chapter 8, experimental results from the application of the proposed independent motion detection methods are presented. In order to facilitate experimentation, a simulation environment has been built. This environment enables the definition of a virtual 3D space, where a binocular observer and a number of objects move with 3D velocities defined by the user. Chapter 8 reports experimental results for all the proposed methods for independent motion detection applied to the data provided by the simulation environment. The proposed methods have also been tested on real world stereoscopic image sequences acquired by TALOS, the binocular robotic platform of ICS FORTH [1].

In chapter 9, the proposed methods are compared and discussed. Furthermore, we present some thoughts on an integration scheme that enables us to tackle the problem of independent motion detection based on the goals and the computational resource limitations of a robot equipped with visual sensors. This integration scheme follows closely the design principles of purposive vision and it is shown that it has several beneficial characteristics regarding

[1]Institute of Computer Science, Foundation for Research and Technology - Hellas

performance, efficiency, robustness and extensibility.

Finally, chapter 10 concludes the dissertation by summarizing its contributions and by providing directions of further research towards building artificial systems possessing visual capabilities.

Chapter 2

Visual Perception of Motion

... several groups of animals that move have evolved a variety of visual mechanisms which prevent collisions, detect the direction of a distant movement and pursue prey or mates... Motion detection is the characteristic of those eyes, and it is this level of complexity without a large brain behind the eye, where there is now scope for copying the processing mechanisms of these eyes into artificial systems with available technology. Human technology is evolving along the same path, in that our cameras have excellent resolution but as yet no brains at all

G. A. Horridge (1987)

2.1 Overview

Until recently, the majority of studies on visual perception of motion were highly influenced by Marr's vision paradigm. In this chapter, a brief overview of such approaches is provided with

emphasis on studies of the visual perception of independent motion. Independent motion is also examined from the point of view of the purposive vision paradigm. The chapter concludes with a brief description of the methods proposed in this dissertation and a summary of its contributions.

2.2 The reconstructionist approach to motion perception

Studies of motion perception in the context of reconstructionist approach to vision consider the following distinct subproblems:

- **The problem of 2D motion computation.** The relative motion of an observer with respect to its 3D environment causes a point of the 3D space to project on different points of the 2D image plane, in a temporal sequence of images. The problem of 2D motion computation amounts to the problem of computing the displacements of the projections of 3D points between consecutive image frames. In many cases, the *correspondence problem* [81, 47, 104] is solved, i.e. the exact position of individual pixels, lines, corners etc is computed in successive frames and the differences in their positions constitute the exact disparity maps. In other cases (especially when the displacements are not too large), the dense *optical flow field* [83, 121, 77, 15, 193, 75, 182, 184, 14, 155, 48], is computed.
- **The structure from motion problem.** The problem of structure from motion amounts to extracting the 3D motion parameters of the observer and the structure of the environment based on the 2D motion field already computed from the previous step.

The above two problems are discussed in more detail in the following sections.

2.2.1 2D motion computation

The relative motion between the 3D scene and the observer (eye or camera) gives rise to changes of the intensity values in a sequence of images. The instantaneous 3D motion of any point

of an object results in a velocity vector which is assigned to the corresponding pixel on the image plane[1]. The problem of 2D motion computation amounts to the computation of these exact image velocities. To solve this problem, either corresponding features between successive frames are identified, or a dense vector field is computed, when the motion between successive frames is small.

The correspondence problem

Correspondence is usually addressed in two steps [104]. First, a type of image primitives is selected and a similarity criterion between these primitives is defined. Primitives that have been employed include raw pixels, lines and contours [202]. In an effort to select tokens that differ from their surroundings, corners have also been employed [121] and points of high interest [117]. Then, for each primitive in one frame, a search is performed which tries to locate this primitive in the other frame. To do this, a similarity criterion is chosen and the primitive is said to be found at the position that gives the best match (in terms of the similarity criterion chosen).

For this search (or mapping), various smoothness assumptions have been made about the spatial and temporal characteristics of motion, which basically restrict either the structure of the scene in view or the motions of the objects in view [47, 103]. Finally, in order to reduce the computational requirements of the search, heuristic rules have been employed. Such a formulation constrains a feature to move in a way similar to neighboring features [29, 62].

There are inherent difficulties related to solving the correspondence problem, which is an *ill-posed* problem[2] [31]. Unfortunately, the constraints imposed by the image acquisition process are inadequate for ensuring a unique solution to the problem. A simple example is presented in Fig. 2.1. Suppose that the two dots in the left image (1) have been moved to the configuration

[1]Note that we can only speak about or measure relative motion. Thus, a 3D motion vector can be assigned even to the ‘‘static” points of the environment due to the motion of the observer.

[2]A problem is characterized as ill-posed [13] when (a) it has no solution, (b) it has many solutions, or (c) its solution(s) do not continuously depend on the input.

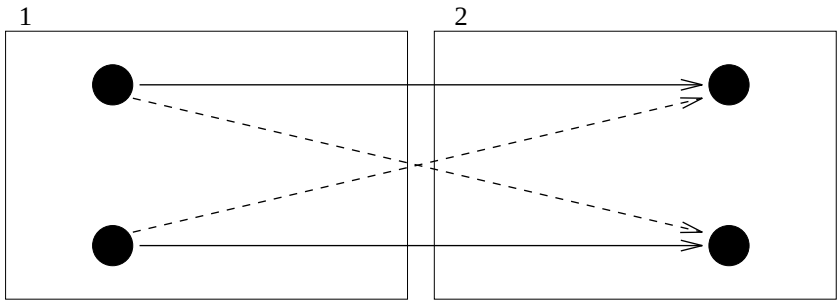


Figure 2.1: The problem of dot correspondence.

of the right image (2). In this case there are two, completely different motions that can give rise to such a result. The dots may have been moved as the horizontal, solid arrows suggest, or as the diagonal, dashed arrows suggest. In order to ensure a unique solution, additional constraints should be imposed, which is achieved through making restrictive assumptions about the environment. Such assumptions are valid in specific types of environments and, wherever they are not valid, the performance of computing the correspondence between features is very poor.

It should also be noted that correspondence techniques usually accept the best among a number of matches, without being able to determine whether a match exists. For example, in cases of a large motion, parts of the scene in view become invisible due to occlusion, or they were previously occluded and become visible. Tokens residing in these areas do not have a corresponding token in the other image and, therefore, any effort to solve the correspondence problem will lead to wrong results.

Optical flow computation

The relative motion of the observer with respect to the scene gives rise to motion of brightness patterns that are formed on the image plane. The instantaneous changes of the brightness patterns are analyzed to derive the optical flow field, a dense, two dimensional vector field reflecting the image displacements.

Similar to the case of the correspondence problem there are inherent difficulties in the computation of optical flow. First of all, the optical flow is defined in terms of the changes of intensity patterns in an image. However, changes in intensity, are not always due to physical movement [82]. For example, consider a camera that acquires images of a stationary sphere illuminated by a moving light source. The motion of the light source will result in changes in the intensity patterns recorded by the camera. On the other hand, if a sphere with no texture is rotating under fixed illumination, then no change in image intensity will be recorded. Thus, in some cases, the computed optical flow field does not correspond to a physical motion and, in other cases, no optical flow can be measured although there is motion in the imaged scene. Verri and Poggio [184] have shown that the motion and optical flow fields are identical in specific cases only.

Even in the cases that these two fields are identical, there are problems regarding the computation of optical flow. The optical flow vector at each point is computed locally, based on a small spatiotemporal neighborhood of that point. However, it can be shown that optical flow cannot be recovered correctly based on local information only. More specifically, the component of the optical flow that is parallel to the feature (i.e. parallel to the tangent of the contour at that point) cannot be recovered (for a detailed, technical explanation see Chapter 3, pp. 44). This is an *inherent* problem of optical flow and does not depend on the technique used to estimate it. In an effort to overcome the problem, techniques for computing optical flow involve two computational steps that exploit two different constraints that are needed for its computation. In a first step, assuming the conservation of some type of information within the image frames, the locally available velocity information is computed. Three different approaches can be distinguished [155]: *gradient based approaches* which assume conservation of image intensity [83, 121], *correlation based techniques* which assume conservation of the local intensity distribution [203, 45, 15, 14] and *spatiotemporal energy-based approaches* [190, 1, 70, 75, 204] which are analogous to gradient-based approaches in spatiotemporal frequency space.

In a second step, and in order to derive a second constraint, additional assumptions have to

be made. Either some kind of smoothness is assumed, or the shape of the scene is geometrically modeled to obtain constraints on the optical flow values. Employing additional smoothness assumptions to recover an unknown function when the available constraints are not sufficient is a general mathematical technique known as *regularization* which has been introduced to computer vision by Poggio [137]. A regularization technique has been used by Horn [83] which is based on the hypothesis that optical flow varies smoothly in most parts of the image. However, regularization cannot, in principle, cope with multiple motions, depth discontinuities or violation of the assumption regarding intensity conservation. Uras [182] requires that the gradient of the image brightness does not change over time. In other cases [163, 120, 193, 192] assumptions about scene geometry have been employed rather than smoothness of the optical flow field. Such a typical assumption is that the world can be locally approximated by planar surfaces.

For a more detailed presentation of different approaches for optical flow estimation the reader is referred to [155]. Furthermore, a comparative study of representative optical flow techniques can be found in [30].

2.2.2 Structure from motion

Assuming that accurate information is available either in the form of correspondences or in the form of a dense optical flow field, the 3D motion and structure of the scene can be computed from equations that relate the measurements of 2D motion to the 3D motion parameters.

The problem of structure from motion has been extensively studied at a theoretical level. A number of results concern the type of information about the motion of the observer and the structure of the environment that can be extracted through a sequence of images [180, 181]. The conditions which guarantee uniqueness of the solution [109, 177] and the behavior of structure from motion methods in the presence of noise [177, 158, 3] have also been investigated. At a practical level, a plethora of algorithms have been designed for the computation of the structure of a scene [80, 191, 192, 198, 178, 159, 172, 151, 166, 197, 105, 136].

In general, the reconstruction of the 3D motion and the structure of a scene cannot be unique because in principle, any arbitrary motion is possible. Therefore, a motion model is required. Since many of the structures of the world are rigid, rigidity [181] or at least piecewise rigidity is usually assumed as the basis for all studies.

The exact form of the equations relating the 2D motion with the 3D motion parameters and structure is determined by the type of projection assumed. From the plethora of projective models, the *orthographic* and the *perspective* models [122] are those most commonly encountered. The orthographic model yields linear equations, but is a realistic approximation in a small area around the camera optical axis, or in the whole visual field of cameras with large focal length. Because of the relative simplicity of the orthographic projection model, it was the first that was considered in attempts at solving the structure from motion problem [181, 80, 10, 80]. On the contrary, the perspective projection model is more realistic, but it results in a non linear relation between the 3D motion and structure and the 2D image motion.

2.2.3 Independent motion detection

Independent motion detection[3] has been often approached as a problem of segmentation of the two dimensional motion representation that has been computed from a temporal sequence of images. The general algorithmic framework that is applied by these methods is the following:

1. Extract some type of motion representation from a temporal sequence of images.
2. Assume that this information is described by some kind of 2D model.
3. Detect the discontinuities of this model and report them as 3D motion discontinuities.

Based on different alternatives regarding each of the above algorithmic steps, a great variety of different methods has been proposed. Step (1) is implemented by any of the available optical

[3]In the remainder of this dissertation and, unless it is explicitly stated otherwise, we will refer to the problem of independent motion detection for the case of a moving observer

flow estimation methods or by employing the spatiotemporal derivatives of the image intensity function. In step (2), an affine or a quadratic model [179] is assumed for the optical flow. According to the affine model, the optical flow (u, v) at a point (x, y) of an image can be modeled by the following equations:

$$u = a_0 + a_1 x + a_2 y$$
$$v = b_0 + b_1 x + b_2 y \tag{2.1}$$

where a_0, a_1, a_2, b_0, b_1 and b_2 are the six parameters of the model. The quadratic model (known also as pseudo-projective), models (u, v) with eight parameters as:

$$u = a + bx + cy + gx^2 + hxy$$
$$v = d + ex + fy + gxy + hy^2 \tag{2.2}$$

In [187], the parameters of an affine model are estimated locally in image patches. Patches are then combined in larger motion segments based on a k-means clustering scheme that merges two layers if the distance of their motion parameters is sufficiently small. Some methods do not compute optical flow, but they are based on the spatiotemporal derivatives of the image intensity function, the so called normal flow. For example Irani et al [93] and Torr and Murray [174] perform independent motion detection based on normal flow rather on optical flow. However, again two dimensional models are employed. Nordlund and Uhlin [127] estimate the parameters of an affine model of 2D motion. By constraining the spatial extent of the independently moving object, they assume that the estimation of the dominant 2D motion will not be affected considerably by the presence of the independently moving object. Therefore, independent motion can be achieved by determining the points where the residual between the measured flow and the flow predicted by the estimated parameters is large. Major emphasis is usually put on step (3), for which a variety of techniques have been proposed. For example Bober and Kittler [34] combine robust statistics with the Hough transform to detect

independently moving objects. In [88], robust statistics and more specifically M-estimators[4] are used to distinguish the dominant 2D motion from the secondary 2D motions. A similar idea is exploited by Ayer et al [22], where two other robust estimators, namely Least Median of Squares and Least Trimmed Squares are used to discriminate the dominant from the secondary 2D motions. Bouthemy and Francois [35] view 2D motion segmentation as a problem of statistical regularization using Markov Random Field models. Some other methods, approach independent motion detection as a problem of *change detection*. Change detection [86] is usually applied in the case of a still observer. In that case, there is no relative motion of the observer with respect to the still background, while there is a relative motion of the observer with respect to any independently moving objects. Change detection methods are used to detect exactly this discriminating property. Paragios and Tziritas [135] also apply change detection in the case of a moving observer. First, the dominant 2D motion is estimated and then compensated. After compensating for the dominant (i.e. observer's) motion, a change detection algorithm is applied to isolate regions that move independently. Odobez and Bouthemy [128], also take the approach of camera motion compensation by employing multiscale MRF models.

Many of these methods produce good results in certain classes of scenes, but have two basic drawbacks that are related to some inherent problems of steps (1) and (2) of the general algorithmic framework presented earlier. The methods that are based on the computation of optical flow tend to eliminate the effect they are trying to detect, because the smoothness assumptions that are employed in order to compute optical flow, tend to eliminate the existing motion discontinuities. Another drawback stems from the simplified 2D motion models that they employ. As it will be shown later in Chapter 3, the projection of 3D motion on the 2D image plane depends on certain characteristics of the observer (e.g. focal length of the system), on the parameters of the relative 3D motion between an object and the observer and, on the depth of the scene. Therefore, discontinuities in the computed 2D motion field are not only due to 3D motion discontinuities (i.e. independently moving objects), but also due to depth discontinuities. The

[4]For a brief introduction to robust estimators see Chapter 3.

simplified 2D models that are employed by all the previous methods do not take into account the depth of the scene. This is equivalent to assuming that depth is constant which is in principle, an unrealistic environmental assumption. For this reason, all the above methods perform well in scenes where the depth variation is small compared to the distance from the observer. In the case of scenes with large depth variations, these methods cannot provide reliable results. For such scenes the segmentation of the 2D optical flow field is not equivalent to a 3D motion segmentation.

In an effort to overcome the inherent problems of employing 2D models of motion, the more accurate 3D models have been employed. By employing 3D models the problem becomes much more complicated because extra variables are introduced regarding the depths of the scene points. Thus, certain assumptions are made in order to provide additional constraints for the problem. Common assumptions of existing methods are related to motion of the observer, on the structure of the scene in view, or both. Most of these methods depend, again, on the accurate computation of the optical flow field. Wang and Duncan [188] present an iterative method for recovering the 3-D motion and structure of independently moving objects from a sparse set of velocities obtained from a pair of calibrated, parallel cameras. Based on stereo correspondences, the method extracts the structure of the viewed scene and identifies independently moving objects by examining the consistency between the stereo disparities, the left/right image flows and the estimated 3-D motion components. Jain [94] has considered the problem of independent 3D motion detection by an observer pursuing translational motion. In addition to imposing constraints on egomotion (the observer's motion cannot have rotational components), knowledge of the direction of translation is required. The method of Clarke and Zisserman [50] is, again, applicable to the case of a purely translational egomotion. Adiv [2] performs segmentation by assuming planar surfaces undergoing rigid motion, thus introducing an environmental assumption. In order to determine the motion parameters of the moving plane and to group the motion vectors, he employs a Hough transform. A very interesting method is proposed by Lobo and Tsotsos [108] which is based on the Collinear Point Constraint (CPC).

The CPC defines a way of selecting projections of optical flow vectors so that the combined motion for these vectors depends only on the translational motion the observer. Given this, the Focus of Expansion and the parameters of rotational motion can be determined. The estimation of egomotion will not be affected much by the presence of an independently moving object that covers a small portion of the field of view. Therefore, such objects can easily be detected because their flow will not be compatible with the one suggested by the estimated egomotion parameters. Torr and Murray [173] approach the problem of independent motion detection by exploiting the epipolar constraint. More specifically, the problem of motion segmentation is transformed into one of finding a set of fundamental matrices which optimally describe the observed temporal correspondences. Thompson and Pong [169] derive various principles for detecting independent motion when certain aspects of the egomotion or of the scene structure are known. Although it is an inspiring work, the practical exploitation of the underlying principles is limited because of the assumptions they are based on and other open implementation issues. Finally, Thompson et all [168] employ the rigidity constraint in order to detect independently moving objects based on robust regression, applied to point correspondences that have been established under an orthographic projection model.

2.3 Purposive approaches to motion perception

The principles of active vision presented in the previous chapter may serve as a starting point for a new approach to visual perception of motion. Based on these principles, Fermüller [64] proposes the *synthetic approach.* According to this approach, a robotic system should initially be equipped with very simple capabilities which should then be progressively enriched by new, more complex ones. The capabilities are related to behaviors that the system should exhibit while navigating, such as egomotion estimation, independent motion detection, obstacle avoidance, motion programming in order to reach a desired destination, etc. Some initial efforts for tackling the first two problems are given in [64]. Obstacle avoidance is treated in [125, 142, 28], while

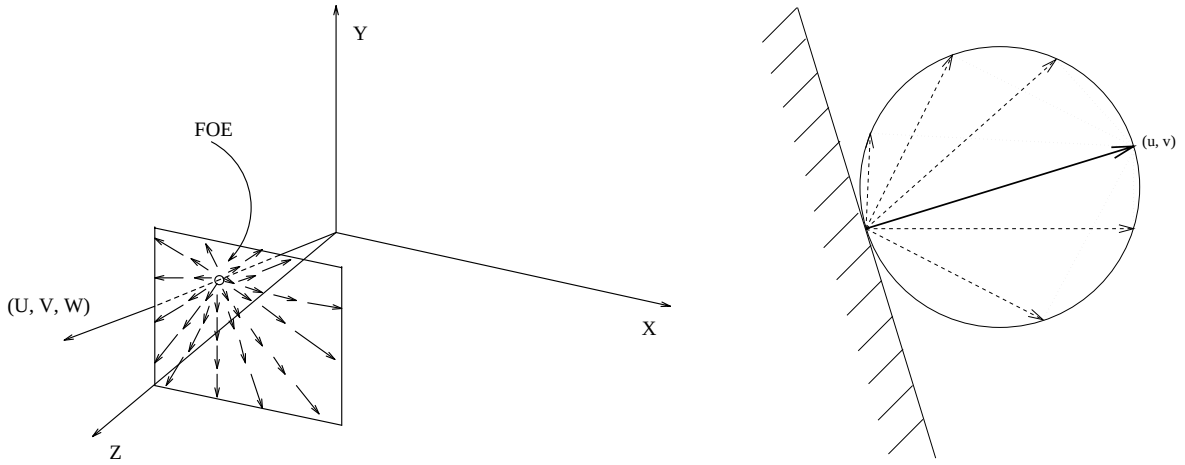
initial efforts towards motion programming are presented in [123, 54]. The synthetic approach of Fermüller resembles the hierarchical *subsumption architecture* proposed by Brooks [37].

The synthetic approach to vision problems conforms with the results of research in the field of neurobiology [205] where experimental results have shown that biological vision systems have sets of neurons that are responsible for specialized capabilities. This approach is also in agreement with the theory of evolution. Species that at a certain stage of their evolution acquired photo-sensitive sensors, probably did not obtained all their visual capabilities at once, but at different times, based on their needs. Consequently, there may be different visual processes implemented by different sets of neurons.

2.3.1 Purposive approaches to independent motion detection

Among the various behaviors that an active, purposive system should exhibit in order to navigate successfully, that of independent motion detection is of primary interest in this work. According to the purposive paradigm, qualitative capabilities should be developed, which may answer simpler questions in a robust manner. From a technical point of view, these capabilities should not rely on the computation of optical flow, which is provably ill-posed in the general case but on simpler, more goal directed representations of the environment.

Methods for independent motion detection that fall within this context have already been proposed. One method has been proposed by Sharma and Aloimonos [149, 150]. However, as in the case of [94], known translational egomotion is assumed. For such an egomotion, it is known that all optical flow vectors emanate from a specific point which is called Focus of Expansion (FOE) if vectors diverge from that point (Fig. 2.2(a)), or *Focus of Contraction (FOC)*, if vectors have the opposite direction. The coordinates of the FOE are directly related to the translational egomotion parameters. Thus, if egomotion parameters are known, then the FOE, and therefore the direction of optical flow at each point is also known. The normal flow vectors due to an optical flow vector are constrained to lie on a half-plane, as shown in Fig. 2.2(b). Since the



(a) (b)

Figure 2.2: (a) Optical flow vectors are parallel to lines intersecting at the FOE, (b) An optical flow vector can only produce normal flows that lie on the half plane that includes the circle.

direction of the optical flow vectors is known, the half plane on which a normal flow should lie is also known. Normal flow vectors that violate this constraint can only be due to independent motion. Nelson [124] presents two methods for independent motion detection, which are also based on the normal flow field. The first of these methods requires *a priori* knowledge of egomotion parameters and assumes upper bounds on the depth of the scene. Based on the above, normal flow vectors are again constrained to lie on specific areas of the plane. Normal flow vectors that do not meet this constraint are reported as due to independent motion. The second method detects abrupt changes of independent motion rather than independent motion itself.

A distinctive feature of the methods presented in this section, which differentiates them from other existing methods, is that they do not rely directly on the estimation of the motion parameters in order to provide answers to the problem of independent motion detection, but rather they rely on geometrical properties of the normal flow field. Thus, the solution of the structure from motion problem is by-passed. However, these methods are based on the restrictive assumption of either known or restricted egomotion.

2.4 Contribution of this work

In order to overcome the limitations of existing methods that rely on restrictive assumptions about the environment and/or the observer, this thesis proposes four new methods for independent motion detection. Each of these methods has specific characteristics (effectiveness, computational performance, representation of results) which favors its application under different circumstances. The integration of these methods under a unified framework provides a very good balance of effectiveness and performance.

The present work makes several contributions in the research area of independent 3D motion detection. First of all, based on the principles of purposive vision, an effort is made towards the definition of an *adequate* representation of visual motion, rather than trying to fully recover motion information. Towards this goal, we develop algorithms for independent motion detection which rely on the computation of normal flow rather than optical flow. It is considered preferable to rely on qualitative information that can be computed without environmental assumptions, instead of relying on general representations that can be robustly constructed in very restrictive environments and situations.

In the rest of this dissertation, it is shown that normal flow, the component of motion in the direction of the image gradient, although less informative than optical flow, may comprise a stable basis for developing independent motion detection capabilities. Based on the specific questions asked, different manipulations of normal flow give rise to proper representations of the motion characteristics of a scene. The set of these methods and representations may be used by a purposive system according to the aspects of the world that are of interest at a particular moment in time, depending on the system's goals. This repertoire of possible representations and capabilities is subject to continuous improvement, as new representations and more capabilities are added.

Based on the choice of normal flow to represent visual motion information, a repertoire of

four novel methods for independent motion detection is presented. Besides the fact that they all employ normal flow, the proposed methods are differentiated from other existing ones in that they achieve simultaneously the following:

- They employ 3D motion models and assume an observer that may move with unrestricted (i.e. both translational and rotational) 3D motion.
- They exploit a stereoscopic configuration without requiring exact knowledge of the stereo configuration parameters (i.e. baseline length, vergence angle).
- They do not rely on knowledge of the observer's motion parameters and do not require independent motion to be rigid.
- They assume perspective projection which is the most realistic projection model.

The proposed methods differ in terms of the information they generate about independent motion and certain aspects of their performance. The first two methods give an accurate map of independently moving objects. The third gives a coarse map of spatial discontinuities of 3D motion, but is computationally less demanding than the first two methods. The fourth method is computationally the most efficient but detects changes of 3D motion rather than independent motion itself.

The first three methods for independent motion detection make use of a binocular vision system. Through a stereoscopic configuration, a vision system is able to record information about the environment which is not subject to dynamic events such as independent motion (both left and right images of a stereo configuration are acquired at the same moment in time). The actual parameters of the stereo configuration (baseline or vergence angle) are not used by any of the proposed methods. The fourth method can be employed by a monocular observer since it does not make any use of stereo information.

Regarding egomotion, no strict simplifying assumptions are made, as is the case with other existing methods (e.g. [94, 149]). Simplifications usually encountered in the literature involve

strict constraints on the type of motion of the observer, exploitation of independent knowledge of egomotion or structure, or a combination of the above. Relying on independent knowledge of the motion parameters is a restricting assumption for many reasons. First, sensors that provide such information [61] are not sufficiently accurate. Second, they provide motion information relative to a coordinate system different from the one of the camera. Thus, any effort to exploit this information should rely on calibration of the system that will define the suitable transformations between coordinate systems. On the contrary, the first two of the proposed methods tackle the general problem of independent motion detection of an observer pursuing unrestricted 3D motion. The third method makes certain assumptions about the egomotion of the observer, which are of qualitative nature. Specifically, certain ordinal relations should hold on some of the motion parameters.

In all four methods, the motion of the independently moving objects is not constrained to be rigid. Thus, all the proposed methods can be used to signal independent non-rigid motion which is very common in dynamic, real world environments.

Finally, an integration framework is proposed which couples the benefits of the four proposed methods. Based on this framework, an artificial autonomous system can actively choose a method for independent motion detection, based on the type of information it needs in order to accomplish its goals, on past observations and on the computational resources that is able to devote to this type of visual processing. This is actually a very basic outcome of this work; the effective solution to vision problems can more readily be achieved through the successful cooperation of loosely coupled processes that tackle certain aspects of the problem rather than by single processes that try to solve general problems as a whole.

Chapter 3

Preliminaries

But life is short and information endless... Abbreviation is a necessary evil and the abbreviator's business is to make the best of a job which, although intrinsically bad, is still better than nothing.

Aldous Huxley

Before proceeding with the description of the proposed independent motion detection methods, issues related to motion representation are discussed. The rationale for the choice to employ the normal flow field in all computations is given, using the 2D velocity equations of the image points (cf. eq. (3.5)). Additionally, a discussion on robust regression is provided, since robust estimation methods are statistical tools that are extensively used by the proposed independent motion detection algorithms.

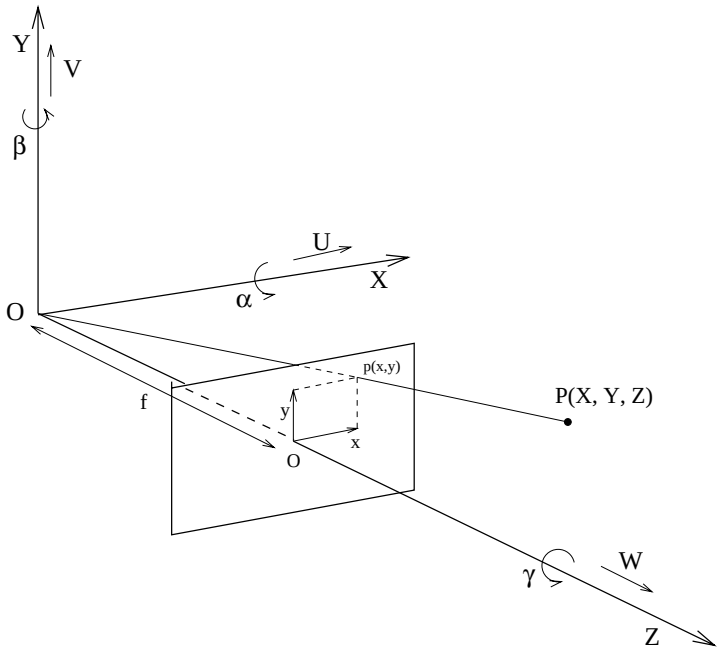


Figure 3.1: The camera coordinate system.

3.1 Visual motion representation preliminaries

3.1.1 The coordinate system

Consider a coordinate system $OXYZ$ adjusted to the optical center (nodal point) of a pinhole camera, such that the axis OZ coincides with the optical axis. Suppose that the camera is moving rigidly with respect to its 3D static environment with translational motion $\vec{t} = (U, V, W)$ and rotational motion $\vec{\omega} = (\alpha, \beta, \gamma)$, as shown in Fig. 3.1. This situation is geometrically equivalent with the case of a static camera and a scene that is moving relative to the camera with translational motion $-\vec{t}$ and rotational motion $-\vec{\omega}$. The velocity $\vec{V}$ of a point $P = (X, Y, Z)$ with respect to coordinate system $OXYZ$ can be computed as follows. Let L be the axis of rotation defined by the vector $-\vec{\omega}$ and d the distance of P from this axis (see Fig. 3.2). Point P rotates around L in a counterclockwise direction. The velocity $\vec{V}$ is the sum of the tangent velocity due to rotation $\vec{q}$ and the translational motion $-\vec{t}$. $\vec{q}$ is perpendicular to the plane defined by L and P and its magnitude is equal to the magnitude of rotational velocity times the distance from P to

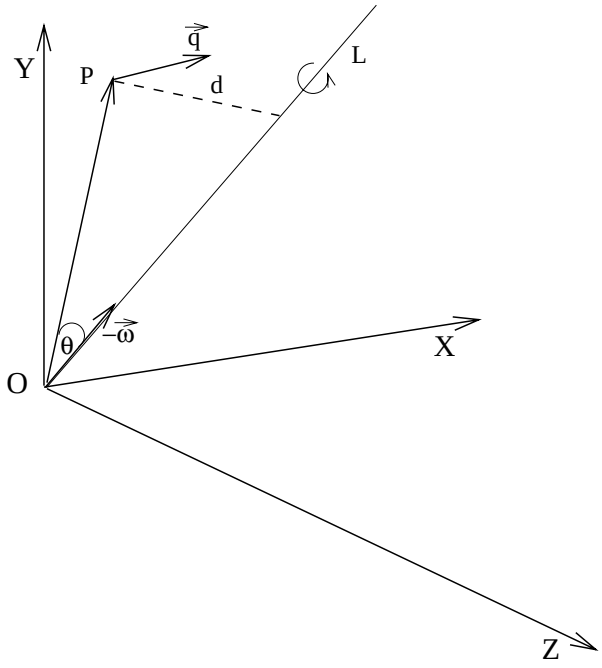


Figure 3.2: Rotation of a point P around axis of rotation L.

L, that is $||\vec{q}|| = || - \vec{\omega}||\ d = || - \vec{\omega}||\ ||P||sin\theta = || - \vec{\omega} \times P||$, where ''$\times$'' denotes cross product. Thus, $\vec{V}$ is

$$\vec{V} = -\vec{t} - \vec{\omega} \times \vec{r} \tag{3.1}$$

where $\vec{r}$ is the column vector $(X, Y, Z)^T$. The above equation can be written in component form as

$$\begin{aligned} X' &= -U - \beta Z + \gamma Y \\ Y' &= -V - \gamma X + \alpha Z \\ Z' &= -W - \alpha Y + \beta X \end{aligned} \tag{3.2}$$

Under perspective projection, the 3D point $P(X, Y, Z)$ projects at the point $p(x, y)$ on the image plane, according to the relations:

$$\begin{aligned} x &= \frac{Xf}{Z} \\ y &= \frac{Yf}{Z} \end{aligned} \tag{3.3}$$

where f represents the focal length of the imaging system. The projection (u, v) of the 3D motion vector on the 2D image at point (x, y), is:

$$\begin{aligned} u &= \frac{dx}{dt} \\ v &= \frac{dy}{dt} \end{aligned} \tag{3.4}$$

By differentiating eq. (3.4) after substitution from eq. (3.3) and eq. (3.2), it turns out that the equations relating the 2D velocity (u, v) of an image point $p(x, y)$ to the 3D velocity of the projected 3D point $P(X, Y, Z)$ are [110]:

$$\begin{aligned} u &= \frac{(-Uf + xW)}{Z} + \alpha\frac{xy}{f} - \beta\left(\frac{x^2}{f} + f\right) + \gamma y \\ v &= \frac{(-Vf + yW)}{Z} + \alpha\left(\frac{y^2}{f} + f\right) - \beta\frac{xy}{f} - \gamma x \end{aligned} \tag{3.5}$$

3.1.2 Motion field - optical flow field

Equations (3.5) describe the 2D motion vector field, which relates the 3D motion of a point to its 2D projected motion on the image plane. The motion field is a purely geometrical concept and, unfortunately, it is not necessarily identical to the optical flow field [82], which describes the motion of brightness patterns observed because of the relative motion between an imaging system and its environment. Verri and Poggio [184] have shown that the motion and the optical flow fields are identical in specific cases only. Even in the cases that these two fields are identical, the computation of the optical flow field is an ill-posed problem since special conditions (such as smoothness) should be satisfied for a unique solution to exist. This issue is further clarified by the following discussion.

In many cases, a sequence of images can be modeled as a continuous function $I(x, y, t)$ of

two spatial (x, y) and one temporal (t) variables. $I(x, y, t)$ actually expresses the image intensity at point (x, y) at time t. Assuming that irradiance is conserved between two consecutive frames, and that u and v are the x- and y-components of the optical flow, we expect that the irradiance will be the same at point $(x + \delta x, y + \delta y)$ at time $t + \delta t$, where $\delta x = u\delta t$ and $\delta y = v\delta t$. That is:

$$I(x + u\delta t, y + v\delta t, t + \delta t) = I(x, y, t) \tag{3.6}$$

By expanding the left hand side of the above equation in a Taylor series, we obtain:

$$I(x, y, t) = I(x, y, t) + \delta x\frac{\partial I}{\partial x} + \delta y\frac{\partial I}{\partial y} + \delta t\frac{\partial I}{\partial t} + e$$

where the term e corresponds to higher order terms and $\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial I}{\partial t}$ are the partial derivatives of I with respect to x, y and t respectively. Canceling out the term $I(x, y, t)$, dividing by δt and taking the limit as $\delta t \to 0$, we obtain

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} = 0$$

Using the abbreviations

$$u = \frac{dx}{dt},\ v = \frac{dy}{dt}$$

and

$$I_x = \frac{\partial I}{\partial x},\ I_y = \frac{\partial I}{\partial y},\ I_t = \frac{\partial I}{\partial t},$$

we obtain

$$I_x u + I_y v + I_t = 0 \tag{3.7}$$

Equation (3.7), which has originally be developed by Horn and Schunk [83], is known as the *optical flow constraint equation* because it gives one constraint for the components u and v of optical flow. Equation (3.7) can be written in the form of a dot product

$$(I_x, I_y) \cdot (u, v) = -I_t \tag{3.8}$$

which can be geometrically interpreted as permitting the computation of the projection of the optical flow along the intensity gradient direction (i.e. the perpendicular to the edge at that point).

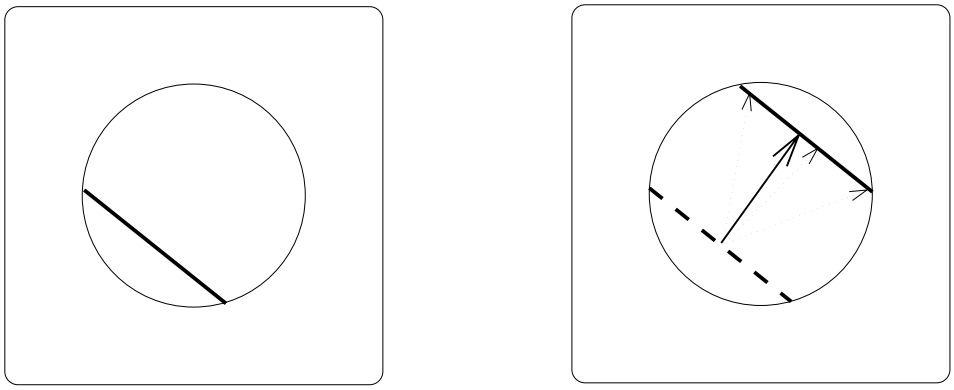


Figure 3.3: A schematic view of the aperture problem. The solid line in the left image has moved to a new position in the right image. Based on the information that is visible through the aperture, it is not possible to decide which of the dashed vectors corresponds to the motion vector of the line. However, whatever the motion vector may be, its projection to the direction perpendicular to the line is unique and is represented by the solid vector.

This projection is also known as *normal flow*. Equation (3.8) can be viewed as the mathematical expression of the *aperture problem* and Fig. 3.3 gives its schematic explanation. Equation (3.8) also shows why extra assumptions should be made in order to compute optical flow. This equation gives only one local constraint on the flow values while there are two unknowns to be recovered for each optical flow vector. Due to this lack of information, all methods that aim at recovering optical flow need to employ additional assumptions, such as smoothness of the optical flow field. However, the optical flow field is not smooth for a number of reasons. First, as eqs. (3.5) show, the smoothness of optic flow depends on the continuity of depth variable Z. In most scenes depth discontinuities do exist and, therefore, the optical flow cannot be regarded as smooth. Second, the smoothness of optical flow also depends on the constancy of 3D motion parameters. Thus, independently moving objects cause discontinuities in the optical flow field. Black and Anandan [33] provide a framework that can be applied to standard formulations of optical flow estimation in order to reduce their sensitivity in the presence of transparancy, depth discontinuities, independently moving objects, shadows and specular reflections. However, in the general case, relying on optical flow for independent motion detection is equivalent to

attenuating an effect that we are trying to detect (i.e. 3D motion discontinuities).

For the above reasons, the proposed methods for independent motion detection do not rely on the computation of the optical flow field, rather on the normal flow field. The normal flow field has also been used in the past for both egomotion estimation [64, 11, 154] and independent motion detection [124, 149].

3.1.3 Normal flow field - normal motion field

The algebraic value of normal flow can be computed from eq. (3.8), by dividing both sides with the gradient magnitude, and is equal to:

$$-\frac{I_t}{\sqrt{I_x^2+I_y^2}} \tag{3.9}$$

The component of the optical flow along the image gradient is called *normal flow* given by:

$$-\frac{I_t}{\sqrt{I_x^2+I_y^2}}\left(\frac{I_x}{\sqrt{I_x^2+I_y^2}},\frac{I_y}{\sqrt{I_x^2+I_y^2}}\right)$$

or

$$\left(-\frac{I_t I_x}{||\nabla I||^2},-\frac{I_t I_y}{||\nabla I||^2}\right)$$

where $||\nabla I||$ is the gradient magnitude. Figure 3.4(a) shows a typical synthetic translational optical flow field and Fig. 3.4(b) shows one possible normal flow field related to that optical flow. There are infinitely many normal flow fields that can result from a particular motion field, because normal flow depends on the image gradient direction and therefore, is scene dependent. The normal flow field of Fig. 3.4(b) corresponds to projections in random directions that are uniformly distributed in the range $[0..2\pi)$. Similarly, Fig. 3.4(c) shows a typical rotational optical flow field and Fig. 3.4(d) shows one possible normal flow field related to that optical flow. For displaying purposes, all fields are subsampled.

The normal flow field is not necessarily identical to the *normal motion field* (the projection of the motion field along the gradient), in the same way that the optical flow is not necessarily

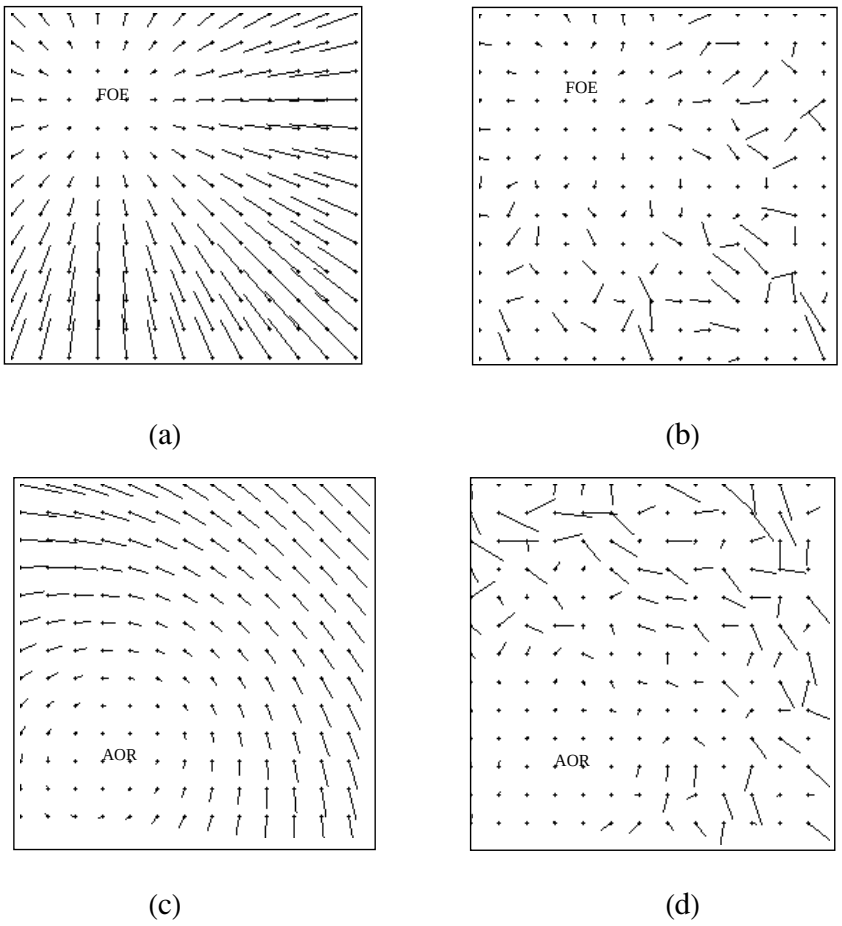


Figure 3.4: Examples of translational and rotational optical and normal flow fields.

identical to the motion field [184]. The algebraic value of normal motion field is equal to

$$\begin{aligned} \left(\frac{dx}{dt}, \frac{dy}{dt}\right) \cdot \frac{(I_x, I_y)}{\sqrt{I_x^2 + I_y^2}} &= \\ \left(\frac{dx}{dt}, \frac{dy}{dt}\right) \cdot \frac{\nabla I}{||\nabla I||} &= \\ \frac{1}{||\nabla I||}\left(I_x \frac{dx}{dt} + I_y \frac{dy}{dt}\right) & \end{aligned} \tag{3.10}$$

The difference between the magnitudes of the normal flow vector and the corresponding normal motion vector, is given by the difference of (3.9) and eqs. (3.10) and is equal to [184]:

$$\frac{1}{||\nabla I||}\frac{dI}{dt} \tag{3.11}$$

Equation (3.11) shows that normal flow is a good approximation of normal motion flow in points where $||\nabla I||$ is large. Consequently, normal flow values are reliable in points where the image

gradient assumes large values. Normal flow vectors at such points can be used as a robust input to 3D motion perception algorithms.

3.1.4 Normal flow field due to motion

Let (n_x, n_y) be the unit vector in the gradient direction. The magnitude u^M of the normal flow vector is given by

$$u^M = un_x + vn_y \tag{3.12}$$

which, by substitution from eq. (3.5), yields:

$$\begin{aligned} u^M &= (-n_x f)\frac{U}{Z} \\ &+ (-n_y f)\frac{V}{Z} \\ &+ (xn_x + yn_y)\frac{W}{Z} \\ &+ \left\{\frac{xy}{f}n_x + \left(\frac{y^2}{f} + f\right)n_y\right\}\alpha \\ &- \left\{\left(\frac{x^2}{f} + f\right)n_x + \frac{xy}{f}n_y\right\}\beta \\ &+ (yn_x - xn_y)\gamma \end{aligned} \tag{3.13}$$

Equation (3.13) highlights some of the difficulties of the problem of independent motion detection. Each image point (in fact, each point at which the intensity gradient has a significant magnitude and, therefore, a reliable normal flow vector can be computed) provides one constraint on the 3D motion parameters. In the case of an observer moving in a static environment, the above equation holds for each point and for one specific unknown set of 3D egomotion parameters (U_E, V_E, W_E), $(\alpha_E, \beta_E, \gamma_E)$. In the case of rigid independent motion, there is at least one more set of motion parameters (U_I, V_I, W_I), $(\alpha_I, \beta_I, \gamma_I)$ that is valid for some of the image points. Furthermore, if no assumption is made regarding the depth Z, each point provides at least one independent depth variable. Evidently, the problem cannot be solved if no additional information is available regarding depth.

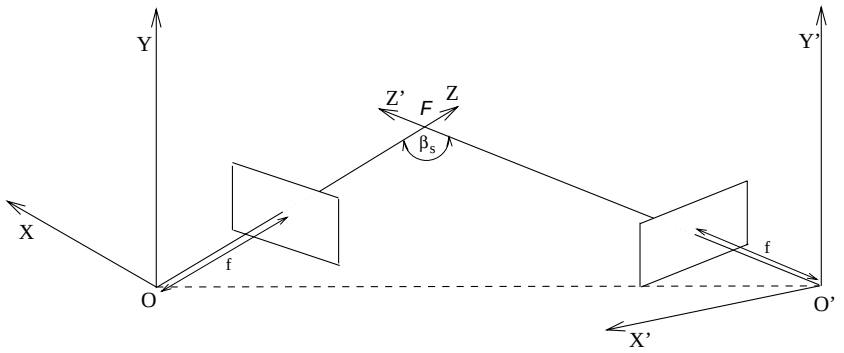


Figure 3.5: The geometry of a fixating stereo configuration.

Equation (3.13) also shows why the problems of egomotion estimation and independent motion detection are considered as chicken-and-egg problems. If egomotion is known, then it can be canceled-out and, therefore, independent motion detection is greatly facilitated. On the other hand, if independently moving objects have not been previously detected, the set of points on which egomotion estimation should rely is unknown. For this reason, most of the proposed techniques for estimating egomotion [64, 11, 154, 44] assume the absence of independently moving objects.

3.1.5 Normal flow field due to stereo

Consider the geometry of a typical stereo configuration of a fixating pair of cameras, as shown in Fig. 3.5. This type of configuration may easily be achieved by existing mechanical heads [131] and is very common in biological organisms, including humans.

A pair of images captured with such a configuration encapsulates information relevant to depth, that manifests itself in the form of *disparities* defined by the displacements of points between images. In [58], Dhond and Aggarwal provide a review of techniques for establishing stereo correspondence. Since the stereoscopic images can be acquired simultaneously, there is no dynamic change in the world that can be recorded by them. It can easily be observed that a stereo image pair is identical to the sequence that would result from a hypothetical (ego)motion

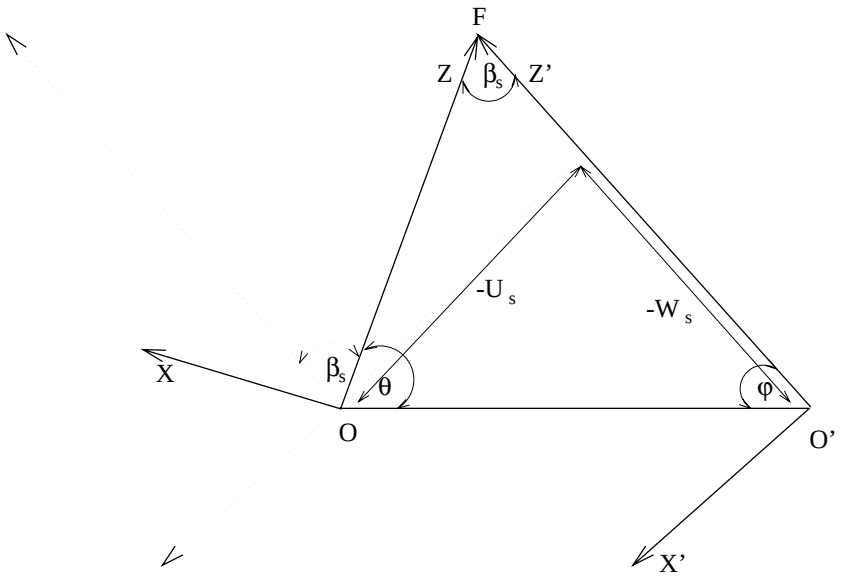


Figure 3.6: The parameters of the motion that transforms the position of the left camera to the position of the right camera. The figure shows a top view of the stereo configuration.

that brings the one camera to the position of the other. This remark enables the analysis of a stereo pair employing motion analysis techniques[1]. The hypothetical motion that transforms the position of one camera to the other is simpler than the one described by the general motion model of eq. (3.5). Fig. 3.6 shows the motion that maps the position of the left camera to the position of the right camera. Evidently, there is no rotation around the X and Z axes, and no translation along the Y axis. Consequently, if $(U_s, 0, W_s)$ and $(0, \beta_s, 0)$ are the translational and rotational parameters of the imaginary motion, then at each point, a normal flow value u^S due to stereo may be computed, which is equal to

$$\begin{aligned} u^S &= (-n_x f)\frac{U_s}{Z} \\ &+ (x n_x + y n_y)\frac{W_s}{Z} \\ &- \left\{\left(\frac{x^2}{f} + f\right) n_x + \frac{xy}{f} n_y\right\} \beta_s \end{aligned} \tag{3.14}$$

In practical situations, the computation of normal flow from a pair of stereo images needs

[1]Regardless if a pair of images is due to a stereo configuration or due to a moving camera, these images can be considered as views of the same scene from different viewpoints.

further consideration. The computation of normal flow is based on the optical flow constraint equation, which does not hold if the two images differ too much. Moreover, normal flow is computed from discrete images through spatial and temporal differentiation with small masks. In the case that 5×5 masks are used, every normal flow that is more than 2 - 3 pixels is not reliable. For the successful computation of normal flow from stereo, two alternative solutions can be proposed:

1. Use of suitable stereo configurations which ensure that normal flow can be robustly computed over the image.
2. Computation of normal flow in lower spatial resolutions.

In the remainder of this section, we will examine each of the above solutions in more detail.

Use of suitable stereo configurations

Equation (3.14) gives the algebraic value of normal flow which depends on the stereo configuration parameters, or equivalently, on the stereo-equivalent motion parameters. Additionally, since the normal flow is the projection of optical flow in a certain direction, the value of normal flow at a certain point, is bounded by the value of optical flow. Thus, we may select suitable stereo configurations for which the maximum obtained optical flow (and consequently the maximum normal flow) does not exceed a certain upper bound. As an example, consider a stereo configuration with a baseline[2] of 7cm. This distance is approximately equal to the distance between the eyes of an adult human. By substituting in eq. (3.5) a typical vergence angle, focal length and image size, we get values of optical flow that do not exceed a number of few pixels, if the minimum depth at the scene exceeds a lower bound.

An interesting special case is the one in which the 3D point at which the optical axes of the cameras intersect (fixation point [66, 42, 52]), lies on the surface of a physical object. In

[2]Refer to Fig. 1.2 for a definition of the parameters of a stereo configuration

such a case, optical flow is zero in a small region around the fixation point and becomes larger as the distance from the center of the image increases. Thus, since the optical flow at a point provides a bound on the algebraic value of the normal flow at the same point, we may restrict the estimation of normal flow in a region around the fixation point.

Lower spatial resolutions

The algebraic value of the optical flow and, consequently, of the normal flow too does not only depend on the motion between the two images and the depth of the scene, but also on the spatial resolution of the images. Therefore, a proper selection of image resolution can be made, so that the magnitude of the optical flow vectors (and therefore the algebraic values of normal flows) is within valid limits. The use of lower spatial resolutions may be useful, if coarse information is to be extracted from stereo for a certain scene.

Lower image resolutions can be used in conjunction with suitable stereo configurations for computing stereo normal flow. Thus, for the case of a fixating stereo, by decreasing the resolution, coarser information can be computed for a larger image area. For the case of a fixating stereo configuration the reduction of resolution can be done in a non-uniform fashion by employing the log-polar image transform [185]. The log-polar transform preserves the original image resolution close to the image center, and reduces it towards image periphery. Thus, it is particularly suitable for computing normal flow in fixated stereo pairs, where optical flow is close to zero near the image center and becomes larger towards the periphery. This idea has also been used by others in order to compute normal flow from a stereo configuration [111].

Another interesting case is the case of a *parallel stereo configuration*, for which the stereo-equivalent motion is a simple translation along the X-axis. Therefore, according to eq. (3.5), the optical flow has just a horizontal component which is given by:

$$u = \frac{-Uf}{Z} \tag{3.15}$$

which, for normal flow yields

$$u^S = (-n_x f)\frac{U_s}{Z} \tag{3.16}$$

From eq. (3.16) it can be seen that the magnitude of the normal flow at a point depends on the stereo baseline length (directly related to U_s) and the scene structure (Z), but not on the coordinates of the point on the image plane. Thus, if a lower bound for the scene structure and an upper bound for the baseline length can be established, an upper bound for the magnitude of normal flow can be guaranteed[3]. If this upper bound is less or equal to the maximum computable normal flow value, then normal flow can be reliably computed for all image points. If a stereo configuration with larger baseline length than the allowable ones is to be used, then again, a proper selection of image resolution can be made, so that the magnitude of normal flow vectors is within valid limits.

3.2 Robust regression

Regression analysis (fitting a model to noisy data) is a very important statistical tool. In the general case of a linear model [144], given by the relation

$$y_i = x_{i1}\theta_1 + \ldots + x_{ip}\theta_p + e_i, \tag{3.17}$$

the problem is to estimate the parameters θ_k, $k = 1, \ldots, p$, from the observations y_i, $i = 1, \ldots, n$, and the explanatory variables x_{ik}. The term e_i represents the error in each of the observations. In classical applications of regression, e_i is assumed to be normally distributed with zero mean and unknown standard deviation. Let $\hat{\theta}$ be the vector of estimated parameters $\hat{\theta_1}, \ldots, \hat{\theta_p}$. Given these estimates, predictions can be made for the observations:

$$\hat{y_i} = x_{i1}\hat{\theta_1} + \ldots + x_{ip}\hat{\theta_p} \tag{3.18}$$

[3] n_x does not violate this assumption since it is a normalized value in the range [0, 1].

Thus, a residual between the observation and the value predicted by the model may be defined as:

$$r_i = y_i - \hat{y}_i \tag{3.19}$$

Traditionally, $\hat{\theta}$ is estimated by the least squares (LS) method, which is popular due to its low computational complexity. LS involves the solution of a minimization problem, namely:

$$Minimize \sum_{i=1}^{n} {r_i}^2 \tag{3.20}$$

and achieves optimal results if the underlying noise distribution is Gaussian with zero mean. However, in cases where the noise is not Gaussian, the LS estimator becomes unreliable. The LS estimator becomes highly unreliable also in the presence of outliers, that is observations that deviate considerably from the model representing the rest of the observations. One criterion for measuring the tolerance of an estimator with respect to outliers is its *breakdown point*, which may be defined as the smallest amount of outlier contamination that may force the value of the estimate outside an arbitrary range. As an example, LS has a breakdown point of 0%, because a single outlier may have a substantial impact on the estimated parameters.

In order to be able to handle data sets containing large portions of outliers, a variety of robust estimation techniques have been proposed. Many of them have been used in computer vision and have been proposed within the vision field [36, 69, 92, 96, 116, 207, 161]. From those, the RANSCAC method [69] is probably the most popular one. Other methods have been borrowed from statistics [32, 102, 115, 144, 156, 160]. Meer et al [115] and Zhang [206] provide excellent reviews of the use of robust regression methods in computer vision.

Probably, the most popular robust estimators are the M-estimators [90]. M-estimators are based on the idea of replacing the squared residuals ${r_i}^2$ by another function of the residuals, yielding

$$Minimize \sum_{i=1}^{n} \rho(r_i), \tag{3.21}$$

where ρ is a symmetric function (i.e. $\rho(-t) = \rho(t)$ for all t) with a unique minimum at zero. It can be shown that minimizing (3.21) is equivalent to solving

$$\sum_{i=1}^{n} \phi(r_i) = 0, \tag{3.22}$$

where ϕ is the first derivative of ρ. Huber's M-estimator [90], uses an *influence function* defined as:

$$\phi(r_i) = \begin{cases} r_i, & \text{if } |r_i| \leq k \\ k, & \text{if } r_i > k \\ -k, & \text{if } r_i \leq -k \end{cases} \tag{3.23}$$

where k is a predefined constant. Since ϕ is a non-linear function, (3.21) needs to be solved by iterative numerical methods. It can be verified that solving (3.22) by using the influence function (3.23) is equivalent to the following weighted least-squares minimization:

$$Min \sum_{i=1}^{n} w_i {r_i}^2, \tag{3.24}$$

where the weights are given as:

$$w_i = \begin{cases} 1, & \text{if } |r_i| \leq k \\ k/r_i, & \text{if } r_i > k \\ -k/r_i, & \text{if } r_i \leq -k \end{cases} \tag{3.25}$$

The Tukey's biweight estimator [144] is another type of M-estimator, which is equivalent to reweighted least squares estimation using the weights:

$$w_i = \begin{cases} (1 - {r_i}^2)^2, & \text{if } |r_i| \leq 1 \\ 0, & \text{if } |r_i| > 1 \end{cases} \tag{3.26}$$

The aim of these influence functions is to protect the estimate from strongly outlying observations. However, M-estimators have two major drawbacks. First, it can be shown that although they behave better that least squares in practical situations, their breakdown point is equal to $1/n$ [144], where n is the number of observations. This approaches zero as n increases[4].

[4]Note that the least squares method is in fact an M-estimator.

Second, it can be shown that they require a reliable initial estimate of the model parameters because otherwise, they can be trapped in local minima.

In an effort to provide robust estimators with a higher breakdown point, Rousseeuw [144] introduced the so-called S-estimators which are defined by minimizing a robust measure of the scatter of the residuals. In general, S-estimators correspond to

$$Minimize_{\hat{\theta}} S(\theta) \tag{3.27}$$

where $S(\theta)$ is a certain type of robust M-estimate of the scale of the residuals $r_i(\theta)$. One special case of S-estimators is the *Least Trimmed Squares(LTS)* estimator, given by

$$Minimize_{\hat{\theta}} \sum_{i=1}^{h} (r^2)_{n:n}, \tag{3.28}$$

where $(r^2)_{1:n} \leq (r^2)_{2:n} \leq \cdots \leq (r^2)_{n:n}$ are the ordered squared residuals. The best robustness properties of LTS are achieved when h is approximately equal to $n/2$, in which case LTS attains a breakdown point of 50%.

Another S-estimator is the *Least Median of Squares(LMedS)* estimator which is described in detail in the next section. LMedS has a breakdown point of 50%, and forms a basic tool for developing the independent motion detection capabilities described in Chapters 4, 5 and 6. It can be demonstrated that 50% is the highest possible breakdown point of an estimator, because for larger outlier contaminations it is impossible to distinguish the ‘‘good’’ from the ‘‘bad’’ data. Recently, a new robust regression method, namely MINPRAN, has been proposed [161] which reports a breakdown point that is higher than 50%. However, MINPRAN makes extra assumptions regarding the distribution of the outliers. More specifically, it assumes a random distribution of the outliers and tries to group data according to a linear model so that the probability of randomness of the grouped data is minimized. Although the concept of MINPRAN is very interesting, it has the disadvantage of a very high computational complexity.

3.2.1 Least Median of Squares (LMedS)

The LMedS method, which was originally proposed by Rousseeuw [143], involves the solution of a non-linear minimization problem, namely:

$$Minimize\{median_{i=1,\ldots,n} {r_i}^2\} \tag{3.29}$$

Qualitatively, LMedS tries to estimate a set of model parameters that best fit the *majority* of the observations, while LS tries to estimate a set of model parameters that best fit all the observations. The above statement gives an idea of the difference in the behavior of the two estimators. The presence of some outliers in a set of observations will not influence LMedS estimation, as long as the majority of the data fit into the particular model. More formally, LMedS has a breakdown point of 50%. Two very characteristic examples of the performance of LMedS relative to LS are shown in Fig. 3.7 (taken from [144]). Figures 3.7(a) and 3.7(b), show two original data sets (five points on the plane) which constitute noisy observations of a straight line. The line fitted in these observations is the line estimated by the least squares method (noted as LS). Figures 3.7(c) and 3.7(d) show the same data sets, but with one of the observations (indicated by a circle) corrupted with high levels of noise. This is actually the outlier of the model. It can be seen (Fig. 3.7(c)) that a single corrupted observation had a substantial impact in the estimation of the least squares method. This deviation is even more profound in Fig. 3.7(d) where the LS estimator completely failed to capture the characteristics of the observations. On the contrary, the LMedS estimation which is shown in Figs. 3.7(e) and 3.7(f) is not influenced by the presence of the outlier.

Once LMedS has been applied to a set of observations, a standard deviation estimate may be derived:

$$\hat{\sigma} = C\sqrt{med{r_i}^2} \tag{3.30}$$

where C is an application dependent constant. Rousseeuw and Leroy [144] suggest a value of

$$C = 1.4826\left(1 + \frac{5}{n-p}\right) \tag{3.31}$$

Based on the standard deviation estimate, a weight w_i may be assigned to each observation

$$w_i = \begin{cases} 1, & \text{if } \frac{|r_i|}{\hat{\sigma}} \leq THR \\ 0, & \text{if } \frac{|r_i|}{\hat{\sigma}} > THR \end{cases} \tag{3.32}$$

All points with weight $w_i = 1$ correspond to model inliers, while points with weight $w_i = 0$ correspond to outliers. The threshold THR controls the sensitivity to outliers. Typically, a value of 2.5 is used. This value reflects the fact that in a Gaussian distribution, very few residuals should be larger than $2.5\hat{\sigma}$. Note that the criterion according to which the labels of inlier and outlier are assigned to data is itself robust, since it involves calculations over the median of residuals. Moreover, the method adapts automatically to the noise levels of the observations. The better the estimated model fits to the observations, the smaller the median residual is and, therefore, the finer the outlier detection becomes.

The computational requirements of LMedS are reported as high [144, 115]. This is because LMedS involves a non linear minimization problem which must be solved by a search in the space of possible estimates generated by the data. However, in Section 4.4, some modifications are proposed, which improve the computational performance of LMedS.

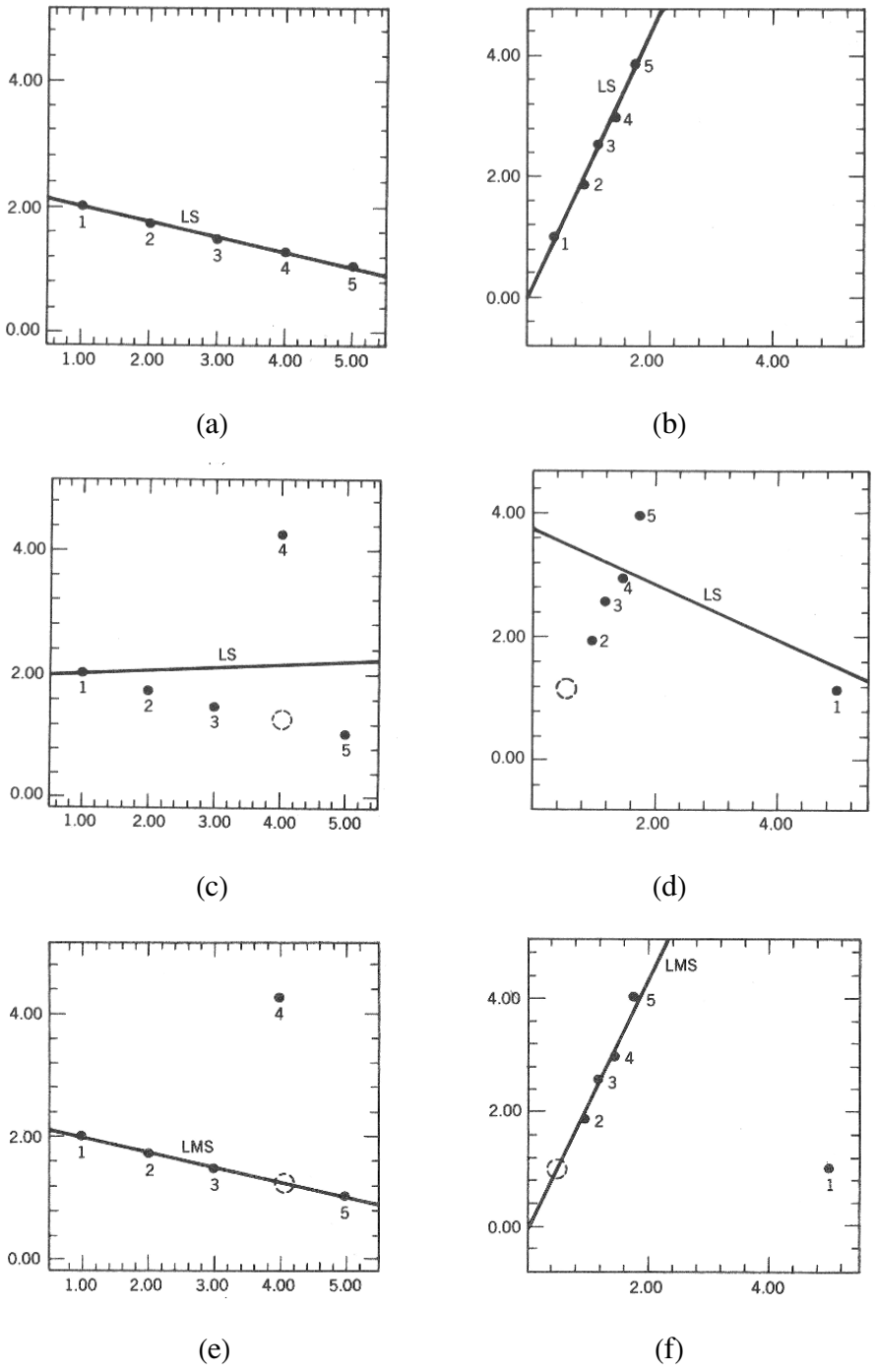


Figure 3.7: Example of least squares (LS) versus least median of squares (LMS) estimation (from [144]).

Chapter 4

Independent Motion Detection Based on Depth Elimination

If among these errors are some which appear too large to be admissible, then those observations which produced these errors will be rejected, as coming from too faulty experiments and the unknown will be determined by means of the other observations which will then give much smaller errors.

Legendre in 1805, in the first publication on least squares

4.1 Method description

Consider a stereoscopic observer that is moving with unrestricted 3D motion in the 3D space. Due to this motion, a reliable normal flow vector can be computed at each point where the image

intensity gradient is large (see eq. (3.11)). The magnitude of this vector is given by eq. (3.13) which is repeated here for convenience:

$$\begin{aligned} u^M &= (-n_x f)\frac{U}{Z} + (-n_y f)\frac{V}{Z} + (xn_x + yn_y)\frac{W}{Z} \\ &+ \left\{\frac{xy}{f}n_x + \left(\frac{y^2}{f} + f\right)n_y\right\}\alpha - \left\{\left(\frac{x^2}{f} + f\right)n_x + \frac{xy}{f}n_y\right\}\beta + (yn_x - xn_y)\gamma \end{aligned} \tag{4.1}$$

Additionally, as it was shown in Chapter 3, a reliable normal flow value due to the stereo configuration can also be computed, with magnitude given by eq. (3.14), which is again repeated here:

$$\begin{aligned} u^S &= (-n_x f)\frac{U_s}{Z} + (xn_x + yn_y)\frac{W_s}{Z} \\ &+ \left\{\left(\frac{x^2}{f} + f\right)n_x + \frac{xy}{f}n_y\right\}\beta_s \end{aligned} \tag{4.2}$$

By solving eq. (4.2) for Z, we obtain:

$$Z = \frac{-n_x f U_s + (xn_x + yn_y)W_s}{u^S - \left[\left(\frac{x^2}{f} + f\right)n_x + \frac{xy}{f}n_y\right]\beta_s} \tag{4.3}$$

In most practical situations, the contribution of the term involving W_s in the above expression is very small[1] compared to the rest of the terms. W_s is usually two orders of magnitude smaller than U_s. Additionally, for specific stereo configurations which are described in Appendix A, eq. (4.3) holds for $W_s = 0$. Thus, this term can be eliminated, giving rise to a simpler expression regarding the depth Z:

$$Z = \frac{-n_x f U_s}{u^S - \left[\left(\frac{x^2}{f} + f\right)n_x + \frac{xy}{f}n_y\right]\beta_s} \tag{4.4}$$

The computation of normal flow involves the computation of the partial image derivatives I_x and I_y, which define the normalized vector (n_x, n_y) at the gradient direction. If, for the computation of both stereo and motion normal flow fields, these derivatives are computed on the same reference frame, then n_x and n_y are the same for both eqs. (4.1) and (4.4). The reference

[1] This is shown in more detail in Appendix A.

frame can be the same if normal flow due to stereo is computed by assuming a left to right transition, and normal flow due to motion is computed in the right image of the stereo pair, from time instant t to time instant $t-1$. Therefore, the substitution of eq. (4.4) to eq. (4.1) results in the following equation:

$$\begin{aligned} u^M &= u^S \frac{U}{U_s} - \left\{ \left(\frac{x^2}{f} + f \right) n_x + \frac{xy}{f} n_y \right\} \left(\frac{U\beta_s}{U_s} - \beta \right) \\ &+ \frac{n_y}{n_x} u^S \frac{V}{U_s} - \frac{n_y}{n_x} \left\{ \left(\frac{x^2}{f} + f \right) n_x + \frac{xy}{f} n_y \right\} \frac{V\beta_s}{U_s} \\ &- \frac{(xn_x + yn_y)u^S}{n_x f} \frac{W}{U_s} \\ &+ \frac{(xn_x + yn_y)\{(\frac{x^2}{f} + f)n_x + \frac{xy}{f} n_y\}}{n_x f} \frac{W\beta_s}{U_s} \\ &+ \left\{ \frac{xy}{f} n_x + \left(\frac{y^2}{f} + f \right) n_y \right\} \alpha + (yn_x - xn_y)\gamma \end{aligned} \tag{4.5}$$

Equation (4.5) is linear on the variables $\phi_1 = \frac{U}{U_s}$, $\phi_2 = \frac{U\beta_s}{U_s} - \beta$, $\phi_3 = \frac{V}{U_s}$, $\phi_4 = \frac{V\beta_s}{U_s}$, $\phi_5 = \frac{W}{U_s}$, $\phi_6 = \frac{W\beta_s}{U_s}$, $\phi_7 = \alpha$, and $\phi_8 = \gamma$. These variables are expressions involving the 3D motion parameters and the stereo configuration parameters. LMedS estimation can be applied to a set of observations of the model of eq. (4.5) as a means to estimate the parameters ϕ_i, $1 \leq i \leq 8$. LMedS will produce a vector of estimated parameters $\hat{\phi}_i$, and a segmentation of the image points into model inliers and model outliers. Model inliers, which are compatible with the estimated parameters $\hat{\phi}_i$, correspond to image points that move with a dominant set of 3D motion parameters. A point may belong to the set of outliers if one (or both) of the following holds:

1. The quantities u^S and/or u^M for this point have been computed erroneously.
2. The 3D motion parameters for this point are different compared to the 3D motion parameters describing the majority of points.

The points of the first class will, in principle, be few and sparsely distributed over the image plane. This is because only reliable normal flow values are considered. The second class of

points is essentially the class of points that are not compatible with the dominant 3D motion parameters. Thus, in the case of two rigid motions in a scene, the inlier/outlier characterization of points achieved by LMedS is equivalent to a dominant/secondary 3D motion segmentation of the scene. In the case that more than two rigid motions are present in a scene, the correctness of 3D motion segmentation depends on the spatial extent of the 3D motions. If there is one dominant 3D motion (in the sense that at least 50% of the total number of points move according to this motion), LMedS will be able to handle the situation successfully. This is because of the high breakdown point of LMedS, which tolerates an outlier percentage of up to 50% of the total number of points. The inlier set will correspond to the dominant motion and the set of outliers will contain all secondary motions. A recursive application of LMedS to the set of outliers may further discriminate the rest of the motions. The recursive application of LMedS should terminate when the remaining points become fewer than a certain threshold. There are two reasons for this. First, if the number of points becomes too small, then the number of constraints provided by eq. (4.5) becomes small and the discrimination between inliers and outliers is subject to errors. Second, at each recursive application of LMedS, the set of outliers does not contain only points that correspond to a motion different than the dominant one, but also points where normal flow values have not been computed accurately.

Note that secondary motion does not necessarily imply independent motion. The problem that can be solved through the processing of visual information is the problem of 3D motion segmentation. This problem is closely related, but not equivalent to the problem of independent motion detection, which, in its general formulation, involves the ability to decide whether a set of 3D motion parameters characterizes the motion of the observer relative to his static background. If the independently moving object moves rigidly and covers most of the visual field of the observer, then LMedS will estimate the parameters of independent motion; inliers will correspond to independent motion and outliers to egomotion. However, the human visual system can also be deceived in similar circumstances[2]. In the context of this work, we assume

[2]An example of such an illusion refers to the case of a human observer, sitting in a train and looking out of his window at another close-by train. If the train he is in starts moving smoothly (so that he has no independent evidence

that egomotion is the motion that has the largest spatial extent, i.e. characterizes the majority of the points in the scene. This is a valid assumption for all practical purposes because in the vast majority of situations, independently moving objects cover a small portion of the visual field. Therefore, the outliers of the linear model of eq. (4.5) correspond to points of independently moving objects and the inliers of this model correspond to points of the static background which appear to be moving due to the egomotion of the observer. In the cases that there is no dominant motion (e.g. there are three motions, each covering 33% of the total number of points), then LMedS cannot handle the situation. As an alternative, other, more complex robust estimators can be used [161], which have higher breakdown points compared to LMedS. However, such estimators rely on probabilistic assumptions regarding the observations of the model and have extra computational overhead. Another possible solution is to apply LMedS in parts of the whole scene, in an effort to change the ratios of the points covered by each motion. Such a decomposition of the scene can be implemented in various ways such as spatial decomposition (e.g. tessellation of the image in rectangular windows and application of LMedS to each of them). However, the decomposition of the scene points can be based on any property. For example, the independent motion detection method that is proposed in Chapter 5 applies LMedS in sets of points that are characterized by an almost constant depth from the observer.

The method for independent motion detection based on depth elimination is presented in pseudocode in Fig. 4.1. In this algorithm, procedure Discriminate_With_LMedS (step 2.1) applies the LMedS estimation technique on the $WorkingSet$ set of observations, and splits it into the $InlierSet$ and $OutlierSet$. The $WorkingSet$ is initialized with the set S of all points in the image where reliable normal flow vectors due to motion and stereo have been computed. The $InlierSet$ includes the points that have a dominant motion at each stage of computations, while the $OutlierSet$ contains the points that should be further processed. The constant $POINTS\_LIMIT$ controls the termination of the iterative application of LMedS. The points that are left unclassified after the last iteration (fewer than $POINTS\_LIMIT$) are

regarding his own motion) he will perceive the other train as moving. Clearly, deciding which of the perceived motions corresponds to egomotion does not rely on visual information only.

0. WorkingSet := S

1. Current_Motion_Segment_ID := 1

2. **While** NumberOfPointsIn(WorkingSet) < POINTS_LIMIT **do**

    2.1 Discriminate_With_LMedS(WorkingSet, InlierSet, OutlierSet)

    2.2 $M^0{}_{Current\_Motion\_Segment\_ID}$:= InlierSet

    2.3 WorkingSet := OutlierSet

    2.4 Current_Motion_Segment_ID := Current_Motion_Segment_ID + 1

3. **EndWhile**

Figure 4.1: Algorithm for independent motion detection based on depth elimination.

attributed to noise and are discarded from further consideration. At the end of the algorithm, each set variable[3] $M^0{}_i$ holds the points belonging to one rigid, 3D motion, from now on referred to as *motion segment*.

4.1.1 Postprocessing

According to the proposed method for independent motion detection based on depth elimination, points are characterized as being independently moving or not based on their conformance to a general rigid 3D motion model. The characterization is made at the point level, without requiring any environmental assumptions, such as smoothness, to hold in the neighborhood of each point. In order to further exploit information regarding independent motion, we would

[3]The superscript 0 in the set variable $M^0{}_i$ denotes that these motion segments refer to the whole image. It is used for notational compatibility with the next chapter, where motion segments will be defined over subsets of points in the scene.

like to refer to connected, independently moving areas rather than to isolated points. There are several reasons why the points of a motion segments do not form connected regions in the segmentation provided by LMedS:

- The normal flow field is usually a sparse field, because normal flow values are considered unreliable in certain cases (e.g. in points with a small gradient value). Thus, there are points in the scene for which no decision can be made regarding their motion.

- Despite the fact that normal flow can be robustly computed, there is always a possibility for misclassifications because some points may become model inliers (or outliers) due to errors in the measurement of normal flow values u^S and u^M and not due to their 3D motion parameters.

- Normal flow is a projection of the optical flow onto a certain direction. Infinitely many other optical flow vectors have the same projection on this direction. Consequently there may be the case that a normal flow is compatible with the parameters of two different 3D motions, and therefore a number of point misclassifications may arise.

We overcome the problem of disconnected motion segments by exploiting the fact that, in the above cases, misclassified points are sparsely distributed over the image plane. A simple majority voting scheme is used. At a first step, the number of inliers and outliers is computed in the neighborhood of each image point. The label of this point becomes the label of the majority in its neighborhood. This allows isolated points to be removed. In the resulting map, the label of the outliers is replicated in a small neighborhood in order to group points of the same category into connected regions.

4.2 An interesting side effect: Egomotion estimation

What is of primary importance for the task of independent motion detection is the segmentation of scene points into motion segments. However, besides the inlier/outlier characterization, LMedS provides estimations $\hat{\phi}_i$ for the parameters ϕ_i of the linear model of eq. (4.5). Each of the model parameters ϕ_i corresponds to expressions involving the 3D motion parameters of the observer [(U, V, W) and (α, β, γ)] and the stereo configuration parameters (U_s, β_s). Thus, the observer is able to relate his own motion parameters to the parameters of his stereo configuration, i.e. to parameters of his own body. This is in agreement with biological experiments [101] according to which biological organisms perform measurements in units that are related to the physiology of their body and to their own mechanical capabilities, rather in some absolute scale that is conventionally defined.

Additionally, the estimated parameters $\hat{\phi}_i$ can also be used to provide knowledge on the 3D motion parameters of the observer. More specifically, the following relations hold:

$$x_0 = \frac{\phi_1 f}{\phi_5},\ y_0 = \frac{\phi_3 f}{\phi_5}$$
$$\alpha = \phi_7,\ \beta = \phi_2 - \phi_1 \frac{\phi_4}{\phi_5},\ \gamma = \phi_8$$

where, x_0 and y_0 are the coordinates of the FOE. Thus, the estimated parameters $\hat{\phi}_i$ of the model, provide knowledge on the observer's motion parameters. Similarly, an estimation of the vergence angle of the stereo configuration is possible:

$$\beta_s = \frac{\phi_4}{\phi_3}$$

4.3 The case of purely rotational egomotion

Consider an observer performing a rigid rotational 3D motion only. Such cases can be detected by using the Collinear Point Constraint (CPC) that has been proposed by Lobo and Tsotsos [108].

The CPC depends only on the translational motion of the observer. Therefore, if it is equal to zero, then the motion of the observer can only be due to a purely rotational motion.

From eq. (3.13) it can be verified that only the translational component of motion is influenced by the depth of the scene. Therefore, in the case of purely rotational egomotion, the part of the independent motion detection method which deals with stereo information is redundant. Independent motion detection may be achieved by just applying the LMedS estimation on the motion normal flow field, whatever the scene structure may be. This leads to a much simpler model (three model parameters rather than eight) and lower computational requirements. However, in practical applications, the case of pure rotational egomotion is very rare.

4.4 Implementation and performance issues

Figure 4.2 summarizes schematically the method for independent motion detection based on depth elimination.

At time t, a pair of images L_t (left) and R_t (right) is acquired by the stereo configuration. Both images are smoothed giving rise to images SL_t and SR_t, respectively. Smoothing is achieved by convolving the input images with a 5×5 Gaussian kernel with standard deviation $\sigma_G = 1.4$. The filter mask is shown in Fig. 4.3. Image smoothing is a characteristic example of low level, data parallel algorithms [194, 167]. The parallelization of such algorithms has been extensively studied [162], and efficient parallel algorithms have been developed for both SIMD [57, 55, 106] and MIMD [60] parallel architectures.

The normal flow field due to the stereo configuration is then computed. To do so, the spatiotemporal derivatives of the image intensity function are computed. Without loss of generality, we compute the temporal derivatives assuming a left-to-right image transition. Based on the interpretation of eq. (3.11), the image intensity gradient is thresholded; all points for

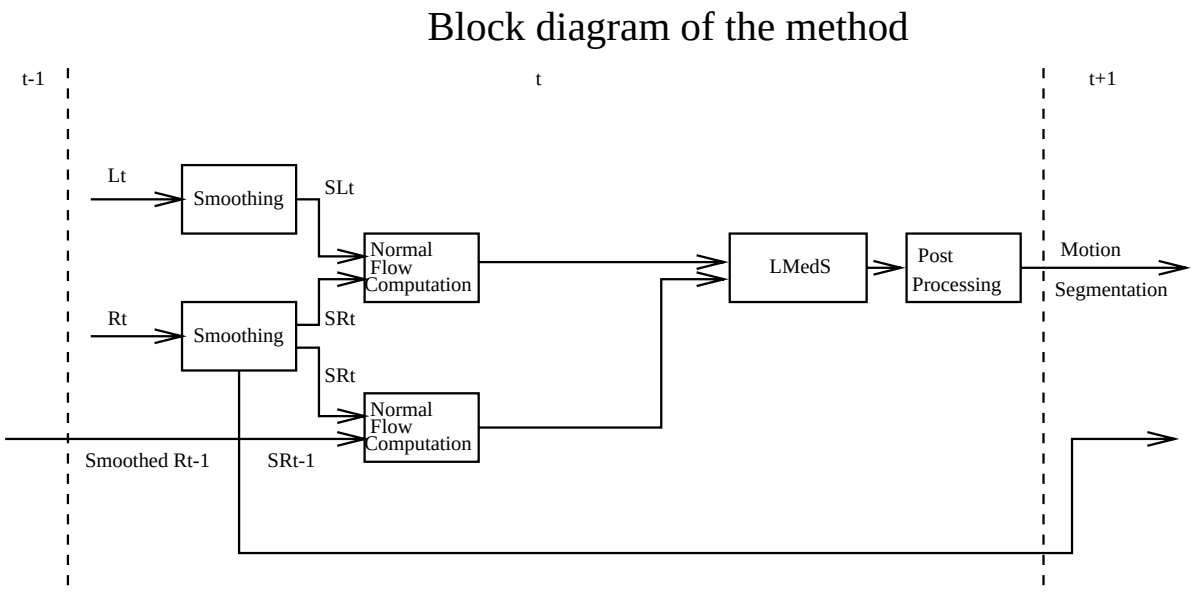


Figure 4.2: A schematic presentation of the method for independent motion detection based on depth elimination.

which the magnitude of intensity gradient is lower than a threshold are removed from further consideration. In a way completely analogous to that of stereo, the motion normal flow field is computed from smoothed images SR_{t-1} and SR_t. The computation of normal flow (either stereo normal flow or motion normal flow), also belongs to the class of low level, data parallel algorithms. Normal flow computation can be decomposed into three independent stages which correspond to the computation of the two spatial and the one temporal derivative of the image intensity function. Each of these stages involves some local mask correlation, which has the same general computational characteristics as image smoothing. In particular, the image gradient is computed by convolving the image with the Sobel operators [73] (masks of Fig. 4.4). The time derivative is computed by subtructing averaged (in 3×3 windows) intensities from the successive frames. The spatial and temporal derivatives are combined through eq. (3.9) to give the normal flow values (for both stereo and motion normal flow).

After normal flow computation, a significant data reduction takes place, because some points are rejected from further consideration due to the low magnitude of intensity gradient.

$$\frac{1.0}{84.0} * \begin{array}{|c|c|c|c|c|} \hline 1.0 & 2.0 & 3.0 & 2.0 & 1.0 \\ \hline 2.0 & 5.0 & 6.0 & 5.0 & 2.0 \\ \hline 3.0 & 6.0 & 8.0 & 6.0 & 3.0 \\ \hline 2.0 & 5.0 & 6.0 & 5.0 & 2.0 \\ \hline 1.0 & 2.0 & 3.0 & 2.0 & 1.0 \\ \hline \end{array}$$

Figure 4.3: The filter used for Gaussian smoothing of images.

-1.0	0.0	1.0
-2.0	0.0	2.0
-1.0	0.0	1.0

(a)

-1.0	-2.0	-1.0
0.0	0.0	0.0
1.0	2.0	1.0

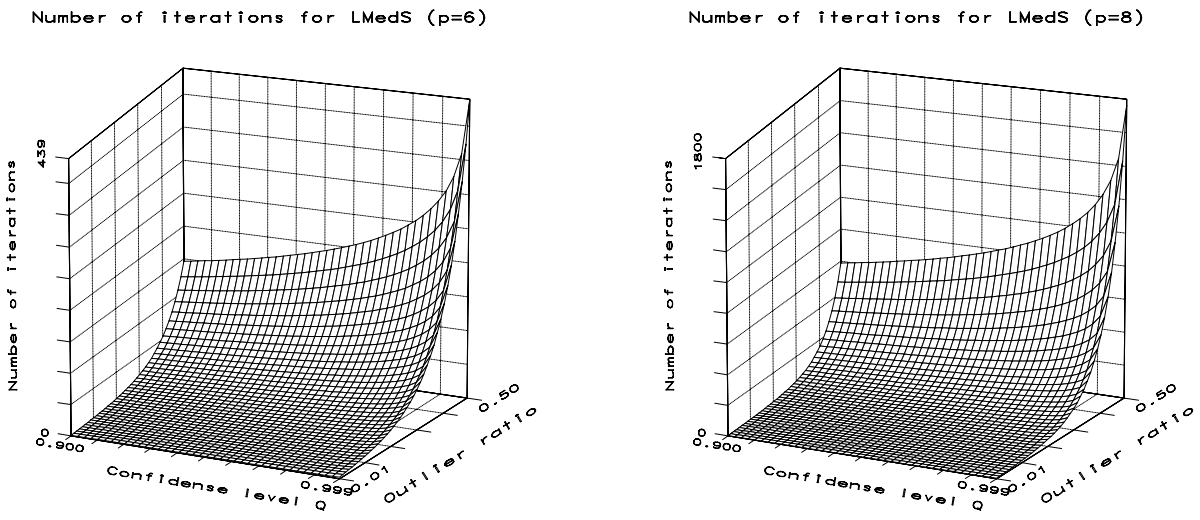
(b)

Figure 4.4: The filters for computing image gradient.

In $N \times N$ images, we may compute n reliable normal flow values ($0 \leq n \leq N^2$). Motion segmentation is now performed based on LMedS estimation over the model of eq. (4.5). LMedS estimation, is the most computationally intensive part of the independent motion detection scheme. LMedS minimization cannot be reduced to a least-squares based solution, but must be solved by a search in the space of possible estimates generated by the data. Let p denote the number of parameters to be estimated; then there are $O(n^p)$ possible p-tuples. Because this search space may become prohibitively large, in practical situations, a Monte-Carlo type of speedup technique is employed [115], in which a certain probability of error is tolerated. If e is the fraction of data contaminated by outliers, then the probability Q that at least one out of m p-tuples has only uncorrupted observations is equal to:

$$Q = 1 - [1 - (1 - e)^p]^m \tag{4.6}$$

Thus, the solution of eq. (4.6) for m, gives a lower bound for the number of p-tuples that should be tried. Note that eq. (4.6) is independent of n. Each of the m trials, requires the



(a) (b)

Figure 4.5: Number of iterations m for LMedS estimation (eq. (4.6)), as a function of Q and e. The number of model parameters is kept equal to (a) p = 6, and (b) p = 8.

selection of candidate parameter values and the computation of the squared residuals between the observations and the predictions of the model. Figures 4.5(a) and 4.5(b) show a 3D plot of the number of required iterations m (for $p = 6$ and $p = 8$, respectively) as a function of the probability Q and the outliers probability e. In our case, $p = 8$, and if a confidence level of 99% is required, then $Q = 0.01$. In the case that no more than 20% of the points are expected to be independently moving, then $e = 0.2$ and solving eq. (4.6) for m gives a value of $m = 26$ iterations. If, however, the ratio of outliers is raised to its maximum value $e = 0.5$, then the number of minimum iterations required becomes 1177. These figures hold for the case that in each iteration, p observations are randomly selected, a linear system is solved and the median of the squared residuals is tested against the minimum of the squared residuals computed so far. Alternatively, candidate solutions can be formed by the results of least squares parameter estimation in rectangles of random dimensions and locations over the input image. In this case, all points with reliable normal flow values contribute to the least squares solution. Both approaches were tested. With the first approach (linear system solution)

a number of computationally cheap iterations are required. The second approach (least squares solution) requires fewer, but computationally more expensive iterations. Experimental results demonstrated that the overall performance is better for the second approach because of the significant reduction of the number of iterations required.

Another algorithmic improvement is achieved through avoiding the use of sorting to compute the median in each of the m iterations. Instead, we use an algorithm that selects the kth largest number out of n numbers, originally suggested in [138, 148]. This algorithm has a time complexity of $O(n)$, rather than the $O(n \log n)$ complexity of the best serial sorting algorithm. Thus, the overall computational complexity of LMedS becomes $O(mn)$.

The performance of LMedS can be greatly improved by exploiting parallel processing techniques. Each of the m iterations does not depend on the outcome of another iteration. All candidate solutions could be evaluated in parallel. An extra stage for comparing the partial results is needed, which amounts to the problem of finding in parallel the minimum in a set of numbers. Going one step further, in each of the m iterations, the computation of the residual for each observation does not depend on the computation of the residual for another observation. Therefore, the n residuals could be computed in parallel.

The resulting motion segments are post-processed towards enhancing the usability of the resulting 3D motion segmentation map. The post processing of the motion segments is again a local operation for which efficient parallel algorithms can be developed.

One interesting computational characteristic of the overall method for independent motion detection stems from the computational nature of LMedS. Through the various iterations over possible parameter sets, LMedS keeps the best solution (the one that minimizes the median of the residuals). Therefore, the whole algorithm can be viewed as an *any time algorithm* (i.e. an algorithm that continuously improves the solution to a problem). This characteristic is very important because the execution time of the algorithm can be appropriately adjusted, taking into account the resource limitations of the system and the hard real time constraints that should be

met.

Additional computational savings can be gained by considering a stereo configuration that does not change over time. In such a case, after few time instances, a robust estimate of β_s can be achieved. Given this estimate, the 8-parameter model of eq. (4.5) can be reduced to a simpler model with fewer parameters. This simplification of the model, leads to lower computational requirements. Similar simplifications can be achieved by considering simpler motion models (e.g. purely translational or purely rotational motion) or a simpler stereo configuration (e.g. a parallel one).

The presented method for independent motion detection owes much of its robustness to the fact that, in all the intermediate stages, computations are applied on image content representations that remain very close to the original image data. According to the recent theories of active and purposive vision [13, 132, 27, 23], this is a very desirable characteristic. Motion segmentation is achieved through a short chain of simple modules. No high level symbolic processing is employed. In fact, the only symbols used are related to the final goal (motion segmentation), and even these are assigned at the pixel level. The above characteristics enhance robustness and make possible the exploitation of parallel processing techniques towards achieving real time performance.

Chapter 5

Independent 3D Motion Detection Through Robust Regression in Depth Layers

In whatever concrete theory, there exist as much science as its mathematics

Immanuel Kant

5.1 Overview

In the previous chapter, a method for independent motion detection was proposed, which was based on the elimination of the depth variable Z from the relation describing the motion normal flow u^M. This elimination has been achieved through the use of additional information in the form of stereo normal flow measurements u^S. The elimination of depth resulted in a

linear system of eight parameters. The LMedS estimation of the model parameters leads to a segmentation of the scene according to the 3D motion parameters of its points. The elimination of depth enables the solution of the independent motion detection problem, without actually computing the depth. This is very important because:

- The accurate computation of depth is very difficult [59].
- The computation of depth cannot be achieved without knowledge of the stereo configuration parameters.

The method proposed in this chapter [18] also avoids the direct computation of depth. However, stereoscopic information is exploited in a different way. In order to clarify the idea exploited by the current method, consider eq. (3.13) which gives the algebraic value of motion normal flow that has been computed from a pair of successive images in time. This equation forms a linear model, when the depth Z and the motion parameters (U, V, W) and (α, β, γ) are constant for all image points. In terms of LMedS estimation, outliers of the linear model will be points for which one, or more, of the following holds:

1. Their depth Z deviates from a dominant depth.
2. Their 3D motion parameters are different from the dominant motion parameters.
3. Noise has been introduced in the computation of normal flow.

For the task of independent motion detection, the second case is of interest. If it is possible to define subsets of observations that correspond to points with the same depth and restrict the application of LMedS in each of these subsets, then outliers should be due to independent motion only. Points where noise was introduced in the computation of normal flow are not of great concern since few such points are expected (only reliable normal flow vectors are considered) and, moreover, they are sparsely distributed over the whole image plane.

5.2 Method description

It is now possible to delineate the following algorithm for independent motion detection, for the case of unrestricted rigid 3D egomotion:

1. Segment the set of image points into *depth layers*, i.e. subsets with approximately constant depth from the observer.
2. Apply the LMedS robust estimation technique to each depth layer in order to identify motion outliers.
3. Combine the results from all depth layers in order to get a global 3D motion segmentation of the scene.

In the remainder of this section, we are going to resolve issues related to each of the steps of the algorithmic scheme described above.

5.2.1 Layering of a scene with respect to depth

Let S be the set of all points p_i, $1 \leq i \leq n$, of the image plane for which reliable normal flow values can be computed. Each point p_i of the image corresponds to a point P_i of the 3D world, whose distance from the observer is Z_i. Each point p_i belonging to S may define a *depth layer* L_i, i.e. a subset of S based on the following relation:

$$L_i = \left\{ p_j : \left| \frac{Z_i - Z_j}{Z_i} \right| < \epsilon \right\} \tag{5.1}$$

Each of the layers L_i can be interpreted as a ‘‘slice’’ of the 3D space, that contains 3D points within a range of depths from the observer. The farthest from the observer, the thicker the depth layers become, for the same value of parameter ϵ. If the parameter ϵ is selected to be sufficiently small, then the depth variations within a specific depth layer are small compared to the distance

from the observer and, the depth variable can be considered as constant:

$$\forall p_j \in L_i, Z_j \simeq C_i. \tag{5.2}$$

In eq. (5.2) C_i is a constant dependent on the layer L_i. The segmentation of a scene S in layers L_i according to the depths of the 3D points from the observer can be achieved either by the application of LMedS on the normal flows computed from a stereo configuration [20, 21], or by considering specific functions of normal flow values. Both approaches are described in the following sections.

Depth layering through robust regression

Consider eq. (3.14) which gives the algebraic values of stereo normal flow vectors that can be computed between two images that are acquired by a binocular imaging system. Since the acquisition of the two stereo images is done at the same time, both images capture the static characteristics of the environment, without any dynamic change (such as independent motion) to be cast. Moreover, as it has been shown in Chapter 3, the stereo-equivalent motion parameters U_s, W_s and β_s, are related to the parameters of the stereo configuration (baseline and vergence angle) and therefore they are the same for all points where normal flow vectors have been computed.

For the above reasons, a set of observations in the form of eq. (3.14) could be fed in a LMedS module as a means to estimate the parameters of the stereo configuration. Equation (3.14) forms a linear model if depth Z is constant over the image plane. Suppose that at least 50% of the image points are at an almost constant depth from the observer. In terms of the depth layer definition of the previous section, this means that at least 50% of the points belonging to S, form one of the depth layers L_i, denoted by L_d. In such a case, all points belonging to L_d will form the stereo inliers, while the rest of the points will be the stereo outliers. The estimation of the stereo configuration parameters through robust regression leads, as a side effect, to the estimation of the *dominant depth layer* L_d. It should be stressed that although LMedS is applied

to estimate the stereo configuration parameters, the estimated parameters are not used in any of the remaining computational steps. What is actually used is the discrimination of the dominant depth layer.

LMedS can be recursively applied to the set of points $L - L_d$ in order to detect the second largest depth layer. This is much alike the recursive application of LMedS in the case of independent motion detection algorithm of Chapter 4. In that context, the recursive application of LMedS was aiming at the detection of motion segments.

The algorithm for depth layering through robust regression is presented in pseudocode in Fig. 5.2.1.

---

0. WorkingSet := S

1. Current_Layer := 1

2. **While** NumberOfPointsIn(WorkingSet) < POINTS_LIMIT **do**

    2.1 Discriminate_With_LMedS(WorkingSet, InlierSet, OutlierSet)

    2.2 $L_{Current\_Layer}$:= InlierSet

    2.3 WorkingSet := OutlierSet

    2.4 Current_Layer := Current_Layer + 1

3. **EndWhile**

Figure 5.1: Algorithm for depth layering based on robust regression.

In this algorithm, procedure Discriminate_With_LMedS (step 2.1) applies the LMedS estimation technique to the $WorkingSet$ set of observations, and splits this set into the $InlierSet$ and $OutlierSet$. The $WorkingSet$ is initialized with the set S of all points at which reliable normal flow vectors have been computed. The $InlierSet$ contains points of the dominant depth layer at

each iteration while the $OutlierSet$ contains points that should be further processed. The constant $POINTS\_LIMIT$ controls the termination criterion of the iterative application of LMedS. The points that are left unclassified after the last iteration (fewer than $POINTS\_LIMIT$) are attributed to noise and are discarded from further consideration.

Direct depth layering

In the previous section, it has been demonstrated how depth layers may be extracted through robust regression. Under such a formulation of the problem, depth layering is achieved as a side effect of the process of estimation of the stereo configuration parameters, although the parameters computed are not directly used. In this section, we describe an alternative approach to depth layering. This approach can be characterized as *direct* in the sense that it surpasses the problem of solving for the stereo configuration parameters and tries to extract information about depth based on specific functions of normal flow. Such functions can be acquired by considering a parallel stereo configuration for which the stereo-equivalent motion is only a translation parallel to the x-axis of the camera coordinate system. Consequently, the normal flow that can be computed from the stereo pair is equal to:

$$u^S = (-n_x f)\frac{U_s}{Z}$$

which can be written equivalently as:

$$-\frac{u^S}{n_x} = \frac{U_s f}{Z} \tag{5.3}$$

Equation (5.3) expresses the fact that for all image points we may compute a function of the form

$$f(Z) = \frac{A}{Z}$$

where $f(Z) = -\frac{u^S}{n_x}$ is a quantity that can be estimated from the stereo pair at each point, and $A = U_s f$ is a constant depending on the stereo configuration and the camera. Suppose now that we want to check whether a point p_j belongs to the layer of a point p_i. According to eq. (5.1), it

is required that:

$$\left|\frac{Z_i - Z_j}{Z_i}\right| < \epsilon \Leftrightarrow \left|\frac{\frac{Z_i}{A} - \frac{Z_j}{A}}{\frac{Z_i}{A}}\right| < \epsilon \Leftrightarrow \left|\frac{\frac{1}{f(Z_i)} - \frac{1}{f(Z_j)}}{\frac{1}{f(Z_i)}}\right| < \epsilon \Leftrightarrow$$

$$\left|1 - \frac{f(Z_i)}{f(Z_j)}\right| < \epsilon \tag{5.4}$$

Therefore, since $f(Z_i)$ and $f(Z_j)$ are computable quantities, we can decide whether two points p_i and p_j belong to the same layer or not. The criterion of eq. (5.4) does not depend on the parameter A, meaning that knowledge of the stereo baseline and the camera focal length is not required.

Comparison of the two depth layering schemes

The main advantage of depth layering based on robust regression is that it can be used regardless of the parameters of the stereo configuration. However, due to the nature of LMedS, a certain relation should hold among the populations of the depth layers. The direct method for depth layering does not impose any assumptions regarding the number of points in each layer but can only be used in the case of a parallel stereo configuration. Consequently, the selection between these two methods should be done based on a detailed analysis of the requirements of an application. As a general remark, the direct approach to depth layering is considered preferable, in the sense that it is related to assumptions regarding the body of the observer (i.e. the parameters of its stereo configuration) and not related to assumptions about the environment (i.e. constraints on the populations of the depth layers).

5.2.2 Motion segmentation of a depth layer

Having already segmented a scene into depth layers, the goal is now to segment each of these layers based on its 3D motion characteristics. From the process of depth layering, it is known that depth variations within a layer are very small compared to the depth from the observer.

Thus, eq. (3.13) is linear with respect to the motion parameters (U, V, W) and (α, β, γ), for the points of a layer and LMedS can be used in order to estimate the dominant 3D motion parameters in this layer. LMedS is actually applied in order to estimate the parameters $(\frac{U}{C_i}, \frac{V}{C_i}, \frac{W}{C_i})$ and (α, β, γ), where C_i is the average depth of each depth layer (see eq. (5.2)). Model inliers will correspond to points of a dominant 3D motion. Model outliers will correspond to points where either normal flow values have been erroneously calculated or to points where the underlying 3D motion parameters are not equal to the ones of the dominant motion. As in the case of the independent motion detection method of Chapter 4, we use the term motion segments to refer to sets of points that are characterized by common 3D motion parameters. Theoretically, up to 50% of outliers can be tolerated in each depth layer. In the case of depth layers with at most two rigid motions, motion segmentation can be successfully achieved, since the one or the other motion will dominate and will be estimated by LMedS estimation. In the case of more than two rigid motions in a depth layer, the segmentation may be recursively applied to distinguish all different motions. However, at each level of estimation, at least 50% of the total number of points should have a common motion parameter set.

5.2.3 Integration of results from the various layers

The motion segmentation of a layer L_i produces μ motion segments ${M_1}^i, {M_2}^i, ..., {M_\mu}^i$, each of which is characterized by a set of parameters $(\frac{U}{C_i}, \frac{V}{C_i}, \frac{W}{C_i})$ and (α, β, γ). In some cases, this type of information is enough to serve the goals of the observer. For example, the observer may be interested to focus his attention to the different 3D motions at a specific depth. If, however, a 3D motion segmentation of the whole scene is required, it should be examined whether two motion segments belonging to different depth layers correspond to the same 3D motion. Unfortunately, the estimated parameters are not pure 3D motion parameters because the translational components of the estimated vectors include also information about depth. Any direct comparison of the estimated parameter vectors across different depth layers is invalid unless additional, quantitative information on depth is available. Moreover, the combination

of results cannot be achieved on the basis of the inlier or outlier characterization of the scene points, because the dominant motion in one layer may appear as a secondary motion in another layer or because one depth layer may contain just one motion segment and no model outliers can be found.

The task of parameter comparison is tackled by reducing the dimensionality of the problem. From each 6-tuple of estimated parameters $(\frac{U}{C_i}, \frac{V}{C_i}, \frac{W}{C_i}, \alpha, \beta, \gamma)$ we get a 5-tuple $(m_1, m_2, m_3, m_4, m_5) = (\frac{U}{W}, \frac{V}{W}, \alpha, \beta, \gamma)$ by dividing the first two coordinates of the 6-tuple with the third one. This 5-tuple depends only on the 3D motion parameters. Therefore, it forms a basis for deciding whether to merge motion segments in different depth layers. Consider the two motion 5-tuples $({m_1}^a, {m_2}^a, {m_3}^a, {m_4}^a, {m_5}^a)$ and $({m_1}^b, {m_2}^b, {m_3}^b, {m_4}^b, {m_5}^b)$ of motion segments a and b, respectively. These motions are considered identical if:

$$\left| \frac{{m_i}^a - {m_i}^b}{max\{{m_i}^a, {m_i}^b\}} \right| < \delta_m, \forall i, 1 \leq i \leq 5, \tag{5.5}$$

where δ_m is a threshold controlling the sensitivity to motion discrimination. According to the criterion of eq. (5.5), two 3D motions are considered identical if they have the same rotational components and the same FOEs. This is not always correct, because in principle, different translational motions can have the same FOE (i.e. all motions of the form (kU, kV, kW), $\forall k \in R$ have the same FOE). However, such cases are rare in practice and, moreover, cannot be tackled without using metric depth information.

In practical situations, the motion 5-tuples compared are not the estimates provided by LMedS. LMedS is recursively applied until all motion segments are identified. Then the 5-tuple of each motion segment is computed by least squares estimation over the whole set of points of this segment. This is because in the absence of outliers, least squares gives more accurate estimates of the model parameters compared to LMedS estimation. It should be stressed, however, that having already segmented a layer with respect to its motion parameters, all algorithms that can solve the egomotion problem would suffice to estimate the motion parameters of a specific segment and subsequently aid towards motion parameter comparison.

The integration of results can be viewed as a problem of connected components labeling [133]. Consider an undirected graph $G = (V, E)$, where the set V of vertices contains all motion segments of all depth layers. Each vertex $v_i \in V$ is characterized by the motion quintuple of the corresponding motion segment. An undirected edge $e(v_1, v_2) \in E$ connects two vertices if the motion quintuples of the vertices satisfy the criterion (5.5). Connected components of graph G correspond to regions of the scene that have common three dimensional motion parameters, independent of depth. Note that two motion segments are not compared if they belong to the same layer and, therefore the corresponding vertices in V cannot be connected. Note also that a connected component of G does not necessarily correspond to connected regions in the image.

5.3 Ordering the depth layers

In the context of the purposive vision paradigm, there is an increasing interest on ordinal (as opposed to metric) representations of depth. Researchers [95, 171, 111] have argued that for a number of tasks, ordering the scene points with respect to depth, is simpler than trying to provide metric information and suffices for solving interesting problems. In this section we present how the proposed method for independent motion can be used to provide such representations of the scene structure. However, instead of providing an ordering of the scene points with respect to depth, we provide an ordering of the depth layers. The ordering of the layers is more robust, because it is achieved through the processing of groups of points rather than isolated points. Additionally, interesting questions of qualitative nature about the depths of objects (e.g. which object is closer to me?) may be answered through reasoning about the relation of the depth layers in which the objects reside.

Suppose that after applying the independent motion detection method, two motion segments a and b that reside in two different layers i and j, respectively, were assigned to the same 3D motion[1]. Let $(\frac{U}{C_i}, \frac{V}{C_i}, \frac{W}{C_i}, \alpha, \beta, \gamma)$ be the estimated sixtuple for the motion segment a. Since the

[1]We ignore the case of two 3D motions with the same FOEs and rotational components at different layers

motion segment b has the same 3D motion parameters, its sixtuple should be $(\frac{U}{C_j}, \frac{V}{C_j}, \frac{W}{C_j}, \alpha, \beta, \gamma)$ By dividing the first three coordinates of the sixtuples one by one, we get $(\frac{C_i}{C_j}, \frac{C_i}{C_j}, \frac{C_i}{C_j})$. Each of the coordinates of this vector, gives the relative depth of the two layers. Even in the case that due to errors, the relative depth is not accurate, the comparison of these ratios with unity can provide information about the ordering of layers in space. Thus, by exploiting 3D motions that are present in more than one depth layers, it is possible to get qualitative information about the scene structure.

5.4 Ambiguities in independent 3D motion detection

A very interesting issue regarding any independent motion detection method, is related to its capability to distinguish among different 3D motions. Two 3D motions can be defined as *ambiguous* with respect to an independent motion detection method, if the method cannot distinguish them, even in the case of perfect sensory input. One class of ambiguous motions for the independent motion detection through depth layering is the set of motions that have the same rotational parameters and the same direction of translation. This type of ambiguity stems from the dimensionality reduction that aided integration of results across the depth layers. The simple example of Fig. 5.2 clarifies this issue. The figure shows an optical flow field that consists of two regions, A and B. Optical flow vectors in each of the regions have a constant direction and magnitude. However, optical flow vectors in region B are larger than those in A. By observing this scene, three different interpretations may be given:

1. Regions A and B have the same motion parameters but different depths. More specifically, region A is at a larger depth than region B.

2. Regions A and B are at the same depth, but have different motion parameters. More specifically, region B moves faster than region A.

3. Regions A and B have different motion parameters and correspond to different depths.

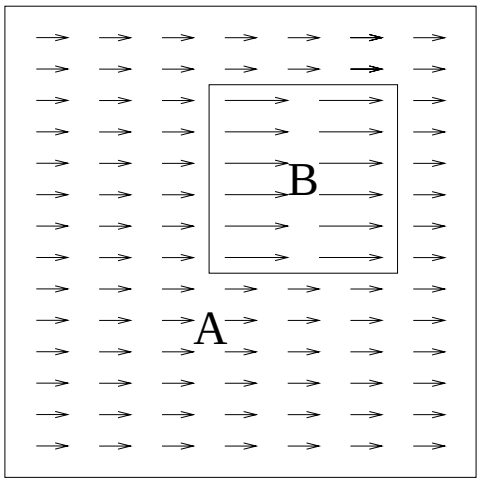


Figure 5.2: An example of a motion field that is ambiguous in terms of motion segmentation.

The proposed method can tackle the second case, because in this case two different motion segments will be part of the same depth layer. However, it cannot discriminate between the cases 1 and 3, since it lacks any knowledge of the relative depths of the depth layers.

5.5 The case of purely rotational egomotion

Consider an observer performing a rigid rotational 3D motion only. Such cases can be detected by using the Collinear Point Constraint (CPC) that has been proposed by Lobo and Tsotsos [108]. The CPC depends only on the translational motion of the observer. Therefore, if it is equal to zero, then the motion of the observer can only be due to a purely rotational motion.

From eq. (3.13) it can be verified that only the translational component of the motion is influenced by the scene structure. Therefore, in the case of purely rotational egomotion, the part of the independent motion detection method which deals with stereo information is redundant. Independent motion detection may be achieved by just applying the LMedS estimation on the motion normal flow field, whatever the scene structure may be. However, in practical applications, the case of pure rotational egomotion is very rare.

5.6 Implementation and performance issues

Figure 5.3 summarizes schematically the method for independent motion detection through depth layering. At time t, a pair of images L_t and R_t is acquired by the stereo configuration.

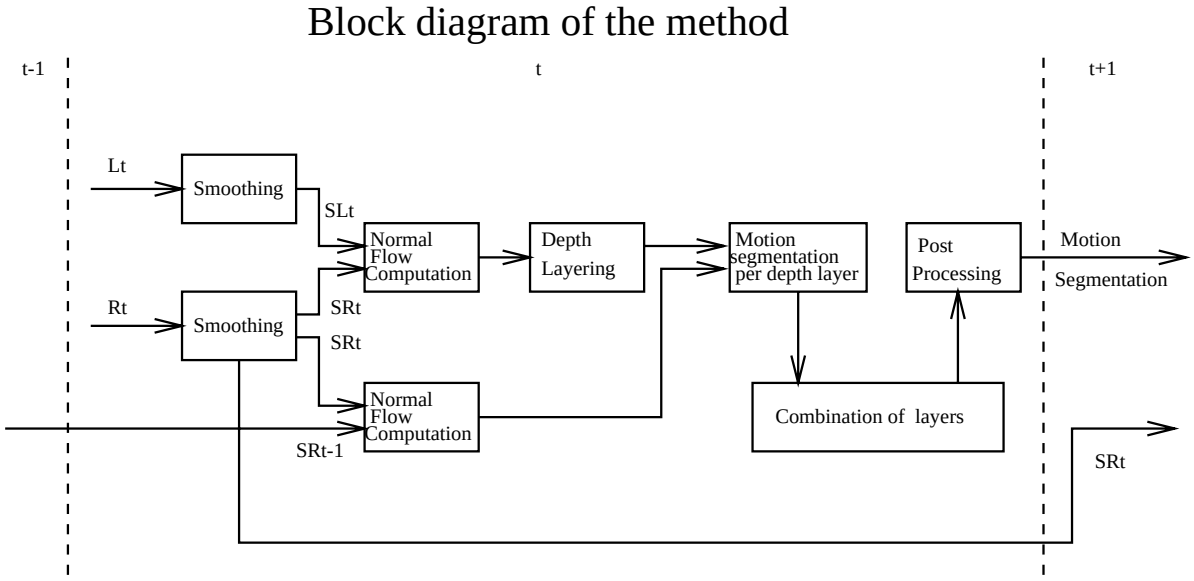


Figure 5.3: A schematic presentation of the method for independent motion detection based on robust regression in depth layers.

Much alike the independent motion detection method based on depth elimination, images are smoothed, the normal flow fields due to stereo and motion are computed and normal flow values are thresholded; all points for which the magnitude of intensity gradient is lower than a threshold, are removed from further consideration. Stereo normal flow is fed into one of the two alternative depth layering modules described, giving rise to a number of depth layers. For each of the depth layers already defined, motion segmentation is now performed based on LMedS estimation. The resulting motion segments are compared based on a least squares estimation of their motion parameters, and merged if the comparison reveals common 3D motion parameters. As already discussed, ambiguities may arise but in restricted situations. The final 3D motion segmentation is post-processed towards enhancing its usability.

Depth layering is achieved either through robust regression or by direct depth comparison. In the first case, the computational complexity of depth layering is determined by the computational requirements of LMedS. In the case that a fixating stereo configuration is used, the number p of the parameters to be estimated is equal to 3. In the case of parallel cameras there is only one parameter to be estimated. In practice, direct depth layering is performed with an iterative scheme. First, a histogram of the function $f(Z)$ is computed. The highest peak of the histogram is determined and the value of the function at this peak becomes the center for the definition of a depth layer. All points which, according to criterion (5.4), belong to this layer are excluded from subsequent consideration. These steps are repeated until all points of the image are assigned to depth layers.

Motion segmentation in each layer has a complexity that is determined by the computational requirements of LMedS. More specifically, motion segmentation involves l applications of LMedS (one for each layer), each of which involves as many observations as the number of points in the current layer. In the general case, the estimation is applied for the six motion parameters of the unrestricted 3D motion, so $p = 6$ (see Fig. 4.5(a) for a 3D plot of the number of required iterations as a function of the probability Q and the outliers probability e, for the case of $p = 6$).

Integration of results has the complexity of connected components labeling ($O(|V| \log |V|)$ where $|V|$ is the number of graph vertices). In practical situations this number is rather small. Finally, the post processing involves local operations for which efficient parallel algorithms can be developed.

Chapter 6

Qualitative Detection of 3D Motion Discontinuities

One of the most interesting aspects of the world is that it can be considered to be made up of patterns. A pattern is essentially an arrangement. It is characterized by the order of the elements of which it is made, rather than by the intrinsic nature of these elements

Norbert Wiener

6.1 Overview

In the two previous chapters, two independent motion detection methods were presented which were based on the indirect comparison of depth information extracted from a stereo configuration and a motion sequence. Both methods made use of robust regression as a means to perform 3D

motion segmentation. Although LMedS is a method for estimating model parameters in a set of noisy data, what was of central importance was not the estimated parameters, but the separation between inliers and outliers which led to motion segmentation. Although the computational cost of these methods is not prohibitively high, we would like to be able to use as fast methods as possible, even if this has some negative impact to the detail with which independent motion is detected. The method proposed in this chapter is towards this direction. Compared to the previous two methods, the method presented in this chapter takes a more qualitative approach, in the sense that solving for the motion or stereo configuration parameters is totally avoided. This is in fact a very important issue in the purposive theory of vision: Problems should be solved directly, by using suitable, specific representations instead of relying on general representations that are difficult to extract accurately. For the specific problem of independent motion detection, this methodological principle leads to an effort to detect independently moving objects through the processing of normal flow and, if possible, without relying on estimations of the 3D motion parameters of the moving objects which amounts to solving the structure from motion problem.

According to the method described in this chapter [19], the independent motion detection problem is reduced to a problem of pattern matching and, more specifically, to a problem of line fitting. The quantities compared are functions of depth, computed from a temporal sequence of image stereo pairs. One qualitative depth function is computed from the stereo configuration and another depth function is computed from motion. Both functions are defined and computed over local image patches. It is shown that these two functions obey a linear relationship in the case that there is only one 3D motion in the image patch under consideration. The violation of this linear relationship in an image patch is attributed to the existence of more than one 3D motions. Consequently, in the general case[1], boundaries of independently moving objects are detected.

[1]Later in this chapter it will be shown that in certain cases the whole image of an independently moving object can be identified

6.2 Method description

6.2.1 Qualitative depth information in image patches due to motion

The current method relies on the assumption that within a small image patch, the translational and rotational components of the egomotion can be accurately approximated with constant vectors. Based on this assumption, eq. (3.5) can be written as:

$$\begin{aligned} u &= \frac{u_T}{Z} + u_R \\ v &= \frac{v_T}{Z} + v_R \end{aligned} \tag{6.1}$$

where (u_T, v_T) and (u_R, v_R) are the constant translational and rotational components of motion for that image patch. This assumption, differs conceptually from the assumption of patchwise constant optical flow. The constancy of the translational and rotational components of egomotion depends only on the observer's 3D motion parameters and poses restrictions on the body of the observer. On the contrary, optical flow constancy implies constancy of depth, which is an environmental assumption. Figure 6.1 clarifies the difference with a characteristic example. Translational components (u_T, v_T) and rotational components (u_R, v_R) of motion are the same at points (x_1, y_1) and (x_2, y_2) in Figs. 6.1(a) and 6.1(b), respectively. However, due to difference in the depths Z_1 and Z_2 of these points ($Z_1 > Z_2$), the optical flow vectors (u_1, v_1) and (u_2, v_2) differ considerably.

The hypothesis for constant translational and rotational components of motion is valid in image areas which are far from the FOE. As demonstrated in Fig. 3.4.(a), in the case of pure translational motion, optical flow vectors emanate from the FOE whose coordinates are $(\frac{Uf}{W}, \frac{Vf}{W})$. In a small image patch far away from the FOE, translational optical flow vectors can be regarded as parallel. The farther the FOE from the image patch, the more accurate the approximation. The approximation becomes perfectly accurate if the W component of translational motion is zero, which is equivalent with the FOE being at infinity. Thus, if the translational component of the

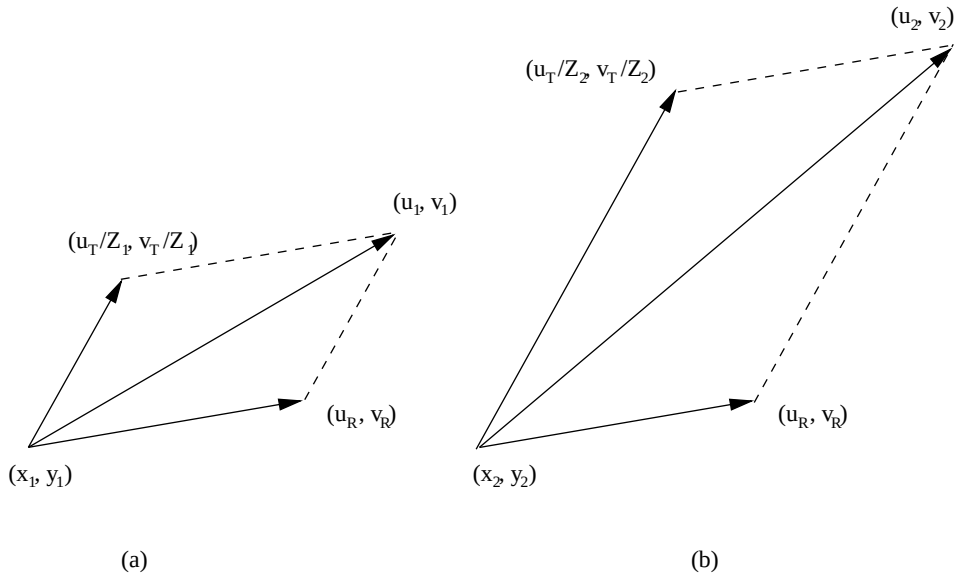


Figure 6.1: An example where the translational and rotational components of motion are constant while optical flow is not constant.

3D motion along the Z axis is very small relative to the 3D motion along the X and Y axes, the approximation holds for all patches that can be defined on the field of view.

Similar arguments hold for the case of rotational motion. As can be verified from Fig. 3.4.(c), the rotational optical flow vectors can be approximated by parallel vectors in a small patch far from the AOR[2] The hypothesis of constant rotational component is even more realistic in practical situations compared to the assumption regarding the translational component because in the case of egomotion, the rotation γ around the Z-axis is usually close to zero and, therefore, the AOR is far outside the field of view.

By substituting u and v of eq. (6.1) in eq. (3.12) we obtain:

$$u^M = \frac{(u_T n_x + v_T n_y)}{Z} + u_R n_x + v_R n_y \tag{6.2}$$

Assuming that $n_x \neq 0$ and dividing eq. (6.2) by n_x, we obtain:

$$\frac{u^M}{n_x} = \frac{(u_T + n v_T)}{Z} + u_R + n v_R$$

[2]The AOR is the point where the rotation axis intersects the image plane. The coordinates of this point are given by $(\frac{\alpha}{\gamma}, \frac{\beta}{\gamma})$.

where $n = \frac{n_y}{n_x}$, is the direction of the image gradient. Thus, for a selected normal flow direction n in a certain image patch, we can compute a depth function of the form:

$$g(Z) = \frac{K}{Z} + L \tag{6.3}$$

where

$$\begin{aligned} K &= u_T + n v_T \\ L &= u_R + n v_R \end{aligned} \tag{6.4}$$

are unknown, constant quantities. Equation (6.3) constitutes the first of the two functions, the comparison of which leads to conclusions about independent motion.

6.2.2 Qualitative depth information in image patches due to stereo

In a way similar to the case of motion, the translational and rotational components of the stereo equivalent motion can be considered constant in image patches. As can easily be shown using geometrical considerations, the translational part of the stereo equivalent motion for a fixating stereo configuration, has only a U_s and a W_s component. In all practical situations, U_s is one or two orders of magnitude greater than W_s[3]. Therefore, the FOE for the stereo equivalent motion is far from the image center. The hypothesis for a constant rotational component is also realistic. The rotational component is due to a rotation β_s around the Y-axis, which produces an almost horizontal flow field. By denoting the translational and rotational components of the stereo equivalent motion with (u_{Ts}, v_{Ts}) and (u_{Rs}, v_{Rs}), respectively, we can derive a depth function, corresponding to each patch, of the form:

$$h(Z) = \frac{A}{Z} + B, \tag{6.5}$$

[3]See Appendix A for a further analysis.

where,

$$A = u_{Ts} + nv_{Ts},$$
$$B = u_{Rs} + nv_{Rs} \tag{6.6}$$

are again unknown, constant quantities. It has also been shown [64] that eq. (6.5) holds with no approximation, for all normal flow vectors for which $xn_x + yn_y = 0$.

6.2.3 Comparison of depth functions

Suppose that for an image point p_i which corresponds to a scene point with depth Z_i, the values $h(Z_i)$ and $g(Z_i)$ of functions h and g are computed. By solving eq. (6.5) for Z and substituting in (6.3) we obtain:

$$g(Z) \quad = \quad \frac{K}{A}h(Z) + \left(L - \frac{KB}{A}\right) \tag{6.7}$$

Let $s = \frac{K}{A}$ and $t = L - \frac{KB}{A}$. Then the above equation can be written as:

$$g(Z) \quad = \quad sh(Z) + t \tag{6.8}$$

Equation (6.8) states that the functions of depth h and g, due to stereo and motion, respectively, have a linear relation for all points with the same gradient direction within an image patch. The scaling parameter s and the shift parameter t depend on the motion parameters and the stereo configuration parameters. Since the quantities A, B, K and L remain constant in an image patch, the same is true for s and t.

Consider now an image patch, in which there are points that correspond to more than one rigid 3D motions. If there are two such motions, eq. (6.3) will hold for some parameters K_1 and L_1, corresponding to points of one motion, and for some other parameters K_2 and L_2, corresponding to points of the other motion. Equation (6.8) will not hold for the same parameters s and t for all points in an image patch. The detection of such situations signals the presence of more than one rigid motions and, therefore, the presence of independent motion.

Comparison through robust regression

A method for comparing the depth functions treats the problem as a problem of line fitting. Each point in a patch, provides one equation of the form of eq. (6.8). $h(Z_i)$ and $g(Z_i)$ are computable quantities, and s and t are unknown parameters. In fact, the set of eqs. (6.8) for all points in an image patch form an overdetermined set of equations over variables s and t, that can be tackled with the LMedS robust estimator. In the presence of two rigid motions, robust regression will estimate the parameters for the majority of the points (dominant motion within that patch). Model inliers will correspond to the points of the dominant motion, while model outliers will correspond to the points of the secondary motion. The absence or the presence of outliers signals one or more 3D motions, respectively. Note that if there are two rigid motions, then the high breakdown point of LMedS suffices to handle correctly the segmentation of the scene. In case that there are more than two rigid 3D motions within a patch, and none is dominant (in the sense that 50% of the total number of points corresponds to that motion) the method will fail to estimate a correct set of parameters s and t. However, it is very likely that whatever the estimated parameters are, outliers will exist and, therefore, discontinuities will be detected.

6.3 Ambiguities in independent 3D motion detection

For the motion discontinuities detection method, there are certain cases of ambiguous 3D motions. According to the previously described method, if no set of parameters s and t can be found such that eq. (6.8) holds for all points of a certain gradient direction in a patch, then there is a 3D motion discontinuity in that patch. However, the reverse does not always hold: If eq. (6.8) holds for all image points of a certain gradient direction within a patch, then it is not guaranteed that no 3D motion discontinuity exists. Assume that in a certain tile, the depth function due to stereo is given by the relation $h(Z) = \frac{A}{Z} + B$. Suppose also that there are two different 3D motions m_1 and m_2, each of which gives a function of depth $g_1(Z) = \frac{K_1}{Z} + L_1$ and

$g_2(Z) = \frac{K_2}{Z} + L_2$, respectively. If, (see eq. (6.7)) the relations

$$\begin{aligned} s &= \frac{K_1}{A} = \frac{K_2}{A} \\ t &= L_1 - \frac{K_1 B}{A} = L_2 - \frac{K_2 B}{A} \end{aligned} \tag{6.9}$$

hold simultaneously, then although m_1 and m_2 are different motions, eq. (6.8) holds for all points of an image patch. Equations (6.9) after substitution from eqs. (6.4), yield:

$$\begin{aligned} (u_{T1} - u_{T2}) &= n(v_{T2} - v_{T1}) \\ (u_{R1} - u_{R2}) + n(v_{R1} - v_{R2}) &= \frac{B}{A}\{(u_{T1} - u_{T2}) - n(v_{T2} - v_{T1})\} \end{aligned} \tag{6.10}$$

For both equations to hold, the gradient direction and the motions m_1 and m_2 should be such that:

$$n = \frac{u_{T1} - u_{T2}}{v_{T2} - v_{T1}} = \frac{u_{R1} - u_{R2}}{v_{R2} - v_{R1}}$$

This is a rather restricted case, because it requires the selected gradient direction to have a special relation with both the translational and the rotational components of both motions. An easy way to avoid these ambiguous situations is to test if eq. (6.8) holds for more than one gradient directions.

6.4 Implementation and performance issues

Figure 6.2 summarizes schematically the method for the detection of discontinuities of 3D motion. As in the case of the previously described methods, two normal flow fields (from stereo and motion) are computed from the smoothed images SR_{t-1}, SL_{t-1} and SR_t and SL_t. The image is then partitioned into tiles. For each tile, a histogram of normal flow directions is computed. The dominant normal flow directions are determined and for these directions, the functions h and g are computed. Functions h and g are then compared by LMedS estimation.

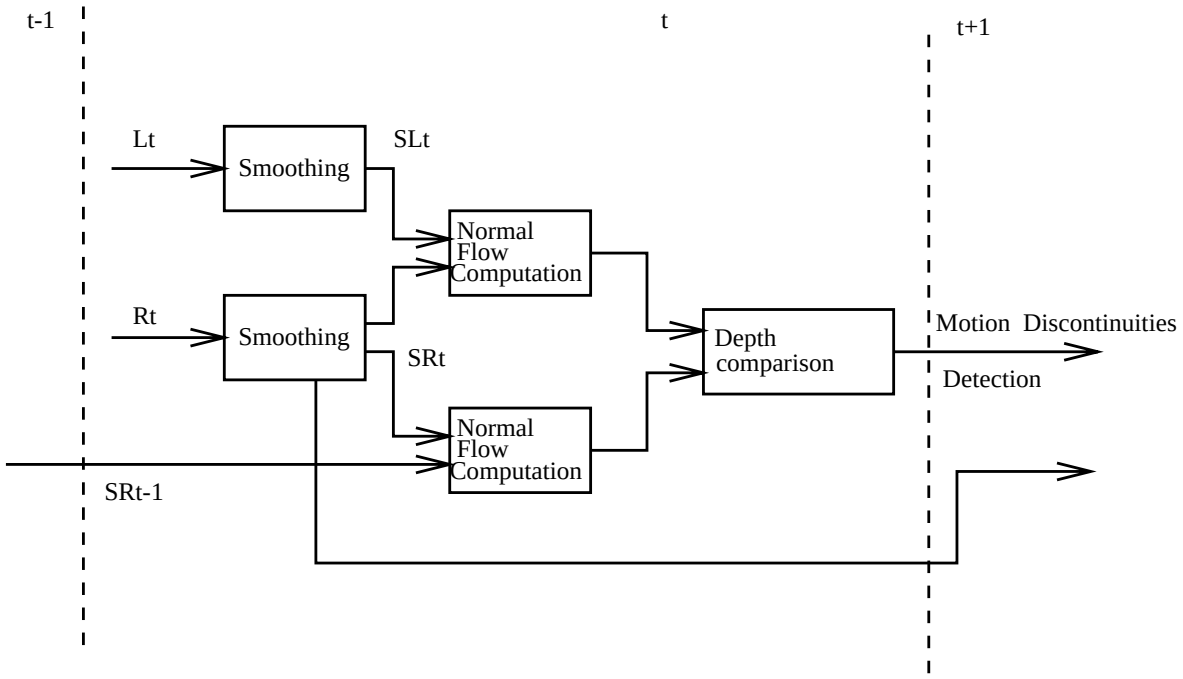


Figure 6.2: A schematic presentation of the method for 3D motion discontinuities detection.

Independent motion detection is reported in an image patch if, for at least one gradient direction, the depth functions due to motion and stereo do not have a linear relationship. As it has been demonstrated in Chapter 4, the time complexity of LMedS estimation is $O(nm)$, where n is the number of observations of the model and m the number of required iterations. In this case, n is small because robust regression is applied to a small image patch as opposed to the whole of an image. Additionally, only vectors at a specific direction are considered. The number of iterations depends on the number of parameters p to be estimated. In this case, $p = 2$. Figure 6.3 shows a 3D plot of the number of required iterations m as a function of the error probability Q (the confidence level to be reached is equal to $1 - Q$) and the outliers probability e.

Regarding the exploitation of parallel processing techniques, it should be emphasized that the processing of one image patch does not depend on the processing of another, and therefore each could be assigned to a different processor of a parallel architecture reducing considerably the execution time of this method.

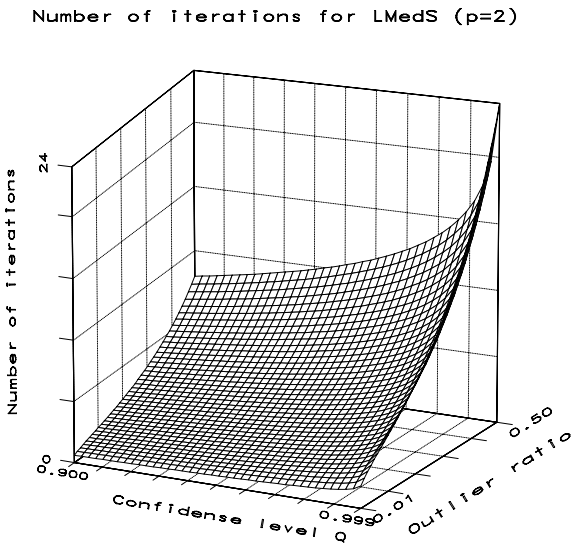


Figure 6.3: Number of iterations m for LMedS estimation (eq. (4.6)), as a function of Q and e. The number of model parameters is kept equal to p = 2.

Chapter 7

Detection of Maneuvering Objects

The unavoidable price of reliability is simplicity

C. Hoare

7.1 Overview

The methods presented in the previous three chapters detect independent motion by exploiting information acquired by a moving binocular observer. The method described in this chapter [17] relies on motion information only, i.e. it does not employ stereoscopic information. Processing is applied over the two normal flow fields that are computed from three successive images in time, in order to detect changes in the 3D motion parameters of the image points. This is important in many applications. For example in certain cases, the observer is expected to maintain a constant 3D motion, while some objects are expected to continuously change their own 3D motion parameters, i.e. they maneuver. More specifically, the proposed method decides whether the 3D motion of a certain point in the first pair of images (images acquired at time instances $t-2$ and $t-1$), remains the same in the second pair of images (images acquired at

	Change detection in intensities	Change detection in normal flow
Goal:	Detecting changes in position	Detecting changes in 3D motion
Assumption:	Constant position of the observer	Constant 3D motion of the observer
Input:	Two image frames	Two normal flow fields (from three frames)
Approach:	Difference of image frames	Difference of normal flow fields

Table 7.1: Change detection in image intensities vs. change detection in normal flow fields.

instances $t-1$ and t).

There are a number of analogies that can be drawn between change detection methods that are used to detect moving objects in the field of view of a static observer and the method for the detection of changes in 3D motion. Table 7.1 summarizes these analogies. Most of the existing methods for change detection aim at detecting changes in the position of an object in the field of view of an observer. The basic underlying assumption is that the observer is static. However, these methods are able to recognize the case of a previously static observer that starts moving, because in that case a change will be detected over the full image plane. In most cases, change detection methods work by taking (weighted) differences in the intensities of two image frames that are acquired at successive time instances.

By complete analogy, the proposed method for the detection of maneuvering objects, aims at detecting changes in the 3D motion parameters of independently moving objects. The basic underlying assumption is that the eye of the observer moves rigidly with constant 3D motion parameters. However, the method is capable of detecting changes in the 3D motion parameters of the observer, because again in that case, a change will occur over the whole image plane.

The proposed method relies on the processing of two normal flow fields that are computed from three successive images in time. The basic problem involved is that, since the observer moves with constant but unknown 3D motion parameters, the spatial registration of the

information to be compared is unknown. In the remainder of this chapter we present a technique that makes this comparison possible, without resorting to the solution of the correspondence problem.

7.2 Method description

Suppose that the motion of the observer remains constant for three frames that are acquired at time instances $t-2$, $t-1$ and t. Suppose also that we compute a normal flow field, from frame $t-1$ to frame t. According to eq. (3.13), the normal flow that can be computed at point (x, y) with gradient direction (n_x, n_y), is equal to

$$\begin{aligned} u^{M\,(t-1\rightarrow t)} &= (-n_x f)\,\frac{U}{Z} \\ &+ (-n_y f)\,\frac{V}{Z} \\ &+ (x n_x + y n_y)\,\frac{W}{Z} \\ &+ \left\{\frac{xy}{f} n_x + \left(\frac{y^2}{f} + f\right) n_y\right\}\alpha \\ &- \left\{\left(\frac{x^2}{f} + f\right) n_x + \frac{xy}{f} n_y\right\}\beta \\ &+ (y n_x - x n_y)\gamma \end{aligned}$$

where (U, V, W) are the translational motion parameters, (α, β, γ) are the rotational motion parameters and f is the focal length of the imaging system. Suppose also that we compute the normal flow from frame $t-1$ to frame $t-2$. Due to the hypothesis of constant egomotion, it turns out that the observer is moving from time $t-2$ to time $t-1$ with motion parameters (U, V, W) and (α, β, γ). Therefore, his motion from time $t-1$ to $t-2$ is described with parameters $(-U, -V, -W)$ and $(-\alpha, -\beta, -\gamma)$. According to eq. (3.13), the normal flow is equal to:

$$\begin{aligned} u^{M\,(t-1\rightarrow t-2)} &= (-n_x f)\left(\frac{-U}{Z}\right) \\ &+ (-n_y f)\left(\frac{-V}{Z}\right) \\ &+ (x n_x + y n_y)\left(\frac{-W}{Z}\right) \end{aligned}$$

$$
\begin{aligned}
&+ \quad \left\{ \frac{xy}{f} n_x + \left(\frac{y^2}{f} + f \right) n_y \right\} (-\alpha) \\
&+ \quad \left\{ -\left(\frac{x^2}{f} + f \right) n_x - \frac{xy}{f} n_y \right\} (-\beta) \\
&+ \quad (y n_x - x n_y)(-\gamma)
\end{aligned}
$$

Therefore,

$$
u^{M(t-1 \to t)} \quad = \quad -u^{M(t-1 \to t-2)} \tag{7.1}
$$

Note that (n_x, n_y) is determined by the spatial derivatives of the image intensity function. Therefore, for a given point (x, y), the gradient direction (n_x, n_y) is the same for both fields, because for both fields the reference frame for the calculation of normal flow is the same (frame $t - 1$), as shown in Fig. 7.1. Equation (7.1) provides a simple, yet effective criterion to check

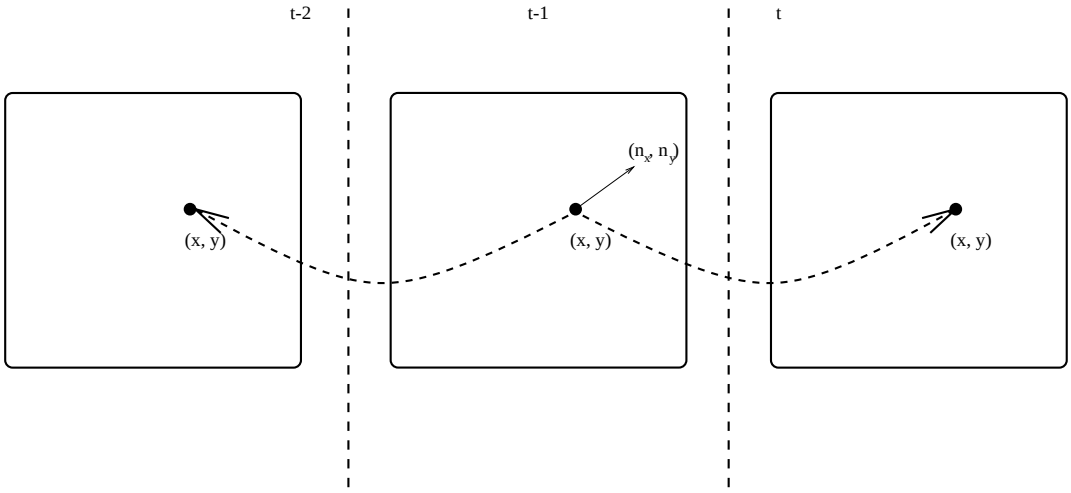


Figure 7.1: A schematic presentation of the method for the detection of maneuvering objects. Comparison of normal flow fields is enabled without solving the correspondence problem, because of the time-reversed computation of normal flow between frames $t - 2$ and $t - 1$.

whether the 3D motion parameters of a point remain constant over three frames in time. Once the two normal flow fields are computed, then for each point the sum of the normal flow values should be equal to zero. A non-zero value signals a change in the 3D motion parameters of the corresponding point. In practical situations, the sum of normal flow values will not be zero due

to errors in the computation of the time derivative. We may however require the absolute value of the sum to be small with respect to the maximum of the absolute normal flow values:

$$\frac{|u^{M(t-1\rightarrow t-2)} + u^{M(t-1\rightarrow t)}|}{max\{|u^{M(t-1\rightarrow t-2)}|, |u^{M(t-1\rightarrow t)}|\}} < \delta_{un} \tag{7.2}$$

where δ_{un} is a threshold controlling the sensitivity to changes in motion in the three frames.

The satisfaction of criterion (7.2) over subsets of scene points, leads to four interesting cases; they are summarized below, where it is assumed that the majority of the scene points correspond to the static world. Let I_P be the set of image points for which reliable normal flow values have been computed; then:

1. **The criterion holds for all image points in I_P.** This is the case where neither the observer, nor any object(s) changed their motion parameters. Note that the change of motion parameters includes the interesting case of previously static objects that have now started moving.

2. **The criterion holds for the majority of image points in I_P.** This is the case where the motion of the observer remained constant. Points where the criterion does not hold, are points of objects that changed their motion.

3. **The criterion holds for the minority of image points in I_P.** This is a special case where both the observer and the independently moving object(s) changed their motion in exactly the same way, so that no relative change can be detected.

4. **The criterion does not hold for any point in I_P.** The motion of the observer has changed. It cannot be decided, however, whether the motion of some objects has also changed.

A label may be assigned to each point with a reliable normal flow value, based on whether criterion (7.2) is satisfied. This label describes whether the 3D motion parameters of this point have changed or not. Thus, despite its simplicity, the method can provide useful information about the motion characteristics of both the observer and its environment. Additionally, as it is

shown in Chapter 9, the provided information can be effectively coupled with other independent motion detection methods.

It is noted that by employing normal flows, only incomplete information about motion is used. A normal flow value is the projection of an optical flow vector at a certain direction. Infinite many other optical flow vectors may have the same projection at this direction. Consequently, there are certain changes in the 3D motion parameters of a point that cannot be recovered through summations of normal flow values. However, in a region where 3D motion changed, it is expected that many different gradient directions exist and, therefore, the concentration of points that do not satisfy criterion (7.2) will be high. This observation leads to the conclusion that some type of post processing is needed on the resulting labeling of points. The type of postprocessing applied is the same as it was in the case of the previous methods for independent motion detection. Again, isolated points are removed. The labels of the resulting maps are repeated in small neighborhoods in order to come up with connected regions containing the maneuvering objects.

7.3 Implementation and performance issues

Figure 7.2 summarizes schematically the method for the detection of maneuvering objects. The computational requirements of the method for the detection of maneuvering objects are extremely low. Practically, the most computationally intensive part of the method is the computation of the two normal flow fields (see Chapter 4 for an analysis of the complexity of normal flow computation). The normal flow comparison is equivalent to one scan of the two normal flow fields in order to test for the satisfaction of criterion (7.2). The extremely low requirements of this method can be exploited in two ways. First, it can be decided very quickly if there are maneuvering objects in front of a moving observer and if the observer has changed his motion parameters. Second, the outcome of the method can be used to control the activation of other independent motion detection algorithms. This is actually the approach taken in Chapter 9,

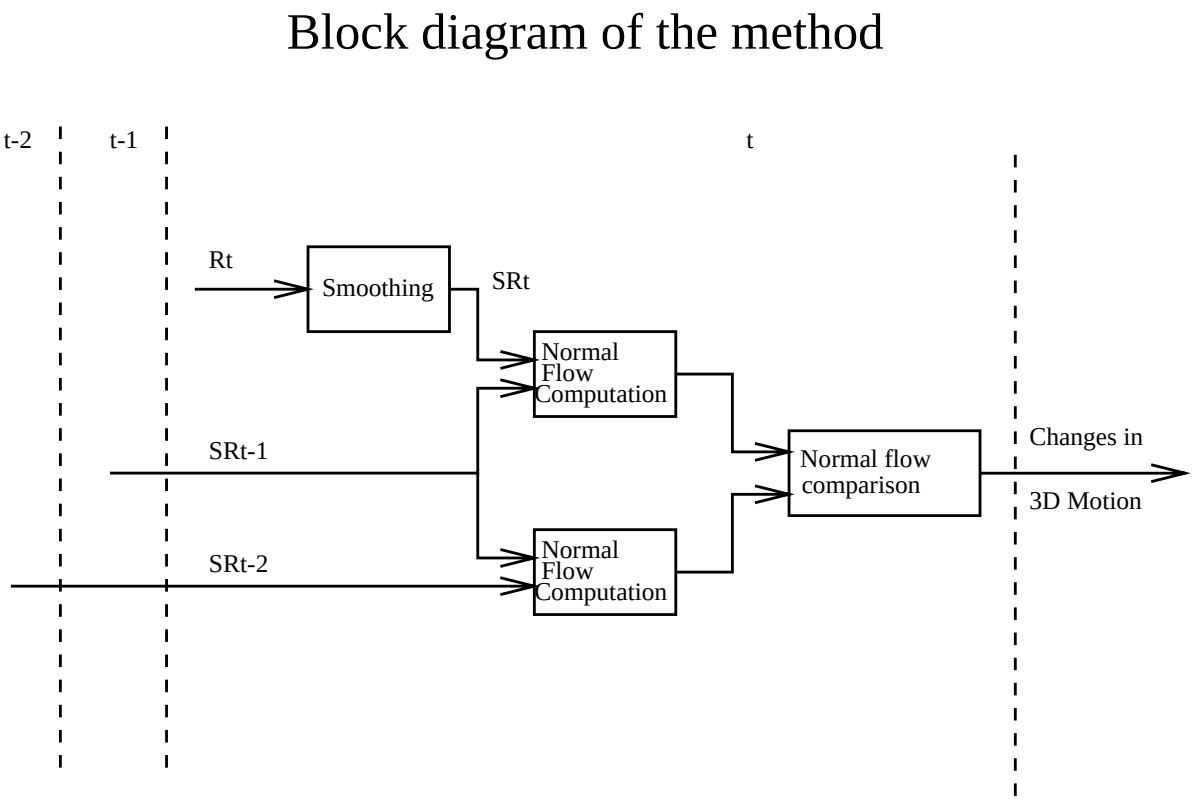


Figure 7.2: A schematic presentation of the method for the detection of maneuvering objects.

where all the proposed methods for independent motion detection are integrated in one unified framework.

Chapter 8

Experimental Results

Well done is better than well said

Anonymous

8.1 Overview

In this chapter, a number of performance issues regarding the proposed independent motion detection methods are studied experimentally. The experimental evaluation of these methods has been based on both simulated data and real image sequences. For notational convenience, we will refer to the proposed methods with the following abbreviations:

- **IMDE:** Independent Motion detection based on Depth Elimination (chapter 4).
- **IMDL:** Independent Motion detection through Depth Layering (chapter 5).
- **MDD:** 3D Motion Discontinuities Detection (chapter 6).
- **DMO:** Detection of Maneuvering Objects (chapter 7).

8.2 Experiments with simulated normal flow fields

8.2.1 Simulation environment

In order to facilitate the experimental evaluation of the proposed methods a simulation environment has been built. This environment enables the creation of synthetic normal flow fields (both stereo and motion) for a certain scene. The characteristics of the observer and of the hypothesized scene are controlled by a series of parameters:

- Image dimensions.
- The coordinates of the point where the optical axis of the hypothesized camera intersects the photosensitive surface (usually close to the image center).
- The focal length of the cameras.
- The parameters of the stereo configuration. These are actually the parameters of the hypothetical 3D motion that transforms the position of the left camera to the position of the right camera.
- The noise level of the resulting normal flow fields. Gaussian noise is assumed, which is modeled by its mean and standard deviation.
- The number of rigidly moving regions.
- For each rigidly moving region, the following parameters are defined:
  - -- Its dimensions on the image (rectangular regions are assumed).
  - -- The mean and the variance of the Gaussian distribution that models the depths of the points in that region. For each point, its depth is a random sample of such a distribution, without making any assumption regarding surface continuity.
  - -- The density of the normal flow field in the region, which is the percentage of the points in the region for which reliable motion and stereo normal flow values are

assumed. This simulates the fact that in a scene, a portion of normal flow vectors will be rejected due to low image gradient.

-- The 3D translational motion parameters of the region relative to a coordinate system that is positioned on the observer.

-- The 3D rotational motion parameters of the region relative to the same coordinate system.

Given the above parameters, a very large variety of scenarios can be effectively simulated. The output of a simulation is one possible normal flow field that can be due to the scene structure and the motion parameters. At each image point, the simulator assumes a random gradient direction which is selected from a uniform distribution in the range $[0.0, .., 2\pi)$.

8.2.2 LMedS estimation of motion parameters in a depth layer

One issue that has been experimentally evaluated using simulated data, is the capability of LMedS to estimate the motion parameters in a scene, in combination with the depth variations in that scene. This is an important issue because of two reasons:

1. If the estimation accuracy does not depend largely on depth variations, then depth variations can be neglected and two dimensional motion models can be effectively used. If, on the contrary, depth variation influences the estimated motion parameters, this favors the use of the 3D motion models as it has been proposed in this thesis.

2. For the IMDL method, the combination of motion information through the various layers depends on the accuracy of the estimated motion parameters in each layer. Thus, studying the motion parameter estimation accuracy of LMedS as a function of the depth variations in a scene, provides experimental evidence that can be exploited for the definition of a depth layer.

The experimental setup assumes images of dimensions 256 $\times$ 256. The focal length of the camera has been set equal to 600 pixels, which roughly corresponds to a focal length of 50mm. The simulation involved only one motion (egomotion) with instantaneous 3D translational parameters $(U, V, W) = (20.0, 20.0, 40.0)$ and instantaneous 3D rotational parameters $(\alpha, \beta, \gamma) = (0.001, 0.0015, 0.0005)$. It has been assumed that reliable normal flow values have been computed in 50% of the total number of points. In order to isolate the effect of depth variations in the estimation of motion parameters, no noise has been added to the normal flow fields. Three different sets of experiments have been conducted. In each of these sets, it has been assumed that the mean of the Gaussian distribution of depth was at 3, 5.5, and 8 meters from the observer, respectively. For each of these mean depths, 11 different runs were conducted, one for each of the different standard deviations of the Gaussian distribution of depth.

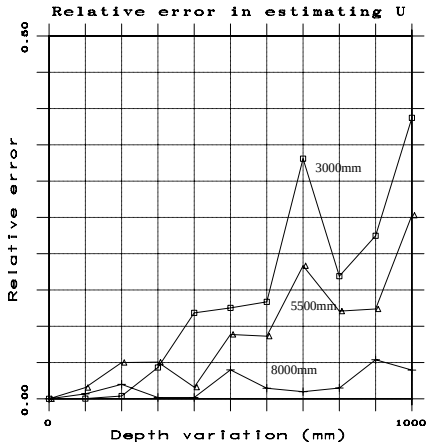
Figures 8.1(a) to 8.1(e) summarize the results of the LMedS estimation for the model of eqs. (6.3). Each figure, (Figs. (a), (b), (c), (d), (e) and (f)) corresponds to one of the motion parameters (U, V, W, α, β and γ). Note that the estimation for the three translational parameters is scaled by a depth factor. The horizontal axis corresponds to the variation of depth in a specific experiment. For example, a value of 100, means that 67% of the scene points are in the range -100 to 100 millimeters around the mean depth for that experiment. The vertical axis corresponds to the relative error E_{rel} in the computation of a parameter, namely:

$$E_{rel} = \left| \frac{p - \hat{p}}{p} \right|, \tag{8.1}$$

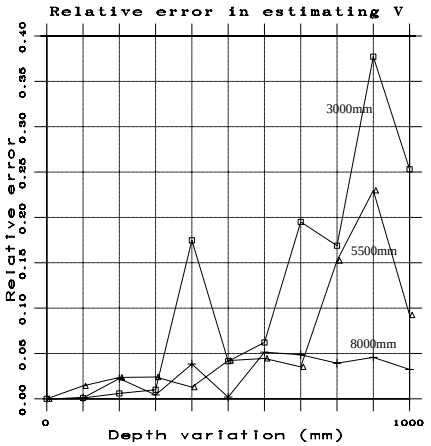
where p is the real value of this parameter and $\hat{p}$ is the estimated value for that parameter. The three curves in each plot correspond to the three different mean values for depth (3, 5.5 and 8m).

Figures 8.1(a) to 8.1(f) show that:

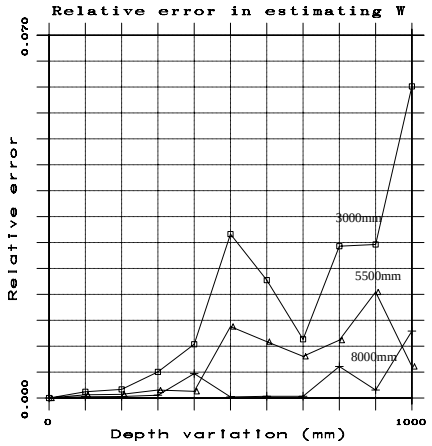
- The relative error of LMedS estimation of the motion parameters increases considerably as the depth variation in a layer increases.



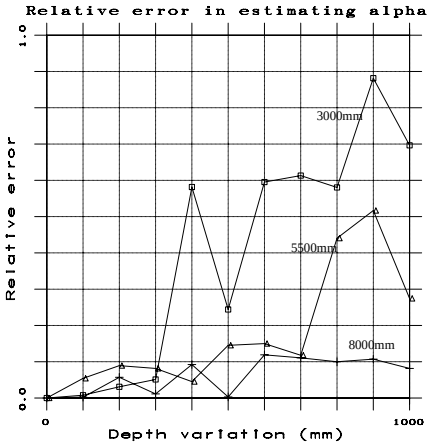
(a) U



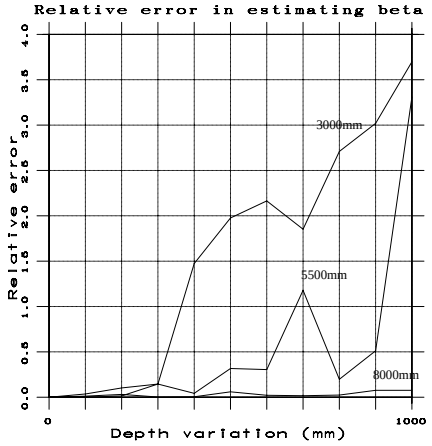
(b) V



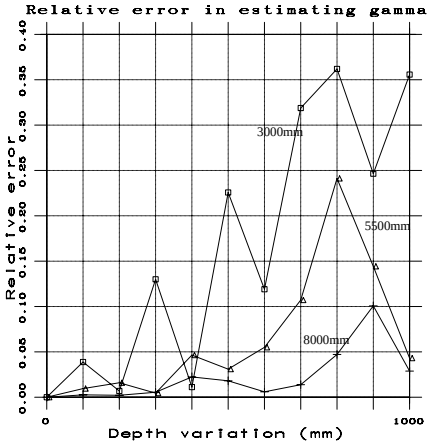
(c) W



(d) α



(e) β



(f) γ

Figure 8.1: Results of motion parameter estimation for various depths and depth variations.

- The relative error of LMedS estimation of the motion parameters increases considerably as the absolute mean depth of a layer decreases.

Regarding the first of questions raised in the beginning of this section, the above results express the need for realistic 3D motion models that exploit independent information on depth, as opposed to the 2D motion models that ignore the scene structure. This is particularly true in cases that the observer is close to the scene in view. Regarding the second of the questions raised, the experimental results favor the depth layer definition of eq. (5.1) that has already been used by the IMDL method. According to this definition, depth layers become thicker as the depth from the observer increases, because the loss of accuracy due to the increased depth variation can be balanced with the accuracy gains due to the higher absolute depth.

8.2.3 Relative performance of the proposed methods

For the purpose of comparing the performance of the proposed methods for independent motion detection, a quantitative performance measure should be defined. In the following section we define such a performance index.

Performance index

Whatever the 3D motion characteristics of a scene may be, each point of its image belongs to one of the following three disjoint sets:

- **The set** D_a, which contains the points where normal flow has been rejected as unreliable.
- **The set** E_a, which contains the points of the static background (i.e. appear to be moving due to egomotion).
- **The set** I_a, which contains the points of the independently moving objects.

The points of set D_a are not considered by the proposed independent motion detection methods because at these points there is no reliable information regarding their motion. Depending on the results of an independent motion detection method and the real motion characteristics of a scene, four different classes of points are defined:

- **Class E_e:** Points that belong to E_a and the independent motion detection method detects them as such.
- **Class I_i:** Points that belong to I_a and the independent motion detection method detects them as such.
- **Class E_i:** Points that belong to E_a but the independent motion detection method employed assigns them to I_a.
- **Class I_e:** Points that belong to I_a but the independent motion detection method employed assigns them to E_a.

Ideally, an independent motion detection method produces empty E_i and I_e sets, i.e. no misclassifications. We define the performance index Π of an independent motion detection method as

$$\Pi = \lambda \frac{\#E_e}{\#E_a} + (1 - \lambda) \frac{\#I_i}{\#I_a}, \tag{8.2}$$

where $\#A$ denotes the cardinality of the set A. The term $\frac{\#E_e}{\#E_a}$ in eq. (8.2) is the ratio of points that were correctly assigned to egomotion while the term $\frac{\#I_i}{\#I_a}$ is the ratio of points that were correctly assigned to independent motion. The factor λ takes values in the range $[0..1]$ and can be interpreted as a significance factor that, depending on its value, shifts the emphasis to the correct detection of the points belonging to I_a ($\lambda = 0$) or to the correct detection of points belonging to E_a ($\lambda = 1$). In classical detection theory [175], the factor λ depends on the a priori probability of a point to belong to a certain class. In the context of this work, λ has the additional meaning of a bias that favors the correct detection of the points that correspond to egomotion or independent motion. The choice of a specific value for λ depends on the application. In certain applications

(e.g. surveillance) it may be crucial to guarantee the detection of independently moving objects. In this case, a method with better performance index for small values of λ must be selected. In other applications (e.g. egomotion estimation) it is crucial to ensure that points identified as moving due to egomotion are really such; in these cases a method with high performance index for large values of λ is preferable. For the purposes of our evaluation, we have chosen a value of $\lambda = 0.5$.

Description of the experiments

A set of experiments have been carried out in order to evaluate the relative performance of the proposed independent motion detection methods, as a function of the noise in the motion and stereo normal flow fields. The DMO method has not been evaluated in these experiments because, as has been shown in Chapter 7, it is sensitive to motion changes rather to independent motion itself. However, the rest of the proposed methods (IMDE, IMDL, MDD) are not only compared relative to each other, but also relative to a motion segmentation method that employs a 2D motion model. This method has very close resemblances to the one proposed by Ayer et al [22], and tries to estimate the parameters of a 2D affine model by employing robust regression on a motion normal flow field. Note that none of the existing 3D motion detection methods can be directly compared to the results of the proposed methods for independent motion because all these methods rely on partial knowledge of the egomotion parameters or of the structure of the environment, which is assumed unavailable by the proposed methods.

In order to evaluate the performance of the independent motion detection methods, a synthetic motion and stereo normal flow field has been constructed by using the simulation environment. The synthetic flow fields refer to 256×256 images. A focal length of 600 pixels has been assumed for both cameras of the simulated stereoscopic observer. The cameras have been arranged in a parallel stereo configuration, with a 7cm baseline. The normal flow values in 50% of the points have been rejected, simulating the rejection of normal flows due to small

image gradient. The simulated scene contains three areas of interest. The layout of the scene[1] can be seen in Fig. 8.2. The black and the gray regions correspond to areas of the static environment,

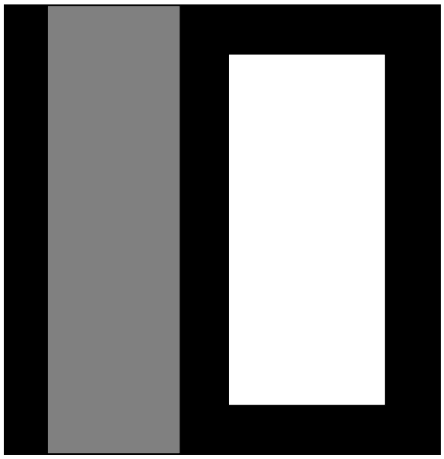


Figure 8.2: The layout of the scene used for the comparative evaluation of independent motion detection methods.

but differ in their average depth. The black area is located at approximately 6m from the observer, while the gray area is at approximately 3m from the observer. The white region corresponds to an independently moving object, which is at the same depth with the black area. Thus, an independent motion detection algorithm should produce a common label for the points of the black and gray regions (egomotion) and another label for the points of the white region (independent motion). The independently moving object covers 32% of the total area of the scene and the close-to-the-observer object covers 29% of the scene. Note also that the independently moving object covers 45% of the scene points that are at a depth around 6m. The observer has been assumed to perform a complex translational and rotational motion with parameters $(U_e, V_e, W_e) = (60.0, 60.0, 6.0)$ and $(\alpha_e, \beta_e, \gamma_e) = (0.001, 0.0, 0.0001)$, while the relative motion between the observer and the independently moving object is $(U_i, V_i, W_i) = (4.0, 40.0, 80.0)$ and $(\alpha_i, \beta_i, \gamma_i) = (0.002, 0.0002, 0.0001)$.

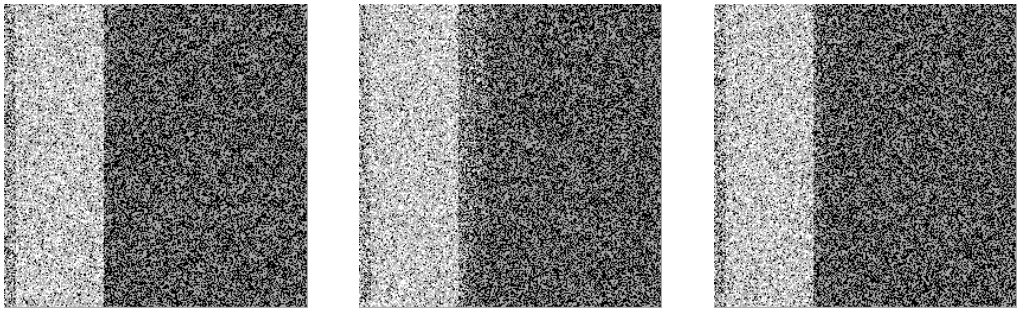
Various simulations were performed, each with different noise added to both motion and

[1]Note that the simulation does not create synthetic images, but synthetic normal flow fields. Thus, Fig. 8.2 is given to illustrate the layout of the regions in the hypothesized scene and it is not relevant to the image intensities.

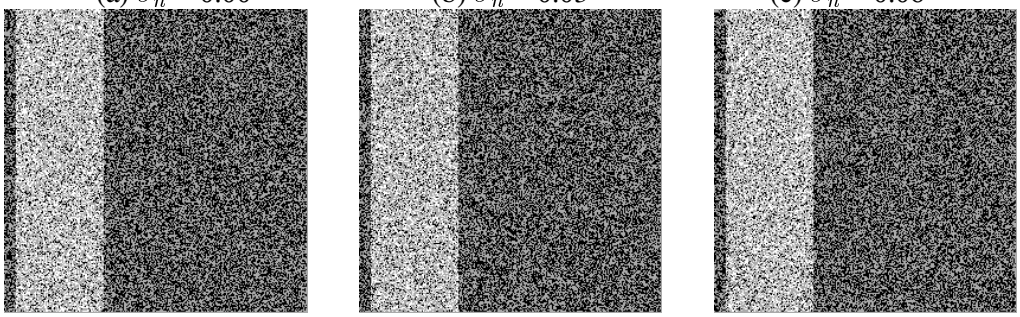
stereo normal flow fields. In all cases, Gaussian noise with zero mean has been hypothesized. The standard deviation σ_n varied in different runs from 0.0 (no noise) to 0.24. This simulates a wide variety of scenarios for noise contamination (from perfect to highly contaminated data). In each experiment, both the motion and the stereo normal flow fields were affected by the same type of Gaussian noise.

Figures 8.3(a) to 8.3(i) illustrate the results of the method for 2D motion segmentation for the noise distributions tested. Each of the images of Fig. 8.3, is a map with dimensions equal to the dimensions of the input images. Each point in this map corresponds to a point in the image and takes one of three possible values: Black, white and dark gray, corresponding to egomotion, independent motion and points with unreliable normal flows, respectively. The results of Fig. 8.3 are characteristic of the inherent weakness of the 2D motion models when applied to scenes with large depth variations. Even in the case of no noise ($(\mu_n, \sigma_n) = (0.0, 0.0)$, Fig. 8.3(a)), the method fails to capture the real 3D motion characteristics of the scene. More precisely, the method perceives the independently moving object as part of the static background, and the static object that is close to the observer (white region in Fig. 8.2) as independently moving. This is because the observed motion of the independently moving object is (in terms of the affine model) more similar to the observed motion of the distant background than it is to the static foreground object. This behavior does not change too much as a function of the noise added to the normal flow field.

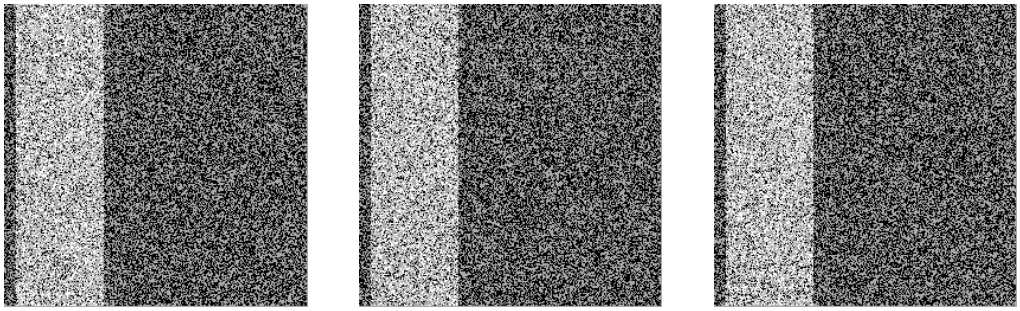
The results of the application of the IMDE method to the same data set are shown in Figs. 8.4(a) through 8.4(i). It can be seen that the correct 3D motion characteristics of the scene are captured by this method. The method recognizes the independent motion of the distant object and, at the same time, can interpret the foreground object as a part of the static background. The method fails to capture these characteristics of the scene only when the noise of the normal flow fields (for both stereo and motion) becomes very high ($\sigma_n > 0.21$). In these cases, the method is unable to clearly identify the independently moving object.



(a) $\sigma_n = 0.00$ (b) $\sigma_n = 0.03$ (c) $\sigma_n = 0.06$

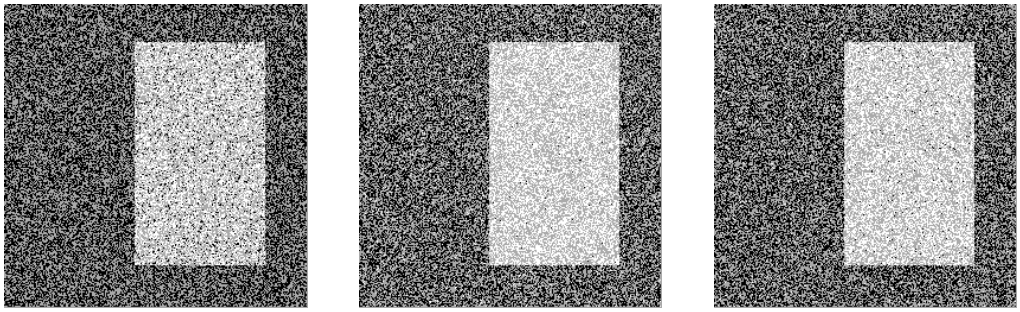


(d) $\sigma_n = 0.09$ (e) $\sigma_n = 0.12$ (f) $\sigma_n = 0.15$

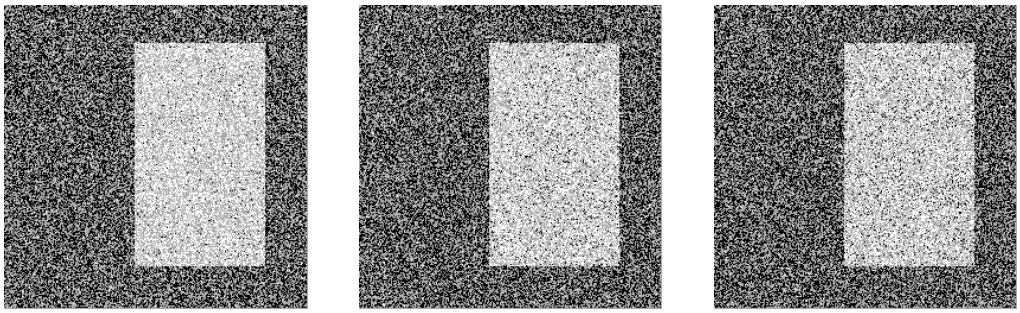


(g) $\sigma_n = 0.18$ (h) $\sigma_n = 0.21$ (i) $\sigma_n = 0.24$

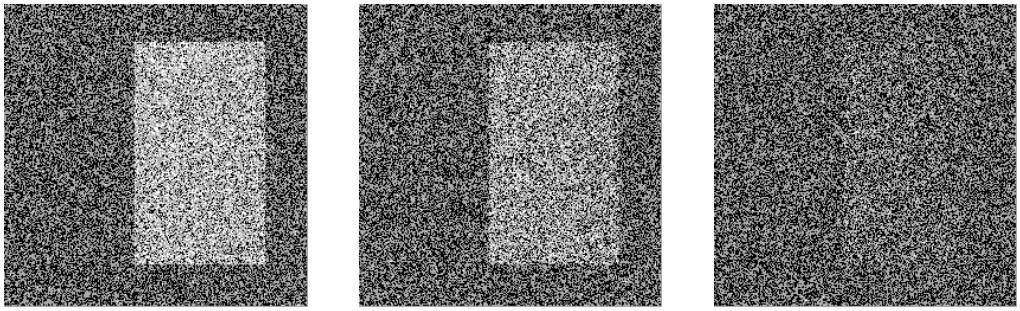
Figure 8.3: Results of 2D independent motion detection for different levels of noise.



(a) $\sigma_n = 0.00$ (b) $\sigma_n = 0.03$ (c) $\sigma_n = 0.06$



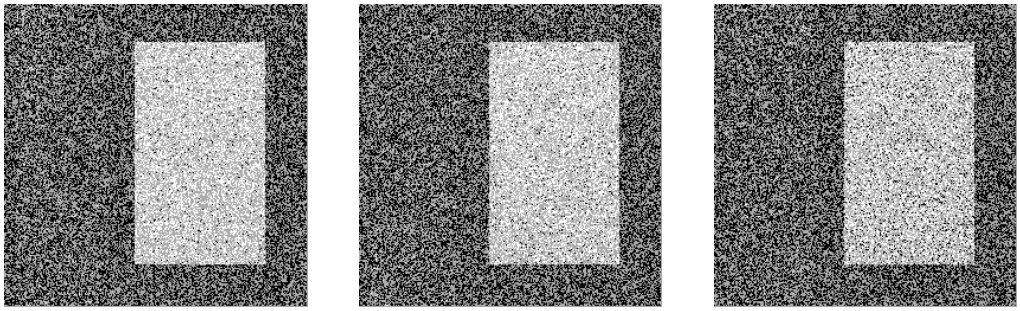
(d) $\sigma_n = 0.09$ (e) $\sigma_n = 0.12$ (f) $\sigma_n = 0.15$



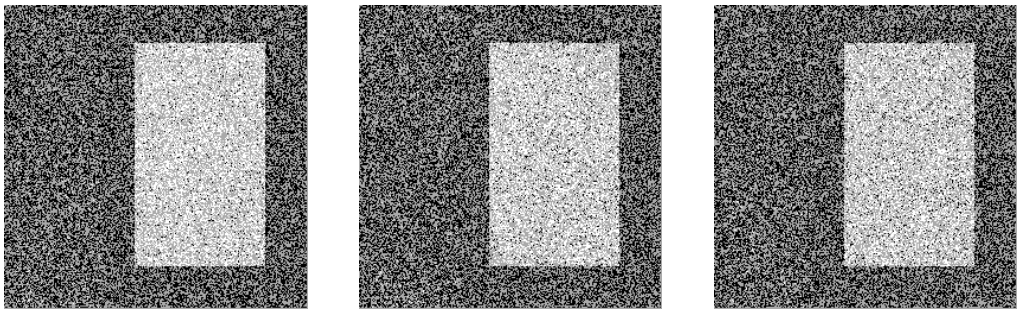
(g) $\sigma_n = 0.18$ (h) $\sigma_n = 0.21$ (i) $\sigma_n = 0.24$

Figure 8.4: Results of IMDE for different levels of noise.

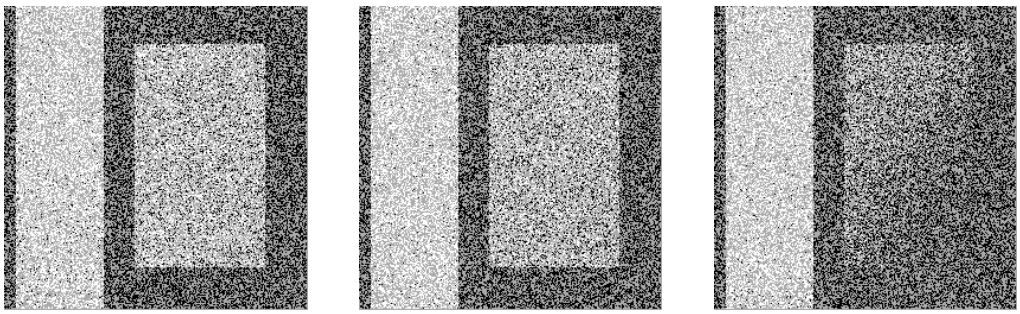
The results of the IMDL method are presented in Fig. 8.5. Again, for most of the noise



(a) $\sigma_n = 0.00$ (b) $\sigma_n = 0.03$ (c) $\sigma_n = 0.06$



(d) $\sigma_n = 0.09$ (e) $\sigma_n = 0.12$ (f) $\sigma_n = 0.15$

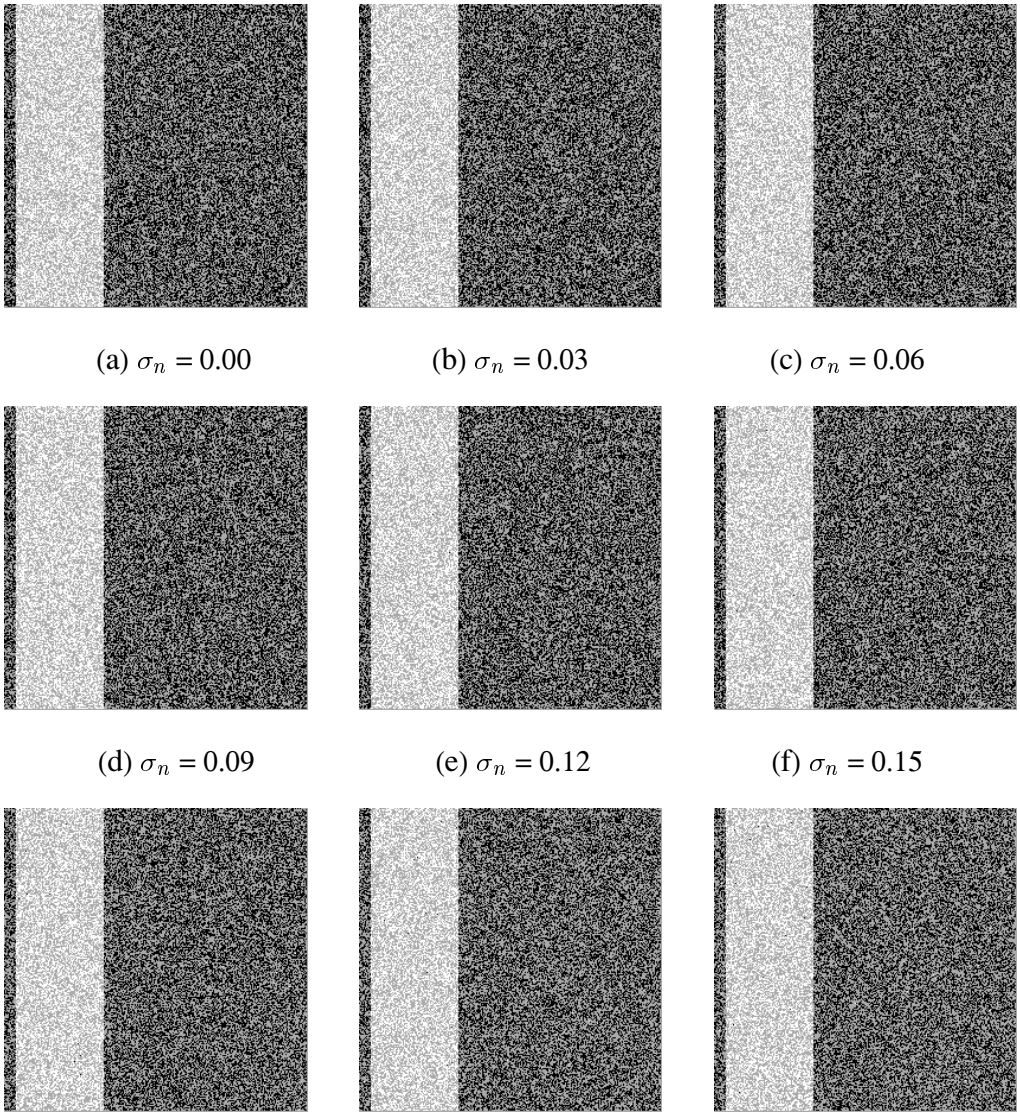


(g) $\sigma_n = 0.18$ (h) $\sigma_n = 0.21$ (i) $\sigma_n = 0.24$

Figure 8.5: Results of IMDL for different levels of noise.

levels, the IMDL method gives the correct labeling of the scene points. However, when the noise increases considerably, the performance of the method degrades. This is because at high noise levels, the estimation of the motion parameters in each layer is subject to errors and the combination of motion information across different depth layers is not robust. This happens despite the fact that the depth layers are correctly identified, and that the motion segmentation in

each depth layer is successful. This is in agreement with results reported by Ayer [22], according to which LMedS succeeds in detecting the outliers of a model even if it fails to accurately estimate its parameters. Therefore, what fails in IMDL at high levels of noise is the stage of integration of results across layers and not the processing of each depth layer. Figure 8.6 presents the results of depth layering for each experiment. It can be seen, that for all levels of noise depth



(a) $\sigma_n = 0.00$ (b) $\sigma_n = 0.03$ (c) $\sigma_n = 0.06$

(d) $\sigma_n = 0.09$ (e) $\sigma_n = 0.12$ (f) $\sigma_n = 0.15$

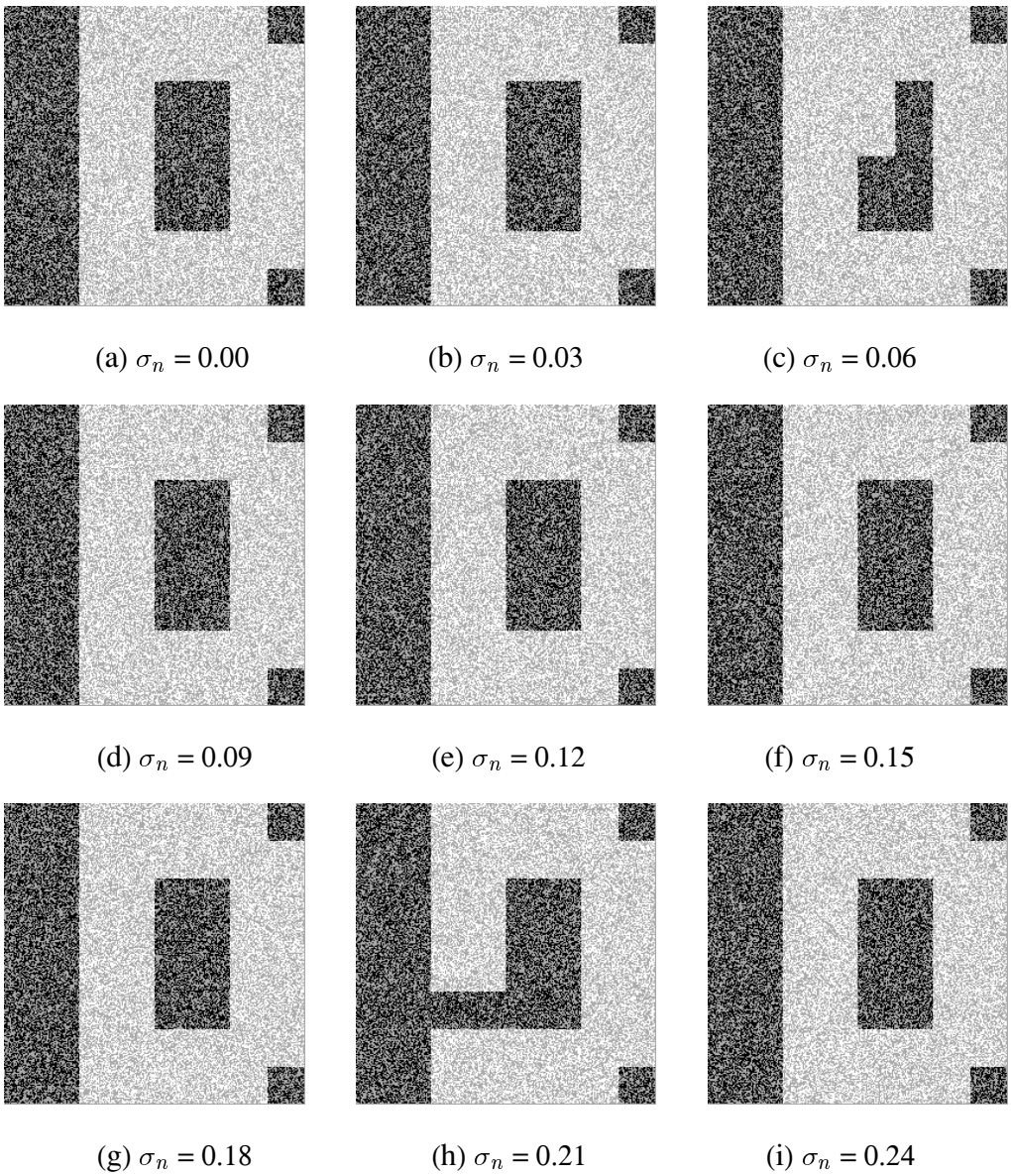
(g) $\sigma_n = 0.18$ (h) $\sigma_n = 0.21$ (i) $\sigma_n = 0.24$

Figure 8.6: Depth layers for the IMDL method, for different levels of noise.

layering is able to detect the two different layers of the scene (white color corresponds to the

closer layer and black color corresponds to the farther layer).

Finally, Fig. 8.7 shows the results of the MDD method. These results demonstrate that for



(a) $\sigma_n = 0.00$ (b) $\sigma_n = 0.03$ (c) $\sigma_n = 0.06$

(d) $\sigma_n = 0.09$ (e) $\sigma_n = 0.12$ (f) $\sigma_n = 0.15$

(g) $\sigma_n = 0.18$ (h) $\sigma_n = 0.21$ (i) $\sigma_n = 0.24$

Figure 8.7: Results of MDD for different levels of noise.

all experiments, the MDD method detects the image tiles containing 3D motion discontinuities. The method is robust in the sense that the same results are produced for most of the different noise levels. However, the result is crude in all cases. Recall that in the MDD method, the

linear relationship of the depth functions due to motion and stereo is valid only for specific directions of normal flow vectors. In this experiment the full range of directions $[0.0, .., 2\pi)$ has been divided into 36 bins. This is equivalent to assuming that two normal flow vectors have the same direction if the absolute difference in their directions is at most 10^o. Assuming a uniform distribution of the directions of normal flow vectors, a 50% covering of a region with normal flows and a minimum of 50 points for robust regression, the above figures result in a minimum of a 60×60 region for robust regression. In the conducted experiments, the size of a tile was kept equal to 64×64, which is a quite large portion of 256×256 images. This explains why the detection of motion discontinuities is so coarse in this example.

Figure 8.8 provides comparative results for the 2D, IMDE, IMDL and MDD independent motion detection methods in the form of plots of their performance index Π. The horizontal axis corresponds to the noise level in each experiment and the vertical axis plots the performance index Π. The four different curves correspond to the IMDE, IMDL and MDD and 2D methods. Each point on each curve corresponds to only one experiment. It can be observed that the IMDE and the IMDL methods have high performance index which gracefully degrades as a function of noise. The IMDE method has in general higher resistance to noise that IMDL. The performance index of MDD method is almost constant with respect to noise, as a consequence of its more qualitative properties. It should be stressed that the use of the performance index Π is not a very ''fair'' measure of the performance of the MDD method. This is because Π is, by construction, an index for evaluating independent motion detection methods rather than 3D motion discontinuities detection methods. Even under these circumstances, MDD outperforms the 2D method, while both are substantially outperformed by the IMDE and IMDL methods.

8.3 Experiments with off-line processing of image sequences

The independent motion detection methods have been also tested using image sequences that (a) have been created synthetically using the RAYSHADE [100] ray tracing tool and (b) have been

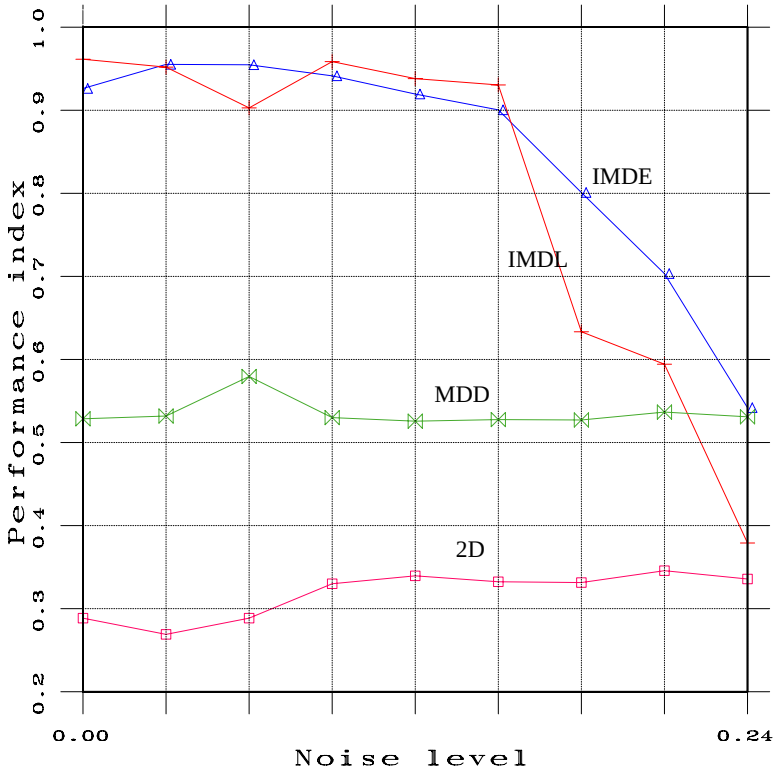
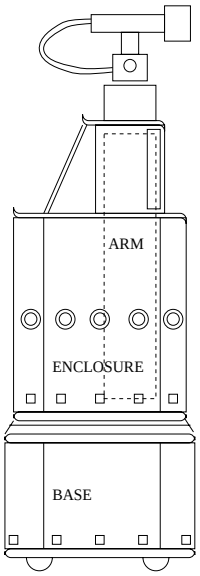


Figure 8.8: Performance index for the independent motion detection methods.

acquired by specialized equipment available at the Computer Vision and Robotics Laboratory (CVRL) of ICS-FORTH. This equipment includes:

- TALOS[2], an RWI B21 mobile robotic platform (equipped with a 486 and a PENTIUM processors running Linux, wireless communications, sonar, infrared and tactile sensors).
- A TRC BiSight active vision head (independent control of pan, tilt and vergence).

Figure 8.9(a) shows a picture of TALOS with the active vision head mounted on it and Fig. 8.9(b) shows the platform geometry. TALOS consists of two parts: the *base* and the *enclosure*. Base



(a) (b)

Figure 8.9: (a) TALOS, the multisensor mobile robot of ICS FORTH and, (b) its geometry.

contains four stepping motors, a rotation motor and four batteries that power the whole system with an autonomy of 4 hours. The enclosure is put on the base and contains the electronic parts of the robot. Besides the two main computers, the enclosure contains the controllers for the sonar, infrared and tactile sensors of the robot.

[2]According to Greek mythology, TALOS (ΤΑΛΩΣ) has been a mechanical giant constructed by the god of fire Hephaestus. TALOS has been assigned the task of patroling and guarding the island of Crete. Therefore TALOS has been the first greek robot

Several experiments have been conducted to test the proposed independent motion detection methods. It should be stressed that during the course of all the experiments the exact values of the camera focal length and image origin were unknown.

As a testbed for evaluating the performance of the proposed algorithms two image sequences have mainly been employed, namely the ''toy-car'' sequence and the ''cart'' sequence. One frame of the ''toy-car'' sequence (right image of the stereo pair at time t) is shown in Fig. 8.10. The scene captured consists of a background wall covered with paper and a toy-car together with some other objects of similar size in the foreground. The toy-car and the various objects are closer to the observer, compared to the background wall which is farther (at a larger depth). The observer (a parallel stereoscopic system) performs translational motion with a right to left direction. Apart from the toy-car, the rest of the scene is stationary. The toy-car is moving across the scene in a left to right direction, with unrestricted 3D motion.

One frame of the ''cart'' sequence (right image of the stereo pair at time t) is shown in Fig. 8.11. In this sequence of images, the observer performs a translational motion with U and W components as well as with a rotational β component. The horizontal translation is the motion that dominates. The observer is again a parallel stereoscopic system. The field of view contains a distant background and a close to the observer foreground. The background contains two independently moving objects: A cart that translates in the opposite direction of the observer (middle of the scene) and a small box (to the right of the scene) that translates at the same direction with the observer, but with different velocity. The foreground of the scene contains a table on which there is a toy car. Both objects are stationary relative to the static environment.

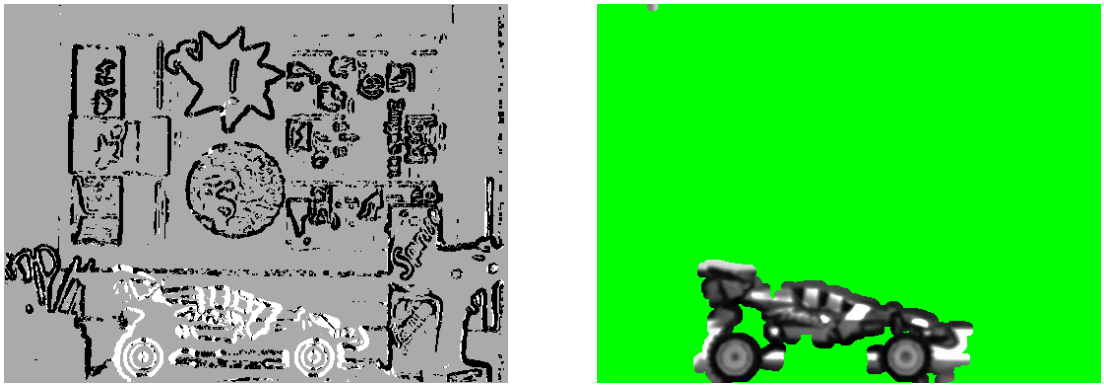
In the following sections, sample results for each of the proposed methods are presented. Results from additional image sequences are presented as the need arises.

Figure 8.10: One frame of the ‘‘toy-car’’ sequence.

Figure 8.11: One frame of the ‘‘cart’’ sequence.

8.3.1 Experiments with independent motion detection based on depth elimination

Figure 8.12 illustrates the results of motion segmentation produced by the IMDE method for the ‘‘toy-car’’ sequence. In Fig. 8.12(a), black color corresponds to egomotion and white color

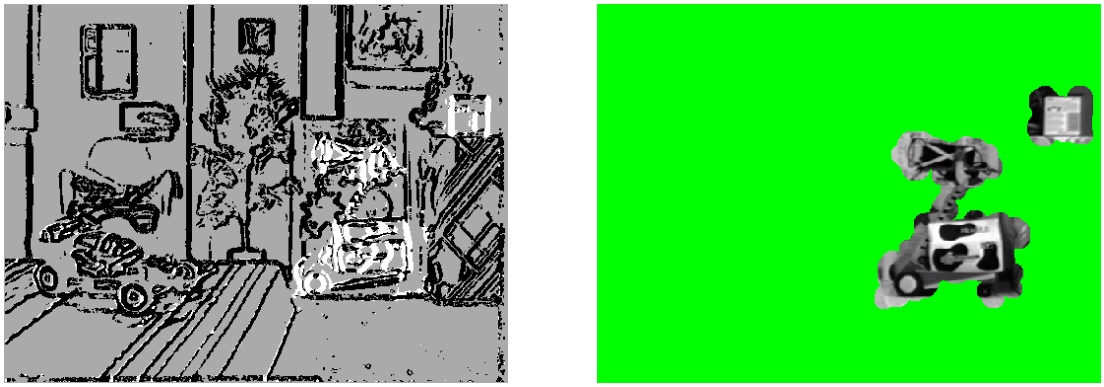


(a) (b)

Figure 8.12: Motion segmentation (IMDE method) for the ‘‘toy-car’’ sequence. (a) before and, (b) after postprocessing.

corresponds to independent motion. Gray color corresponds to points where no decision can be made due to low image gradient and, therefore, lack of normal flow vectors. It can be verified that the largest concentration of white (i.e. independently moving) points is over the regions of the independently moving points. Figure 8.12(b) presents the results of Fig. 8.12(a) after postprocessing, which eliminates isolated outliers (inliers) in large populations of inliers (outliers) and, in the resulting map, dilates the labels of independently moving points. In Fig. 8.12(b) areas that are detected as independently moving appear with the intensities that they have in the original image, while all other areas are masked out. It can be seen that after this type of postprocessing the body of the car has been successfully identified as independently moving by the IMDE method.

Figure 8.13 illustrates the results of motion segmentation of the ‘‘cart’’ sequence by the IMDE method. It can be seen that the points that are not identified as independently moving, although they belong to an independent motion, are those belonging in horizontal edges. This is



(a) (b)

Figure 8.13: Motion segmentation (IMDE method) for the ‘‘cart’’ sequence (a) before and, (b) after postprocessing.

because the model of eq. (4.5) does not hold for $n_x = 0$, which is the case of vertical gradients or, equivalently, horizontal edges ($(n_x, n_y) = (0, 1)$). Figure 8.13(b) presents the results of Fig. 8.13(a) after postprocessing which eliminates isolated outliers (inliers) in large populations of inliers (outliers).

8.3.2 Experiments with independent motion detection through depth layering

A first result refers to synthetically generated images. Figure 8.14 shows one frame of the ‘‘buildings’’ sequence. The composed scene contains 4 artificial ‘‘buildings’’ on a checkered ground. All buildings have the same physical dimensions. The leftmost and rightmost buildings are at the same depth from the observer. The left-middle building is at a larger depth from the observer (compared to the depths of the leftmost and rightmost buildings); the right-middle building is at an even larger depth from the observer. The observer performs a translational motion along the Z axis approaching the scene in view. At the same time, the two buildings in the right half of the image perform independent motions on their own. The rightmost building performs an independent translational motion along the Y axis and the right-middle building performs a composite translational and rotational motion. Figures 8.14(a),(b),(c) show the results

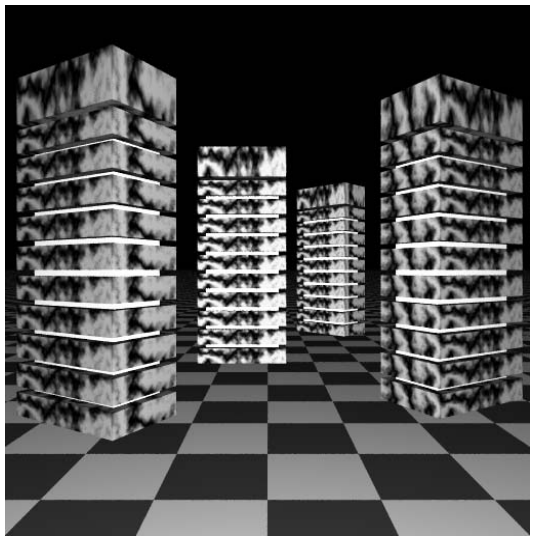
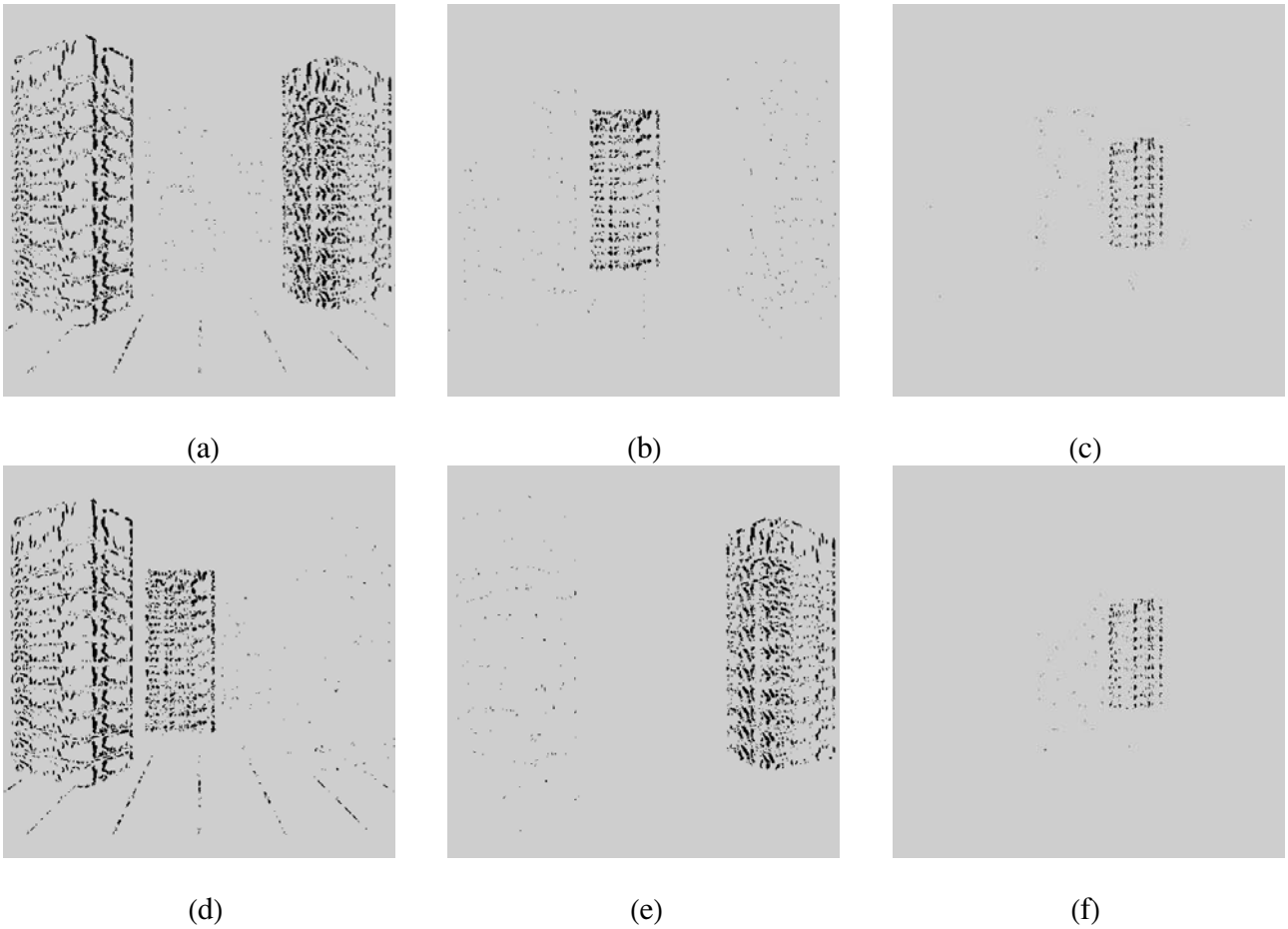


Figure 8.14: One frame from the ‘‘buildings’’ synthetic, stereoscopic sequence.

of depth layering. As can be verified from this figure, the three different depth layers have been successfully detected and outlined. The points corresponding to each layer have a black color. The first layer corresponds to the two closer objects (leftmost and rightmost buildings), the second to the object in intermediate depth (left-middle building) and the third to the object furthest from the observer (right-middle building). Figures 8.14(d),(e),(f) show the results of 3D motion segmentation. Two independent motions have been revealed (Figs. 8.14(e),(f)). Egomotion is shown in Fig. 8.14(d). It can be observed that successful discrimination of all different 3D motions in the scene has been accomplished, although they appear at different depths. Moreover, the 3D motion of the middle-left building has been successfully characterized as being identical to that of the leftmost building.

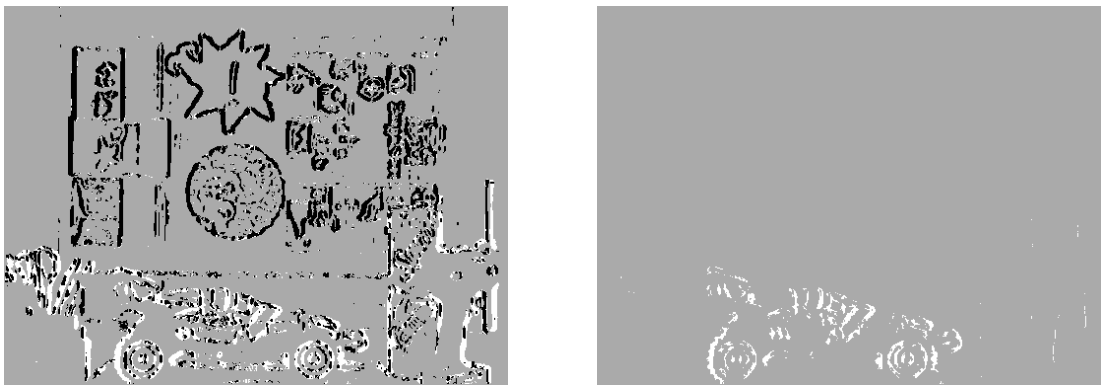
A second result refers to the ‘‘toy-car’’ sequence (Fig. 8.10). The process of depth layering resulted in two depth layers, that are shown in Fig. 8.16(a). Black color corresponds to points of the distant layer and white color corresponds to the layer close to the observer. The outliers of motion segmentation in the second layer appear in Fig. 8.16(b). These outliers correspond to the motion of the toy-car. The 3D motion segmentation of the whole scene is the result of combination of motion information through the various depth layers. The motion segmentation



(a) (b) (c)

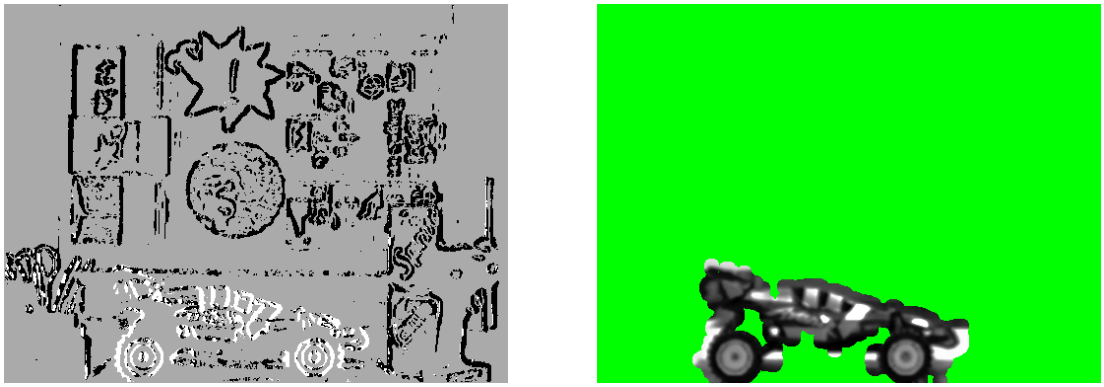
(d) (e) (f)

Figure 8.15: IMDL results for the ‘‘buildings’’ sequence (a), (b), (c) depth layers, (d), (e), (f), motion segmentation.



(a) (b)

Figure 8.16: (a) Depth layers detected by IMDL for the ‘‘toy-car’’ sequence and (b) the outliers of motion segmentation in the second (closer to the observer) depth layer.

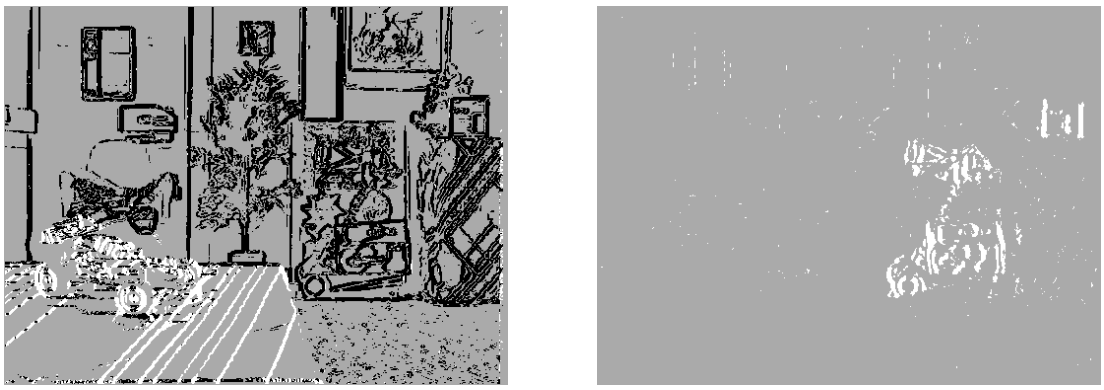


(a) (b)

Figure 8.17: IMDL motion segmentation (a) before and, (b) after postprocessing for the ‘‘toy-car’’ sequence.

is shown in Figs. 8.17(a) and (b), before and after postprocessing, respectively. Again, it can be observed that correct discrimination of the two different depth layers, as well as of the independent motion of the toy-car, has been achieved.

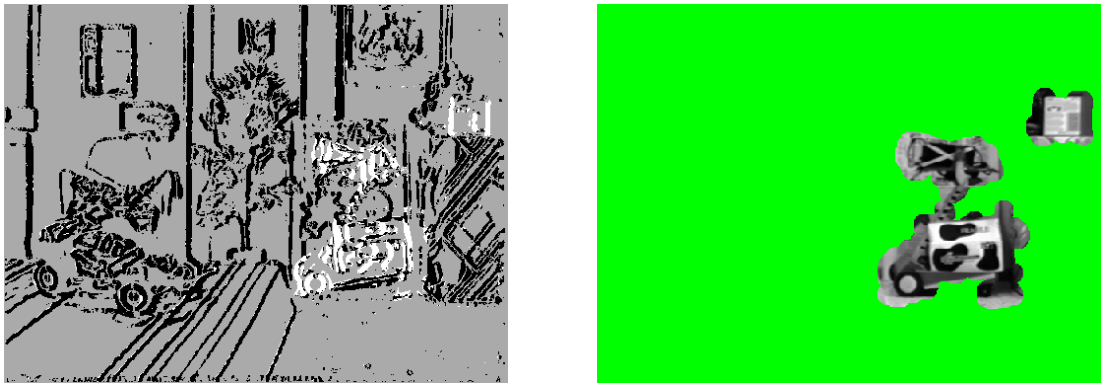
IMDL has also been applied to the data set of the ‘‘cart’’ sequence (Fig. 8.11). The method defined two depth layers, that are shown in Fig. 8.18(a). Robust regression within each of



(a) (b)

Figure 8.18: (a) Depth layers detected by IMDL for the ‘‘cart’’ sequence and (b) the outliers of motion segmentation within the first (distant) depth layer.

these layers gives rise to an estimation of the parameters of the dominant motion and to a set



(a) (b)

Figure 8.19: IMDL motion segmentation (a) before and, (b) after postprocessing for the ‘‘cart’’ sequence.

of outliers. The combination of maps that are acquired in each depth layer give the integrated 3D motion segmentation map that is shown in Fig. 8.19. As it can be verified from Fig. 8.19, the IMDL method is capable of detecting the independent motion that is present in the scene, without misinterpreting the observed motion due to depth variation as independent motion.

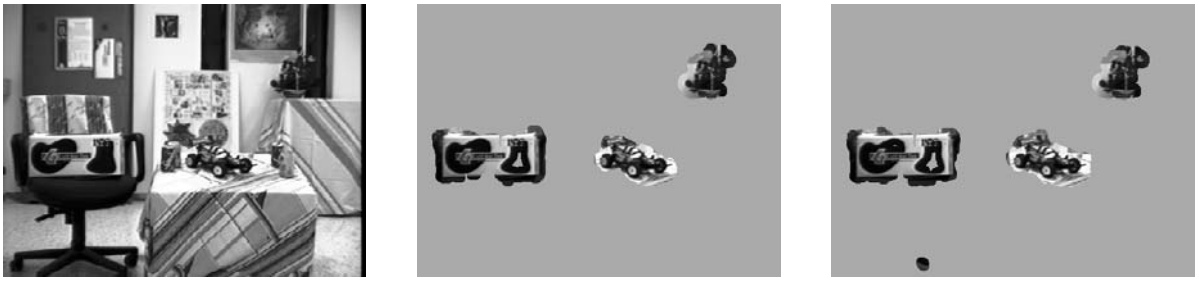
Figures 8.20, 8.21 and 8.22 provide additional results from applying the IMDE and IMDL methods in three more scenes.

(a) (b) (c)

Figure 8.20: (a) One frame of the ‘‘moving-car’’ image sequence: Egomotion has translational (U) and rotational (γ) components and the toy car is independently moving (U-, W-translation). (b) IMDE results and, (c) IMDL results.

(a) (b) (c)

Figure 8.21: (a) One frame of the ‘‘moving tableau’’ image sequence: Egomotion has translational (U) and rotational (α) components and the tableau is independently moving (V-translation). (b) IMDE results and, (c) IMDL results.



(a) (b) (c)

Figure 8.22: (a) One frame of the ‘‘multiple motions’’ sequence: Egomotion has translational (V, U) and rotational (β) components. Three objects are independently moving, the tableau (V-translation), the toy-car (U-, W-translation) and the flowers (U-translation). (b) IMDE results and, (c) IMDL results.

8.3.3 Experiments with motion discontinuities detection

The MDD method has been applied to the ‘‘toy-car’’ image sequence. The 3D motion segmentation results are presented in Fig. 8.23. The results demonstrate that correct discrimination of the independent motion of the toy-car has been achieved. The results of the motion segmentation

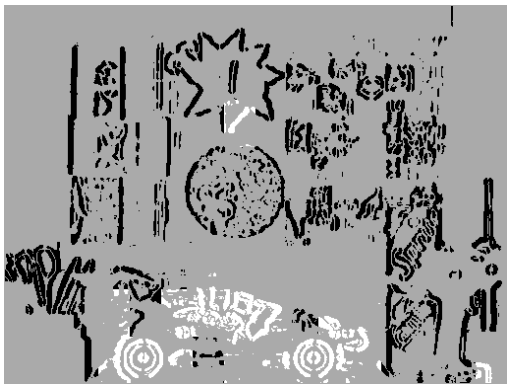


Figure 8.23: Motion discontinuities detection for the ‘‘toy-car’’ image sequence.

Figure 8.24: Motion discontinuities detection for the ‘‘cart’’ image sequence.

of the ‘‘cart’’ sequence are shown in Fig. 8.24. It can be seen that the method has correctly identified most of the tiles where more than one 3D motions are present. Although the results are complete for the case of the moving box, the method fails to detect the elongated (upper) part of the cart because there are not enough normal flows (in terms of the requirements of the method) to support the existence of independent motion.

8.3.4 Experiments with independent motion detection based on a 2D method

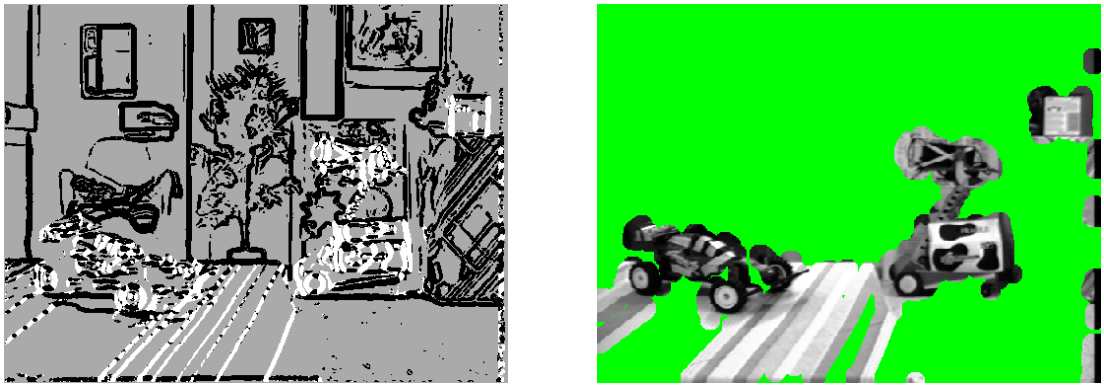
At this point, we present the results of the 2D method when applied to the ‘‘toy-car’’ and ‘‘cart’’ sequences. Figure 8.25 illustrates the results of motion segmentation produced by the 2D method for the ‘‘toy-car’’ sequence. It can be seen that the 2D method detects the independent motion

(a) (b)

Figure 8.25: Motion segmentation (2D method) for the ‘‘toy-car’’ sequence, (a) before and, (b) after postprocessing.

of the toy-car, but also detects all the static points of the image foreground as independently moving. This is because, due to the depth variation, the observed motion of the static objects of the foreground ‘‘outlies’’ compared to the dominant motion of the distant background.

Figure 8.26 illustrates the results of motion segmentation produced by the 2D method for the ‘‘cart’’ sequence. The method is able to detect the independent motion that is present in the distant background. However, it also detects the toy car and the table as independently moving, although they belong to the static background. This is because the 2D method does not exploit any information regarding depth. It actually assumes that all points in the scene are at an equal distance from the observer. Thus, the observed motion of the toy-car which is due to its relatively small distance from the observer, is erroneously interpreted as being independent.



(a) (b)

Figure 8.26: Motion segmentation (2D method) for the ‘‘cart’’ sequence, (a) before and, (b) after postprocessing.

8.3.5 Experiments with detection of maneuvering objects

A set of experiments have been conducted in order to test the performance of the DMO method. Representative results from these experiments are given in this section. In a first experiment, a real image sequence (the ‘‘coca-cola’’ image sequence) has been employed. Figure 8.27 shows one frame of this sequence, at time instance t. The camera moves with translational motion

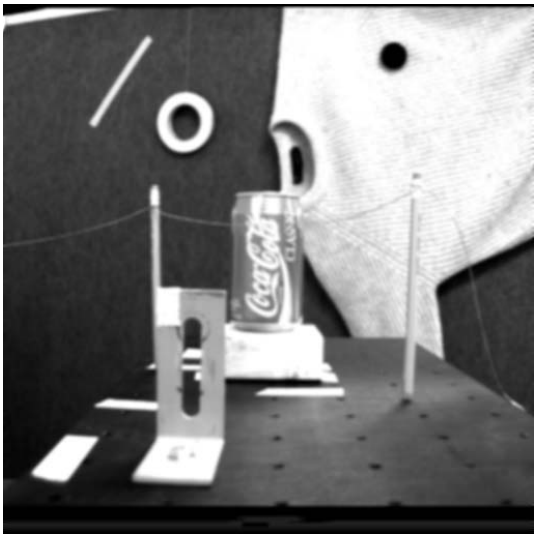


Figure 8.27: One frame of the ‘‘coca-cola’’ sequence.

approaching the scene. A rotational motion has been added in the area of the coca-cola can in order to simulate a static object that started moving in the field of view of the observer. More specifically, in the third frame (frame at time t), the coca-cola can has been moved relative to the second frame by synthetically adding rotational motion. Rotational motion has been employed because it does not depend on the (unknown) scene structure.

After smoothing the images, the two normal flow fields were obtained and the left part of criterion (7.2) was computed for all image points with a reliable normal flow value. Figure 8.28 shows a three dimensional plot of the values of the left part of criterion (7.2). x and y dimensions

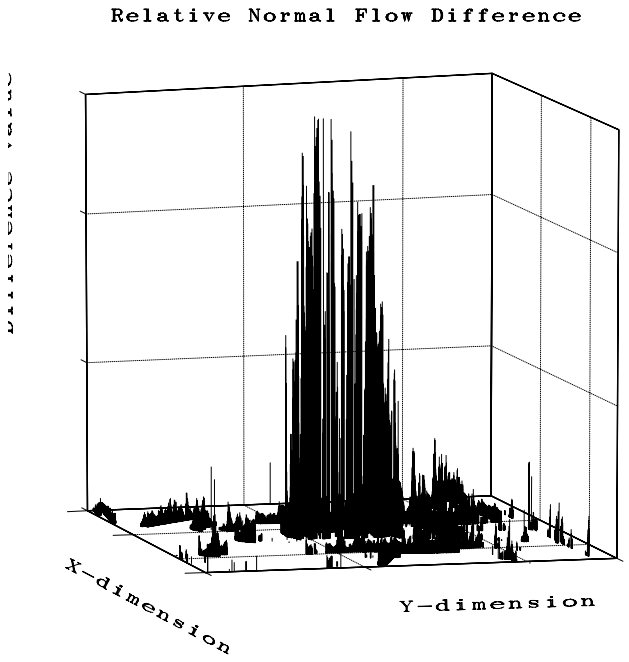


Figure 8.28: 3D plot of the image points with respect to criterion for changes in 3D motion.

of the plot correspond to the x and y dimensions of the image while the third dimension corresponds to the values of the left part of criterion (7.2). It is evident that in the points of the coca-cola can where a motion change occurs, criterion (7.2) gives distinguishably different values compared to the rest of the image points which move due to the constant egomotion. Figure 8.29 shows the final labeling of the image points. White pixels correspond to points where the 3D motion has changed, black pixels correspond to points which kept the same 3D

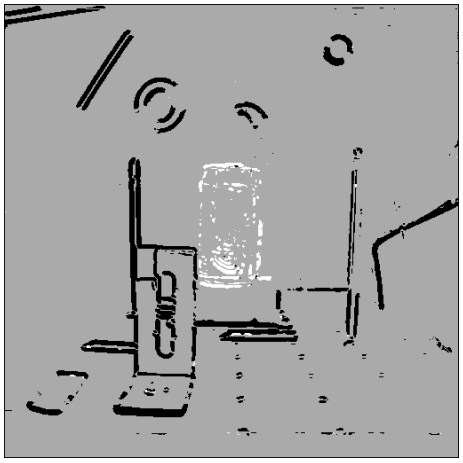


Figure 8.29: Characterization of points with respect to the constancy of their 3D motion parameters (see text for explanation).

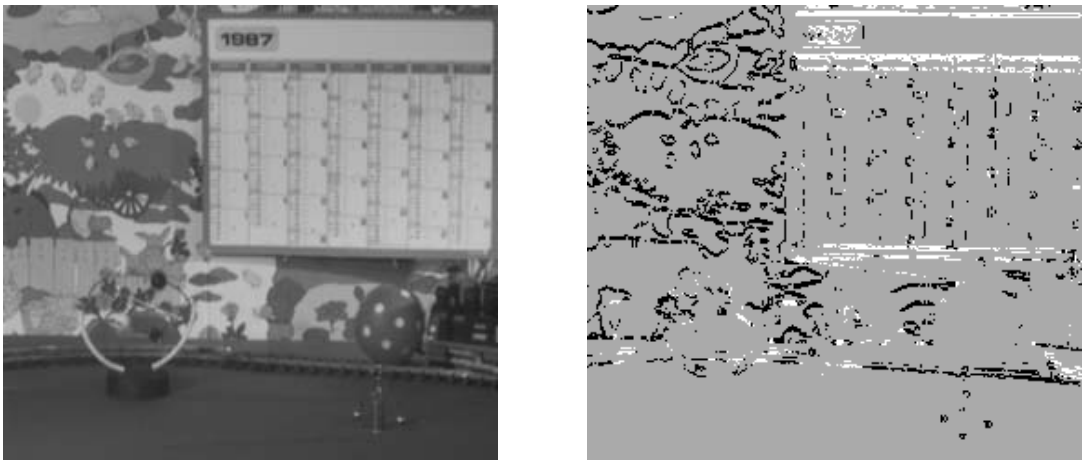
motion parameters and gray pixels correspond to points where no reliable normal flow vectors could be computed.

In a second experiment, the ‘‘interview’’ sequence has been employed. In order to simulate a change in the motion parameters of the observer, four consecutive frames were selected and the third one was dropped. Thus, the frames used correspond to time instances $t-3$, $t-2$ and t. The omission of a frame is equivalent to a change in the observer's 3D motion parameters, since in the original sequence his motion is a constant one. Fig. 8.30(a) shows the frame at time instance $t-2$ (the middle of the three frames used) and the results after the application of criterion (7.1) are illustrated in Fig. 8.30(b). Assuming the same coloring scheme as in Fig. 8.29(b), all points where normal flow has been reliably computed appear in white, signaling the change of the 3D motion parameters of the observer.

A third experiment employs the ‘‘calendar’’ sequence. In this sequence, the calendar appearing on the top-right of the images is moving upwards and, subsequently, its motion is modified and oriented down and to the right with respect to the image frame. All other objects are moving with constant motion parameters. Fig. 8.31(a) shows the middle of the three frames used from this sequence and the results regarding motion changes are presented in Fig. 8.31(b).

(a) (b)

Figure 8.30: (a) A frame of the ‘‘interview’’ sequence (b) characterization of points with respect to the constancy of their 3D motion parameters (see text for explanation).



(a) (b)

Figure 8.31: (a) A frame of the ‘‘calendar’’ sequence (b) characterization of points with respect to the constancy of their 3D motion parameters (see text for explanation).

As can be verified, the points of the calendar that contribute to the normal flow field have been successfully detected as points where 3D motion parameters change.

8.4 Experiments with on-line processing of image sequences

A set of experiments has also been conducted with to test the on-line performance of the IMDL method when robust regression is used as a means for depth layering [21]. In all these experiments the on-board PENTIUM processor of TALOS has been used; due to the limited processing power and for reducing the overhead for image acquisition, processing and results dumping on disk, the image size has been kept small, namely 144×106. Moreover, the number m of iterations of LMedS for motion segmentation has been kept relatively small, namely $m = 100$. However, this did not affect the motion segmentation results since the spatial extent of the motion was rather small.

A sample result from these experiments is presented in Figs. 8.32 and 8.33. Fig. 8.32 shows twelve frames of a sequence that correspond to the left frames of the stereo pairs. These frames are not consecutive in time but show intermediate snapshots of the whole sequence. As can be observed, in the scene in view there is a man who is initially sitting on a chair (right of the scene). The man then stands up, moves to the cart, takes it to the leftmost part of the room and then returns to his initial position. Meanwhile, TALOS is moving with W and U components (i.e. approaches the scene and also moves to the left). Figure 8.33 shows the motion segmentation (after post processing) that has been achieved by independent motion detection. This figure demonstrates clearly the correct segmentation that has been achieved. It is also worth noting that the small table in the foreground (bottom-left of the scene) has not been misinterpreted as moving, although it is placed at substantially different depth from the rest of the scene.

Figure 8.32: Twelve images from the ‘‘moving man’’ sequence.

Figure 8.33: Results of independent motion detection for the ‘‘moving man’’ sequence.

Chapter 9

Integrating the Proposed Methods

The whole is greater than the sum of its parts

Max Wertheimer

9.1 Overview

Through the presentation of the proposed methods for independent motion detection and through the study of the experimental results obtained by their application in simulated and real situations, the advantages and the disadvantages of each approach have been demonstrated. In this chapter, a comparative overview of these methods is provided. Additionally, the study of this overview reveals the possibility of integrating the independent motion detection methods in a unified framework, in order to exploit the merits of each of them.

9.2 Comparative study of independent motion detection methods

Table 9.1 summarizes the characteristics of each method with respect to a number of important criteria. As Table 9.1 shows, all the proposed methods for independent motion detection assume unknown, unrestricted (both translational and rotational) rigid motion. This is a valid assumption because, even if the observer himself moves non-rigidly, all biological eyes and, especially, man-made cameras have rigid structure and move rigidly. The only constraint is put by the MDD method, for which the FOE and the AOR of the egomotion should meet some constraints. Additionally, the DMO method assumes that the egomotion of the observer is constant for three image frames. The generality of the egomotion model assumed by the proposed independent motion capabilities should be contrasted with the strict assumptions of other approaches, where the egomotion is either assumed to be known, or restricted (e.g. only translational or only rotational components).

With respect to the independent motion that can be detected, the constraints put are even looser. Independent motion is not even constrained to be rigid for IMDE, MDD and DMO. A non-rigid motion will also be detected. Points are characterized as independently moving if they do not conform to the rigid egomotion model. A point may outlie either because of an independent rigid motion that differs from egomotion, or because of a non-rigid motion. For the MDD method, the detection of non-rigid motion is not only possible but easier, too. For the case of non-rigid motion, the hypothesis of eq. (6.8) does not hold. Therefore, for an image patch that contains points of a non-rigid motion it is very likely that the criterion of eq. (6.8) will not hold for the same parameters s and t for all points n the patch. This should be contrasted to the case of a rigid independent motion: In this case only the 3D motion discontinuities can be effectively detected. Finally, for the IMDL method, rigidity is assumed for the independent motion, as a means to enable the integration of motion information through the various depth layers.

The four proposed methods differ considerably in the representation of their output. The output of IMDE and IMDL is a 2D map of labels for the image points. Different labels correspond

Criterion	**IMDE**	**IMDL**	**MDD**	**DMO**
Egomotion implied:	Unrestricted rigid 3D motion	Unrestricted rigid 3D motion	FOE and AOR far from the image patch	Constant 3D egomotion for three frames
Independent motion that can be detected:	No constraint (rigid or nonrigid)	Unrestricted rigid 3D motion	No constraint (rigid or nonrigid)	Changes in motion
Representation of output:	3D motion segmentation map	3D motion segmentation map	3D motion discontinuities map	3D motion changes map
Maximum number of independently moving points	50% of the normal flows of a scene	50% of the normal flows of a scene	No constraint	No constraint
Stereo config. used:	Fixating or parallel	Fixating or parallel	Fixating or parallel	No stereo conf. used
Noise tolerance:	Very high	High	Very high	Moderate
Motions that can be discriminated:	All rigid motions	Rigid motions with different FOEs	All rigid motions	Maneuvering objects, changes in egomotion
Additional info provided:	Parameters of dominant 3D motion (egomotion)	Ordering of depth layers	No	Changes of observers' 3D motion
Dependence on parameter thresholds:	(a) Unreliable normal flow rejection, (b) Outlier char. thresh.	(a) Unreliable normal flow rejection, (b) Outlier char. thresh., (c) Width of each layer, (d) 3D motion distance	(a) Unreliable normal flow rejection, (b) T_{IM} threshold	(a) Unreliable normal flow rejection, (b) ''change occurred'' threshold
Computational Requirements:	High	High	Medium	Very low

Table 9.1: Comparative overview of the proposed independent motion detection methods.

to different 3D motions. For MDD, the labels map is binary valued and coarse. Labels are assigned to image patches (as opposed to image points). The two possible values represent the two different cases that can be detected by the method: Either all points of a tile belong to the same 3D motion, or there is a 3D motion discontinuity. For the DMO method the labels map is binary valued and the labels are assigned at the pixel level. Each label represents constancy or change of the 3D motion parameters of a point within 3 image frames.

The high breakdown point of LMedS, enables IMDE to detect an independent motion that covers up to 50% of the total number of points with reliable normal flow vectors. This constraint (which is already very loose in practical situations), is transferred, in the case of IMDL, to the points of one depth layer. Thus, there are certain configurations of independently moving objects that cannot be detected by IMDE but which can be detected by IMDL and vice versa. The local operations that are performed by the MDD and DMO methods do not limit the spatial extent of the detectable independent motions.

Besides the DMO method that exploits motion information only, all other methods use some form of stereoscopic configurations. More specifically, the IMDE and MDD method can work with both a fixating and a parallel stereo configuration. The IMDL method uses in principle a parallel stereo configuration. In Chapter 4 it has been demonstrated that depth layering may also be performed through robust regression in the normal flow field produced by a fixating stereo configuration. However, the breakdown point of LMedS (50%), imposes certain relations of the number of points per depth layer. This is, in fact, an environmental assumption which should be avoided whenever possible.

The experimental results have shown that the IMDE algorithm performs best compared to the rest of the methods. From the rest of the methods, IMDL has a very good performance at each depth layer, but it encounters some problems in the phase of combination of results, if the error levels in the normal flow fields become too high. The MDD method is fast and robust but produces coarse maps of the 3D motion discontinuities.

For IMDE and MDD there are no ambiguous motions (according to the definition of motion ambiguity in section 5.4). This is not the case with IMDL which cannot discriminate motions that have the same rotational components and their translational components differ by a constant factor. The DMO method reports on the constancy of 3D motion rather than on the existence of independent motion.

Except from the task of independent motion detection, some of the proposed methods provide additional useful information regarding the observer or the environment. Thus, IMDE gives an estimation of the dominant 3D motion parameters (egomotion parameters) and an estimation of the vergence angle of the stereoscopic system, IMDL gives an ordering of the depth layers and DMO gives information on the changes of egomotion.

An important characteristic of the proposed methods, is that their performance does not depend on the fine tuning of a large number of parameters. A prerequisite for all methods is the computation of reliable normal flow fields, which is achieved through the rejection of the normal flow vectors for which the temporal and spatial derivative do not meet certain conditions. Besides this threshold, the IMDE method uses only one additional threshold that is used to discriminate the outliers from the inliers. The criterion of eq. (3.32) performed very well in all conducted experiments. Therefore, for all practical purposes, it can be treated as an algorithmic constant rather an algorithmic parameter. The MDD method uses a threshold to decide on the existence of more than one 3D motions within an image patch. The DMO method uses one threshold in order to decide whether the 3D motion of a point has been changed within three frames. The IMDL method depends on more parameters compared to the other methods. Two extra thresholds need to be employed, one that determines the width of a depth layer and a threshold on the distance measure between two motion segments (eq. (5.5)).

Finally, from a computational performance point of view, DMO method is extremely fast, because it involves the computation and the pointwise comparison of two normal flow fields. On the other side, IMDE is the method with the largest computational requirements because it

employs robust regression over the full set of points with reliable normal flow vectors. However, the execution time for the full method is shorter than the execution of many state of the art algorithms that just compute optical flow. IMDL and MDD lie in the middle of this ordering, since they employ robust estimation, but in subsets of the points with reliable normal flow vectors. The computational performance of all methods depends on the number of reliable normal flow vectors. IMDL also depends on the distribution of points in depth layers. In the one extreme, all points lie in one depth layer and the performance is similar to the performance of IMDE. If the scene points are distributed in a large range of depths, then the computational performance of the algorithm is improved. This is because the overhead that is due to the manipulation of more layers is small compared to the gain because of the application of LMedS in smaller data sets.

Besides the independent interest in each of the proposed methods, what is extremely important is the fact that they can be combined into a general framework for independent motion detection. This combination leads to an integrated system that shares the merits of all the proposed methods.

9.3 Putting pieces together

In Chapter 7 it has been demonstrated that by using DMO we are able to recover information on the constancy of motion of the observer and/or of some independently moving objects. Suppose that a robot equipped with visual sensors is able to use the four independent motion detection schemes already described. Suppose also that the robot starts moving with constant 3D motion parameters in a static environment. As long as the robot preserves its motion parameters and nothing starts moving in the environment, the situation described in case (1) of the discussion in Section 7.2 will be encountered. Therefore, there is no need for any other independent motion detection algorithm to be invoked. If an object starts moving, then the change of its motion parameters will be detected by the DMO mechanism. If the independently moving object moves

non rigidly, or in a continuously changing manner, then again, DMO is able to detect it (case (2) in Section 7.2) and no other independent motion detection mechanism is needed. If, however, the independently moving object ends up moving at a constant set of 3D motion parameters, then DMO will no more be able to detect it. Therefore, one of the three other independent motion detection schemes must be invoked, depending on the information required for the goals of system and on the available computational resources that can be devoted to this type of processing.

Figure 9.1 illustrates the general layout of the scheme for independent motion detection. At

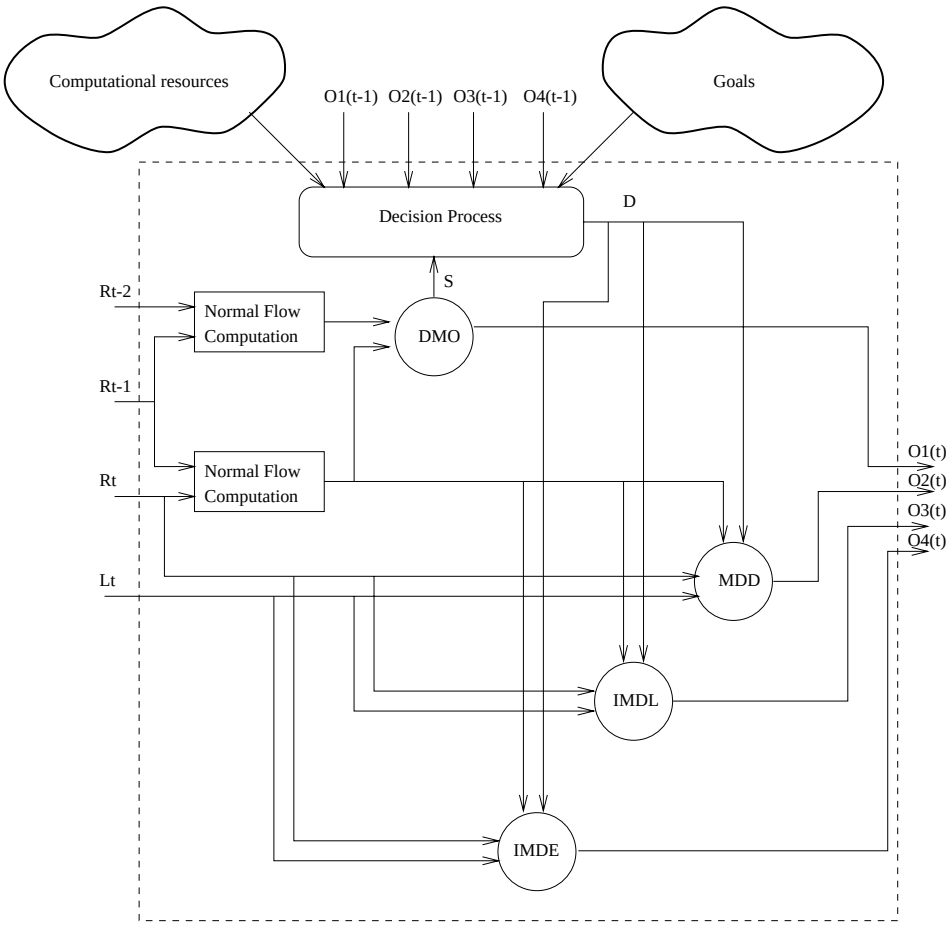


Figure 9.1: A schematic view of the unified framework for independent motion detection.

each moment in time, the system computes the normal flow fields between the images of the right camera, corresponding at time instances $t-1$ and $t-2$ and $t-1$ and t. From these normal

flow fields, the DMO method extracts a map O_1 of maneuvering objects. This map can be further used by other visual capabilities (i.e. methods for recognizing the maneuvering object) Another important outcome of DMO, is the signal $\mathcal{S}$, which codes the 4 different situations presented in Section 7.2. Signal $\mathcal{S}$ is input to a simple decision process which implements the logic described earlier. This process also takes account the systems' goals and computational resource limitations in order to decide which of the three independent motion detection algorithms should be employed. The output of decision process is a combined signal D which actually controls the activation of one off the IMDE, IMDL or MDD methods. For example, if the observer needs a full map of independently moving objects then he can use the IMDE method. If he is only interested in a specific depth layer, then he can use the IMDL method for independent motion detection and exploit only the results concerning the depth layer of interest. Finally if just needs coarse information about independent motion, then he can exploit the MDD method which results in computational savings.

9.4 Characteristics of the unified framework

The proposed framework for integrating independent motion detection capabilities has a number of advantages. A key characteristic of the integration is that it follows very closely the principles of purposive vision. The different capabilities provide a number of alternative representations for independent motion from which the system may choose, based on its goals and resource limitations. The goals are in essence those which determine the problem to be solved: If simpler representations suffice, those are provided; as a consequence, lower computational overhead is paid. But the idea also works the other way around: If there are not enough computational resources to be devoted for this task, then it may be desirable to get a coarse idea with a simpler algorithm rather than having no information at all.

The integrated independent motion detection system, is a characteristic example of a system that is built based on progressively more complicated competences. There is a *horizontal*

organization of the system processes rather than a *vertical organization*. If, for example, one of the four capabilities is excluded from the system, then the system will continue to operate, probably with a reduced overall performance. This is completely different to systems with vertically organized modules, where, eliminating a module is catalytic for the operation of the whole system. Moreover, the activation of modules is not predetermined but dynamically evolving and dependent on the systems' goals and computational resources.

The system described treats the task of independent motion detection with considerable computational gains over the IMDE, IMDL and MDD methods. Performance gains are achieved by careful consideration of the problem at hand. Most of the time the system handles independent motion detection with the solution of a computationally trivial problem, that of computing and comparing normal flow fields in a pointwise manner. More elaborate and computationally intensive methods are involved only when necessary. In many cases (which may well be the majority of cases), independent motion detection will be handled by the DMO method which has very low computational requirements. In these cases, even the computation of stereoscopic information is avoided.

Last but not least, the presented integration framework is easily extensible. Capabilities that are related with independent motion detection can easily be integrated to the existing scheme. For example, a change detection algorithm could be easily incorporated into the system in order to provide evidence on whether the observer is moving or not. In this case, simple change detection over two consecutive image frames suffices to detect independent motion, without the need to compute any motion related representation of the scene.

Chapter 10

Conclusions

Success is never final

Winston Churchill

Every moving biological organism with the sense of vision has the ability to detect independent 3D motion [74]. Robots, like animals, should operate in dynamic environments that consist of both stationary as well as moving objects. The perception of independent motion is of crucial importance because it provides useful information on where attention should be initially focused and then maintained. A moving object may represent a target to be tracked or avoided. In any case, independent motion is an interesting event for which an appropriate and possibly immediate reaction of the observer is needed.

In this dissertation, independent 3D motion detection was based on *precategorical* visual processing, i.e. the observer was not required to recognize the independently moving object(s). This is actually in agreement with experiments according to which biological organisms are capable to perceive motion before they understand what is really moving. Thus independent

motion detection was based only on motion and depth information that the observer may acquire as he moves in the 3D space. Four novel independent motion detection methods have been proposed which share a number of very important characteristics:

- They employ 3D motion models rather 2D motion models. As such, they are able to perform satisfactorily even in scenes with considerable depth variations. Moreover, they do not make any assumptions regarding the external world in the form of restrictive geometrical modeling.
- They are based on the computation of the normal flow field (which can be accurately computed from sequences of images) and not on the computation of optical flow field or on the solution of the correspondence problem, which are ill-posed problems. As it has been discussed in Chapter 3, the computation of optical flow field is based on environmental assumptions which do not hold in practice, especially when independently moving objects are present in the scene.
- They do not make any assumptions about the motion of the observer[1], which can move with unrestricted translational and rotational 3D motion.
- They do not rely on any knowledge regarding the baseline and the vergence angle of the stereoscopic configuration. Moreover, they perform very well even in case that the intrinsic camera parameters (image origin, focal length) are not precisely known.
- They are capable to detect rigid and non-rigid independent motion, being tolerant to high levels of noise.

Besides the beneficial characteristics of each of the proposed methods, a scheme for integrating these methods has been presented. This scheme tries to couple the computational characteristics of the simpler methods with the rich information that is provided by the more computationally intensive ones. This integration follows closely the principles of purposive

[1]Except the MDD method which places some qualitative constraints on egomotion.

vision, because the methods are activated based on the goals of the system and its resource limitations.

The limitations of the methods described should be clear. First, computing normal flow from stereo can be achieved robustly with a parallel stereo configuration with a small baseline, or by a fixated stereo configuration, in a portion of the scene around the fixation point. This limits the choices of configurations that can be used in practice. Second, although the approach taken by the proposed independent motion detection methods follows the purposive theory of vision, none of the methods exploits action (in the form of controlled movement) in order to simplify the problem that should be solved.

10.1 Future work

Future work in the area of independent motion detection could take several directions. Some of them have already been stated implicitly, since they originate from the limitations of the proposed methods.

Methods that will give rise to depth functions by using a broader class of stereo configurations should be further investigated. Ongoing research tries to exploit multiresolution strategies for the computation of normal flow. Certain efforts are focused towards non-uniform image transformations (e.g. log-polar transformation) which is particularly suited for computing stereo normal flow by a fixated stereo configuration.

Tracking of a moving object by means of gaze holding mechanisms, could lead to the solution of simpler problems regarding independent motion detection. Tracking alone cannot effectively handle the problem of independent motion detection. This is because it only enables the maintenance of attention on an independently moving object, but cannot aid the initial attraction of attention. The incorporation of such mechanisms in the unified framework could prove beneficial for the successful solution of the overall independent motion detection problem.

Change detection methods can also be incorporated in the proposed integration framework. In the same way that the DMO method provided useful information on the constancy of the 3D motion parameters of the observer and of some independently moving objects, change detection algorithms applied to the image intensities can provide useful signals on the constancy of the location of the observer and on the presence of independently moving objects.

A dual effect of the detection of independently moving objects is the isolation of image areas that appear to be moving only due to egomotion. This enables the effective isolation of points that can be used as robust input for the estimation of the egomotion. It should be noted that the vast majority of methods for 3D egomotion estimation assume that there are no independently moving objects in the field of view of the observer. Thus, having already identified the independently moving objects of a scene, enables all these methods to become operational under looser assumptions.

Another axis of future work is related to the exploitation of a larger time window. All the proposed methods are able to extract information about independent motion from two successive frames in time. The use of more frames provides redundant information that can be exploited in several ways. One such way is the qualitative characterization of the motion of independently moving objects. Certain higher level characteristics of the motion of independently moving objects cannot be recovered by processing just two frames, but require a longer sequence of images. For example, the movement of animals is characterized by specific periodical patterns of motion [126]. The recognition of such patterns requires the processing of long image sequences.

Finally, in the more general context of building machines with space perception capabilities, it is very interesting to investigate how other visual cues (e.g. color, texture e.t.c) are combined with motion information in order to provide the rich perceptual information that enables the interaction with a dynamic world and the achievement of the goals of the system. Towards this direction, it should be studied whether the proposed integration framework can be support the successful coordination and integration of many processes that will cooperate in order

to solve specific problems. At this level of analysis, a number of important issues are raised regarding memory organization, learning capabilities, action selection, control strategies, resource allocation and many others. Each of these issues constitutes a broad area for research on its own right. It is evident that despite the important results in understanding the various aspects of perception and cognition, machines are far from competing with their simplest biological counterparts.

Appendix A

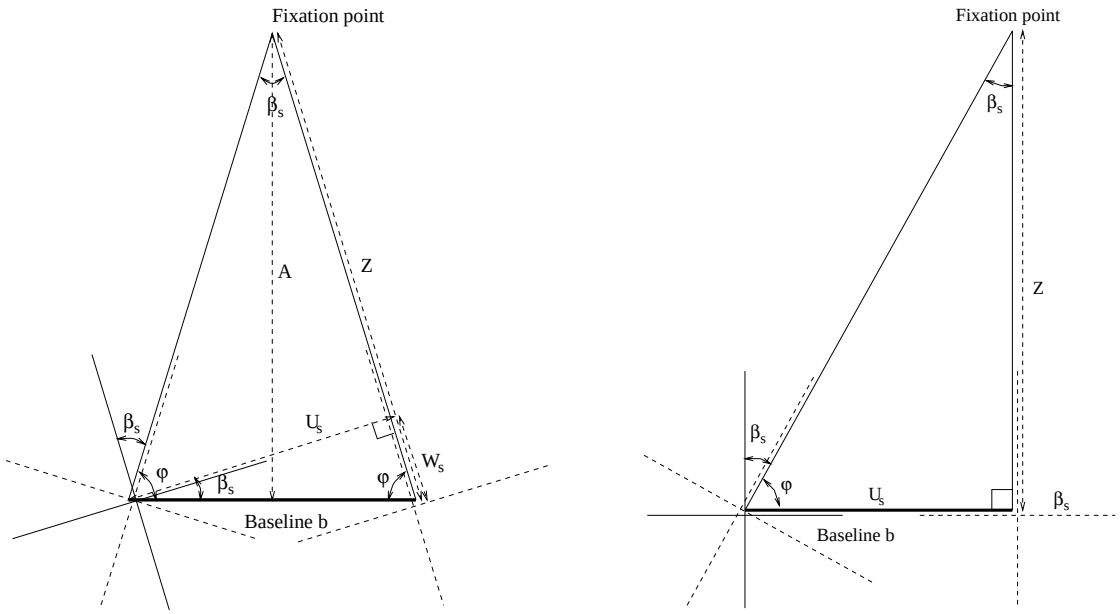
The significance of the term W_s for stereo normal flow

Consider the stereo configuration of Fig. A.1(a). In this stereo configuration, the optical axes of the two cameras form an equilateral triangle. One side of the formed triangle is the stereo baseline and the other two sides are equal to the distance of each camera from the fixation point. The stereo equivalent motion that transforms the position of the one camera to the position of the other includes a rotation β_s, and two translations U_s and W_s. However, W_s is, in the general case, very small compared to U_s. This is because in practical situations the distance Z of each of the cameras from the observed scene points is much larger compared to the length of the baseline. This, causes the vergence angle β_s to be a very small angle and the angles ϕ to be approximately right angles. Assume for example a scene at a distance of $A = 3m$ from the observer's baseline and a baseline $b = 7cm$ (the typical baseline for an adult human). By using elementary geometry (see Fig. A.1(a)) we get:

$$tan(\phi) = \frac{A}{b/2} \Rightarrow \phi = 89.33^\circ$$

Therefore, $\beta_s = 180^\circ - 2\phi = 1.34^\circ$. From Fig. A.1(a) it can also be verified that:

$$tan(\beta_s) = \frac{W_s}{U_s} \Rightarrow W_s = 0.0234 U_s$$



(a) (b)

Figure A.1: (a) An equilateral and (b) a right-angled stereo configuration.

which means that W_s is two orders of magnitude smaller than U_s. Note also that in eq. (3.14), U_s is multiplied by $T_{U_s} = \frac{n_x f}{Z}$ and W_s is multiplied by $T_{W_s} = \frac{x n_x + y n_y}{Z}$. Given that in practical situations, f is a few times larger than the maximum x or y coordinate of a point in the image, it is evident that the term that contains W_s is negligible compared to the remaining terms of eq. (3.14).

In the case of the right angled stereo configuration of Fig. A.1(b), W_s is exactly equal to zero. A rotation β_s and a translation $U_s = b$ suffice to transform the coordinate system of the right camera to the coordinate system of the left camera. Such a right angled stereo configuration has also been used by others [152].

Bibliography

[1] E.H. Adelson and J.R. Bergen. Spatiotemporal Energy Models for the Perception of Motion. *Journal of the Optical Society of America A*, 2:284--299, 1985.

[2] G. Adiv. Determining Three Dimensional Motion and Structure from Optical Flow Generated by Several Moving Objects. *IEEE Transactions on PAMI*, PAMI-7(4):384--401, July 1985.

[3] G. Adiv. Inherent Ambiguities in Recovering 3D Motion and Structure from a Noisy Flow Field. *IEEE Transactions on PAMI*, PAMI-11(5):477--489, May 1989.

[4] N. Ahuja and A.L. Abbott. Active Stereo: Integrating Disparity, Vergence, Focus, Aperture, and Calibration for Surface Estimation. *IEEE Transactions on PAMI*, 15(10):1007--1029, October 1993.

[5] P.K. Allen, A. Timcenko, B. Yoshimi, and P. Michelman. Automated Tracking and Grasping of a Moving Object with a Robotic Hand-Eye System. *IEEE PAMI*, 9(2):152--165, April 1993.

[6] Y. Aloimonos. PEGASUS: A Real-time System of Intelligent Agents. Personal Communication.

[7] Y. Aloimonos. Purposive and Qualitative Active Vision. In *Proceedings DARPA Image Understanding Works.*, pages 816--828, 1990.

[8] Y. Aloimonos. Introduction: Active Vision Revisited. In Y. Aloimonos, editor, *Active Perception*, pages 1--18. Erlbaum Associates, Hilsdale, New Jersey, 1993.

[9] Y. Aloimonos and A. Bandopadhay. Active Vision. In *IEEE 1st International Conference on Computer Vision*, pages 35--54, June 1987.

[10] Y. Aloimonos and C.M. Brown. On the Kinetic Depth Effect. *Biological Cybernetics*, 60:445--455, 1989.

[11] Y. Aloimonos and Z. Duric. Estimating the Heading Direction using Normal Flow. *International Journal of Computer Vision*, 13(1):33--56, 1994.

[12] Y. Aloimonos and D. Tsakiris. On the Visual Mathematics of Tracking. *Image and Vision Computing*, 9(4):235--251, August 1991.

[13] Y. Aloimonos, I. Weiss, and A. Bandopadhay. Active Vision. *International Journal of Computer Vision*, 2:333--356, 1988.

[14] P. Anandan. A Computational Framework and an Algorithm for the Measurement of Visual Motion. *International Journal of Computer Vision*, 2:283--310, 1989.

[15] P. Anandan and R. Weiss. Introducing a Smoothness Constraint in a Matching Approach for the Compuation of Optical Flow Fields. In *3rd International Workshop on Computer Vision: Representation and Control*, pages 186--194, 1985.

[16] M.A. Arbib. Perceptual Structures and Distributed Motor Control. In Brooks VB, editor, *Handbook of Physiology - The nervous system II. Motor Control*, pages 1449--1480, Bethesda, MD, 1981. American Physiological Society.

[17] A.A. Argyros, M.I.A Lourakis, P.E. Trahanias, and S.C. Orphanoudakis. Fast Visual Detection of Changes in 3D Motion. In *IAPR MVA '96, Tokyo, Japan*, November 12-14 1996.

[18] A.A. Argyros, M.I.A Lourakis, P.E. Trahanias, and S.C. Orphanoudakis. Independent 3D Motion Detection Through Robust Regression in Depth Layers. In *British Machine Vision Conference (BMVC '96), Edinburgh, UK*, September 9-12 1996.

[19] A.A. Argyros, M.I.A Lourakis, P.E. Trahanias, and S.C. Orphanoudakis. Qualitative Detection of 3D Motion Discontinuities. In *IROS '96, Tokyo, Japan*, November 4-8 1996.

[20] A.A. Argyros, P.E. Trahanias, and S.C. Orphanoudakis. Robust Regression for the Detection of Independent 3D Motion by a Binocular Observer. Technical Report TR 158, FORTH-ICS, 1995.

[21] A.A. Argyros, P.E. Trahanias, and S.C. Orphanoudakis. Robust Regression for the Detection of Independent 3D Motion by a Binocular Observer. *Real Time Imaging (submitted)*, 1996.

[22] S. Ayer, P. Schroeter, and J. Bigun. Segmentation of Moving Objects by Robust Motion Parameter Estimation over Multiple Frames. In *European Conference on Computer Vision*, 1994.

[23] R. Bajcsy. Active Perception. *Proceedings of the IEEE*, 76(8):996--1005, August 1988.

[24] R. Bajcsy and M. Campos. Active and Exploratory Perception. *CVGIP:Image Understanding*, 56(1):31--40, July 1992.

[25] D. Ballard. Animate Vision. Technical report, University of Rochester, 1990.

[26] D. Ballard and C. Brown. Principles of Animate Vision. *Computer Vision, Graphics and Image Processing*, 56(1):3--21, 1992.

[27] D.H. Ballard and C.M. Brown. Principles of Animate Vision. In Yiannis Aloimonos, editor, *Active Perception*, page 254. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

[28] Y. Baram and Y. Barniv. Obstacle Detection by Recognizing Binary Expansion Patterns. Technical Report CIS-9321, Technion, Israel, September 1993.

[29] S.T. Barnard and W.B. Thompson. Disparity Analysis of Images. *IEEE Transactions on PAMI*, 2:333--340, 1980.

[30] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of Optical Flow Techniques. *International Journal of Computer Vision*, 12(1):43--77, 1994.

[31] M. Bertero, T. A. Poggio, and V. Torre. Ill-Posed Problems in Early Vision. *Proceedings of the IEEE*, 76(8):869--889, August 1988.

[32] P.J. Besl, R.C. Jain, and L.T. Watson. Robust Window Operators. In *Proceedings of IEEE International Conference on Computer Vision*, pages 591--600, 1988.

[33] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75--104, January 1996.

[34] M. Bober and J. Kittler. Estimation of Complex Multimodal Motion: An Approach Based on Robust Statistics and Hough Transform. *Image and Vision Computing*, 12:661--668, December 1994.

[35] P. Bouthemy and E. Francois. Motion Segmentation and Qualitative Dynamic Scene Analysis from an Image Sequence. *International Journal of Computer Vision*, 10(2):157--182, 1993.

[36] K.L. Boyer, M.J. Mirza, and G. Ganguly. The Robust Sequential Estimator: A General Approach and its Application to Surface Organization in Range Data. *IEEE Transactions on PAMI*, PAMI-16:987--1001, 1994.

[37] R.A. Brooks. A Robust Layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation*, RA-2(7):14--23, April 1986.

[38] R.A. Brooks. Intelligence Without Reason. Technical Report AILAB Memo 1293, Massachusetts Institute of Technology Artificial Intelligen ce Laboratory, April 1991.

[39] C. Brown. The Rochester Robot. Technical report, University of Rochester, 1988.

[40] C. Brown, D. Coombs, and J. Soong. Real-Time Smooth Pursuit Tracking. In A. Yuille A. Blake, editor, *Active Vision*, Artificial Intelligence, chapter 8, pages 123--136. MIT Press, Cambridge, Mass., 1993.

[41] V. Bruce and P. Green. Perceiving Depth and Movement. In Bruce V and Green P, editors, *Visual Perception Physiology, Psychology and Ecology*, chapter 6, pages 129--161. Erlbaum Associates, 1985.

[42] K. Brunnstrom, J.O. Eklundh, and T. Uhlin. Active Fixation for Scene Exploration. *International Journal of Computer Vision*, 17(2):137--162, February 1996.

[43] E. Brunswik. Perception and the Representative Design of Psychological Experiments. Univ. of California Press, Berkeley, 1956.

[44] W. Burger and B. Bhanu. Estimating 3-D Egomotion from Perspective Image Sequences. *IEEE Trans. on Image Processing*, 12(11):1040--1058, Nov. 1990.

[45] P.J. Burt and C. Yen X. Xu. Multiresolution Flow Through Motion Analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 246--252, 1983.

[46] P. Cavanagh. Reconstructing the Third Dimension: Interactions Between Color, Texture, Motion, Binocular Disparity, and Shape. *Computer Vision and Image Processing*, 37:171--195, 1987.

[47] C.J. Cheng and J.K. Aggarwal. A Two-stage Hybrid Approach to the Correspondence Problem via Forward Searching and Backword Correcting. In *International Conference on Pattern Recognition*, pages 173--179, 1990.

[48] T. Chin, W. Karl, and A. Willsky. Probabilistic and Sequential Computation of Optical Flow Using Temporal Coherence. *IEEE Transactions on Image Processing*, 3:773--788, November 1994.

[49] J. Clark and N. Ferrier. Attentive Visual Servoing. In A. Yuille A. Blake, editor, *Active Vision*, Artificial Intelligence, chapter 9, pages 137--154. MIT Press, Cambridge, Mass., 1993.

[50] J. C. Clarke and A. Zisserman. Detection and Tracking of Independent Motion. *Image and Vision Computing*, 14:565--572, 1996.

[51] T.S. Collet, E. Dillmann, A. Giger, and R. Wehner. Visual Landmarks and Route Following in Desert Ants. *Journal of Comparative Physiology*, 170:435--442, 1992.

[52] D.J. Coombs. *Real-time Gaze Holding in Binocular Robot Vision*. PhD Dissertation, Computer Science Department, University of Rochester, 1992.

[53] D.J. Coombs and C.M. Brown. Real-Time Binocular Smooth Pursuit. *International Journal of Computer Vision*, 11(2):147--164, 1993.

[54] B. Crespi, C. Furlanello, and L. Stringa. A Memory-based Approach to Navigation. *Biological Cybernetics*, 69:385--393, 1993.

[55] R. Cypher and J.L.C. Sanz. SIMD Architectures and Algorithms for Image Processing and Computer Vision. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(12):2158--2172, December 1989.

[56] J. Van der Speigel, G. Kreider, C. Claeys, I. Debusschere, G. Sandini, P. Dario, F. Fantini, P. Bellutti, and G. Soncini. A Foveated Retinal-line Sensor Using CCD Technology. In C. Mead and M. Ismail, editors, *Analog VLSI and Neural Network Implementations*, Boston, 1989. DeKluwer Pubs.

[57] R. Deriche. Fast Algorithms for Low Level Vision. *IEEE Transactions on PAMI*, 12(1):78--87, January 1990.

[58] U. R. Dhond and J. K. Aggarwal. Structure from Stereo - A Review. *IEEE Trans. on Systems, Man and Cybernetics*, 19(6):1489--1510, November/December 1989.

[59] R. Dutta. *Depth from Motion and Stereo: Parallel and Sequential Algorithms, Robustness and Lower Bounds.* PhD dissertation, Department of Computer Science, University of Massachusetts Amherst, 1994.

[60] P.J. Hatcher et al. Data Parallel Programming on MIMD Computers. *IEEE Transactions on Parallel and Distributed Systems*, 3(2):377--383, July 1991.

[61] H.R. Everett. *Sensors for Mobile Robots.* A K Peters Ltd, 1995.

[62] J.Q. Fang and T.S. Huang. Some Experiments on Estimating the 3-D Motion Parameters of a Rigid Body from Two Consecutive Frames. *IEEE Transactions on PAMI*, 6:547--554, 1984.

[63] M. Farah. *Visual Agnosia: Disorders of Object Recognition and What They Tell us About Normal Vision.* MIT Press, Cambridge, MA, 1990.

[64] C. Fermuller. *Basic Visual Capabilities.* PhD Dissertation, Center for Automation Research, University of Maryland, 1993.

[65] C. Fermuller. Navigational Preliminaries. In Yiannis Aloimonos, editor, *Active Perception*, chapter 3. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

[66] C. Fermuller and Y. Aloimonos. The Role of Fixation in Visual Motion Analysis. *International Journal of Computer Vision*, 11(2):165--186, 1993.

[67] C. Fermuller and Y. Aloimonos. Vision and Action. *Image and Vision Computing*, 13(10):725--744, December 1995.

[68] G.D. Fischbach. Mind and Brain. *Scientific American*, 267(3):24--41, September 1992.

[69] M.A. Fischler and R.C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *CACM*, 24:381--395, 1981.

[70] D.J. Fleet and A.D. Jepson. Computation of Component Image Velocity from Local Phase Information. *International Journal of Computer Vision*, pages 77--104, 1990.

[71] P. Fornland. Direct Obstacle Detection and Motion from Spatio-Temporal Derivatives. In *CAIP96*, pages 874--879, Prague, Czech Republic, September 1995.

[72] J. Gibson. The Ecological Approach to Visual Perception. Houghton, Mifflin, Boston, 1979.

[73] R.C. Gonzalez and R.E. Woods. *Digital Image Processing*, chapter 7, pages 418--420. Addison-Wesley, third edition, 1992.

[74] R.L. Gregory. *Eye and Brain: The Psychology of Seeing*. Princeton University Press, Princeton, New Jersey, 1990.

[75] D. Heeger. Optical Flow Using Spatiotemporal Filters. *International Journal of Computer Vision*, 1:279--302, 1988.

[76] H. Helmholtz. Handbuch der physiologischen optik. Leopold Voss, 1869.

[77] E. Hildreth. Computations Underlying the Measurements of Visual Motion. *Artificial Intelligence*, 23:309--354, 1984.

[78] G.E. Hinton. How Neural Networks Learn from Experience. *Scientific American*, pages 105--109, September 1992.

[79] J. Hochberg. Machines Should Not See as People Do, but Must Know How People See. *Computer Vision, Graphics and Image Processing*, 37:221--237, 1987.

[80] D.D. Hoffman. Infering Local Surface Orientation from Motion Fields. *Journal of the Optical Society of America A*, 72:880--892, 1982.

[81] R. Horaud and T. Skordas. Stereo Correspondence Through Feature Grouping and Maximal Cliques. *IEEE Transactions on PAMI*, 11(11):1168--1180, November 1989.

[82] B.K.P. Horn. *Robot Vision*. MIT Press, Cambridge, MA, 1986.

[83] B.K.P. Horn and B. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17:185--203, 1981.

[84] A. Horridge. The Evolution of Visual Processing and the Construction of Seeing Systems. In *Proc. of the Royal Society, London B 230*, pages 279--292, 1987.

[85] I.D. Horswill and R.A. Brooks. Situated Vision in a Dynamic World: Chasing Objects. In *Proceedings of AAAI*, 1988.

[86] Y. Hsu, H.H. Nagel, and G. Rekkers. New Likelihood Test Methods for Change Detection in Image Sequences. *Computer Vision, Graphics and Image Processing*, 26:73--106, 1984.

[87] L. Huang and Y. Aloimonos. The Geometry of Visual Interception. Technical Report CAR-TR-622, CS-TR-2893, University of Maryland, April 1992.

[88] Y. Huang, K. Palaniappan, X. Zhuang, and J.E. Cavanaugh. Optic Flow Field Segmentation and Motion Estimation Using a Robust Genetic Partitioning Algorithm. *IEEE Transactions on PAMI*, 17(12):1177--1190, December 1995.

[89] D. Hubel and T. Wiesel. Receptive Fields and Functional Architecture of the Monkey Striate Cortex. *Journal of Physiology*, 195:215--243, 1968.

[90] P.J. Huber. *Robust Statistics.* John Wiley and Sons Inc., New York, 1981.

[91] A. Hurlbert and T. Poggio. Do Computers Need Attention? *Nature*, 321:651--652, June 1986.

[92] J. Illingworth and J. Kittler. A Survey of the Hough Transform. *Computer Vision, Graphics and Image Processing*, 44:87--116, 1988.

[93] M. Irani, B. Rousso, and S. Peleg. Computing Occluding and Transparent Motions. *International Journal of Computer Vision*, 12(1):5--16, 1994.

[94] R.C. Jain. Segmentation of Frame Sequences Obtained by a Moving Observer. *IEEE Transactions on PAMI*, PAMI-7(5):624--629, September 1984.

[95] E. Johnston. Systematic Distortions of Shape from Stereopsis. *Vision Research*, 31:1351--1360, 1991.

[96] J.M. Jolion, P. Meer, and S. Bataouche. Robust Clustering with Applications in Computer Vision. *IEEE Transactions on PAMI*, 13:791--802, 1995.

[97] J.W.Y. Kam. A Real-time 3D Motion Tracking System. Master's thesis, Computer Science Department, University of British Columbia, April 1993.

[98] P. Kanerva. *Sparse Distributed Memory*. MIT Press, Cambridge, MA, 1988.

[99] W. Kohler. Gestalt psychology. Liverght, New York, 1947.

[100] C.E. Kolb. *Rayshade User's Guide and Reference Manual*, 0.4 edition, January 1992.

[101] J. Konczak. Towards an Ecological Theory of Motor Development: The Relevance of the Gibsonian Approach to Vision for Motor Development Research. In J. E. Clark and J. H. Humphrey, editors, *Advances in Motor Development Research*, volume 3, pages 201--224. 1991.

[102] R. Kumat and A.R. Hanson. Robust Methods for Estimating Pose and a Sensitivity Analysis. *CVGIP:Image Understanding*, 1994.

[103] T. Lawton. Processing Translational Motion Sequences. *Computer Vision, Graphics and Image Processing*, 22:116--144, 1983.

[104] C.H. Lee and A. Joshi. Correspondence Problem in Image Sequence Analysis. *Pattern Recognition*, 26(1):47--61, 1993.

[105] L. Li and J. Duncan. 3-D Translational Motion and Structure from Binocular Image Flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-15(7):657--667, July 1993.

[106] J.J. Little and G.E. Blelloch. Algorithmic Techniques for Computer Vision on a Fine Grained Parallel Machine. *IEEE Transactions on PAMI*, 11(3):244--257, March 1989.

[107] M. Livingstone. Segregation of Form, Color, Movement, and Depth Processing in the Visual System: Anatomy, Physiology, Art and Illusion. In Cohen B and Bodis-Wollner I, editors, *Vision and the Brain*, pages 119--138. Raven Press Ltd., 1990.

[108] N. V. Lobo and J. K. Tsotsos. Computing Egomotion and Detecting Independent Motion from Image Motion Using Collinear Points. *Computer Vision and Image Understanding*, 64(1):21--52, July 1996.

[109] H.C. Longuet-Higgins. A Computer Algorith for Reconstruction of a Scene from Two Projections. *Nature*, 293:133--135, 1981.

[110] H.C. Longuet-Higgins and K. Prazdny. The Interpretation of a Moving Retinal Image. In *Proceedings of the Royal Society*, pages 385--397. London B, 1980.

[111] M. Lourakis. An Alternative Approach for Studying Space Perception: Use of Ordinal instead of Metric Depth. Master's thesis, University of Crete, Computer Science Department, 1995.

[112] D. Marr. *Vision*. W. H. Freeman, San Francisco, 1982.

[113] B. Marsh, C.M. Brown, T.J. LeBlanc, M.L. Scott, T.J. Becker, P. Das, J. Karlsson, and C.A. Quiroz. Operating System Support for Animate Vision. Technical Report 374, -, University of Rochester, June 1991.

[114] R.A. McCallum. Learning with Incomplete Selective Perception. Master's thesis, University of Rochester, Dept. of Computer Science, Department of Computer Science, University of Rochester, March 1993.

[115] P. Meer, A. Mintz, and A. Rosenfeld. Robust Regression Methods for Computer Vision: A Review. *International Journal of Computer Vision*, 6(1):59--70, 1991.

[116] T. Mintz. Robustness by Consensus. Technical Report CAR-TR-576, Center for Automation Research, University of Maryland, 1991.

[117] H. Moravec. Towards Automatic Visual Obstacle Avoidance. In *International Joint Conference on Artificial Intelligence*, pages 584--585, 1977.

[118] A. Movshon. *Images and Understanding*, pages 122--137. Cambridge University Press, 1990.

[119] D. Murray and A. Basu. Motion Tracking with an Active Camera. *IEEE PAMI*, 16(5):449--459, May 1994.

[120] D.W. Murray and B.F. Buxton. Reconstructing the Optic Flow from Edge Motion: An Examination of Two Different Approaches. In *First Conference on AI Applications*, 1984.

[121] H.H. Nagel. Displacement Vectors Derived from Second order Intensity Variations in Image Sequences. *Computer Vision, Graphics and Image Processing*, 21:85--117, 1983.

[122] V.S. Nalwa. *A Guided Tour of Computer Vision*, chapter 8. Addison-Wesley, 1993.

[123] R.C. Nelson. *Visual Navigation*. PhD Dissertation, University of Maryland, 1988.

[124] R.C. Nelson. Qualitative Detection of Motion by a Moving Observer. *International Journal of Computer Vision*, 7(1):33--46, 1991.

[125] R.C. Nelson and Y. Aloimonos. Obstacle Avoidance Using Flow Field Divergence. *IEEE Transactions on PAMI*, PAMI-11(10):1102--1106, October 1989.

[126] R.C. Nelson and R. Polana. Qualitative Navigation of Motion Using Temporal Texture. *CVGIP:Image Understanding*, 56(1):78--89, July 1992.

[127] P. Nordlund and T. Uhlin. Closing the Loop: Detection and Pursuit of a Moving Object by a Moving Observer. *Image and Vision Computing*, 14:267--275, 1996.

[128] J.M. Odobez and P. Bouthemy. Detection of Multiple Moving Ojects Using Multiscale MRF with Camera Motion Compensation. In *1st IEEE International Conference on Image Processing, ICIP*, 1994.

[129] B.A. Olshausen and C. Koch. Selective Visual Perception. In M.A. Arbib, editor, *The handbook of brain theory and neural networks*, pages 837--840. MIT Press, Cambridge, Mass, 1995.

[130] T.J. Olson and D.J. Coombs. Real-Time Vergence Control for Binoculare Robots. *International Journal of Computer Vision*, 7(1):67--89, 1991.

[131] K. Pahlavan and J.O. Eklundh. A Head-Eye System - Analysis and Design. *CVGIP: Image Understanding, Special Issue on Purposive, Qualitative Active Vision*, 56:41--56, 1992.

[132] K. Pahlavan, T. Uhlin, and J.O. Eklundh. Active Vision as a Methodology. In Yiannis Aloimonos, editor, *Active perception*, chapter 1. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

[133] C.H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice Hall Co., 1982.

[134] N.P. Papanikolopoulos, P.K. Khosla, and T. Kanade. Visual Tracking of a Moving Target by a Camera Mounted on a Robot: A Combination of Control and Vision. *IEEE Transactions on Robot. and Automation*, 9(1):14--25, February 1993.

[135] N. Paragios and G. Tziritas. Detection and Location of Moving Objects Using Deterministic Relaxation Algorithms. In *ICPR96*, Vienna, Austria, September 1996.

[136] G. Patras, N. Alvertos, and G. Tziritas. Joint Disparity and Motion Field Estimation in Stereoscopic Image Sequences. In *ICPR96*, Vienna, Austria, September 1996.

[137] T. Poggio, H. Vorhees, and A. Yuille. Regularizing Edge Detection. Technical Report A.I. Memo 776, MIT, A.I. Lab, Cambridge, MA, 1984.

[138] W.H. Press, S.A. Teukolsky, A.W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C.* Cambridge University Press., New York, 1992.

[139] S. Reddi and G. Loizou. Analysis of Camera Behavior During Tracking. *IEEE PAMI*, 17(8):765--778, August 1995.

[140] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-Free Attentional Operators: The Generalized Symmetry Transform. *International Journal of Computer Vision*, 14:119--130, 1995.

[141] R. Rimey and C. Brown. Task Oriented Vision. In A. Yuille A. Blake, editor, *Active Vision*, Artificial Intelligence, chapter 14, pages 221--233. MIT Press, Cambridge, Mass., 1993.

[142] D.L. Ringach and Y. Baram. A Diffusion Mechanism for Obstacle Detection from Size-change Information. *IEEE Transactions on PAMI*, PAMI-15(12), December 1993.

[143] P.J. Rousseeuw. Least Median of Squares Regression. *Journal of American Statistics Assocciation*, 79:871--880, 1984.

[144] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection.* John Wiley and Sons Inc., New York, 1987.

[145] S. Sarkar and K. Boyer. Perceptual Organization in Computer Vision: A Review and a Proposal for a Classificatory Structure. *IEEE Transactions on Systems, Man and Cybernetics*, 23(2):383--399, 1993.

[146] I. Schmidt, T.S. Collet, F.X. Dillier, and R. Wehner. How Desert Ants Cope with Enforced Detours on Their Way Home. *Journal of Comparative Physiology*, 171:285--288, 1992.

[147] E.L. Schwartz. -. *Presentation given at the NSF Active Vision Workshop*, August 1991.

[148] R. Sedgewick. *Algorithms.* Addison-Wesley, Reading, MA, 1988.

[149] R. Sharma. Robust Detection of Independent Motion: An Active and Purposive Solution. Technical report, Center for Automation Research, University of Maryland, CAR TR-534, College Park, MD, 1991.

[150] R. Sharma and Y. Aloimonos. Early Detection of Independent Motion from Active Control of Normal Image Flow Patterns. *IEEE Transactions on SMC*, SMC-26(1):42--53, February 1996.

[151] A. Shashua. Projective Structure from two Uncalibrated Images: Structure from Motion and Recognition. Technical Report A.I. Memo 1363, MIT, A.I. Lab, September 1992.

[152] Y.Q. Shi, C.Q. Shu, and J.N. Pan. Unified Optical Flow Approach to Motion Analysis from a Sequence of Stereo Images. *Pattern Recognition*, 27(12):1577--1590, 1994.

[153] T. Shipley and H. Shore. The Human Texture Visual Field: Fovea-To-Periphery Pattern Recognition. *Pattern Recognition*, 23(11):1215--1221, 1990.

[154] D. Sinclair, A. Blake, and D. Murray. Robust Estimation of Egomotion from Normal Flow. *International Journal of Computer Vision*, 13(1):57--69, 1994.

[155] A. Singh. *Optical Flow Computation: A Unified Perspective*. PhD Dissertation, Department of Computer Science, Columbia University, New York, NY, 1990.

[156] S.S. Sinha and B.G. Schunk. A Two Stage Algorithm for Discontinuity-Preserving Surface Reconstruction. *IEEE Transactions on PAMI*, PAMI-14:36--55, 1992.

[157] K. Skifstad and R. Jain. Illumination Independent Change Detection for Real World Image Sequences. *Computer Vision, Graphics and Image Processing*, 46:387--399, 1989.

[158] M.E. Spetsakis and Y. Aloimonos. Optimal Motion Estimation. In *IEEE Workshop on Visual Motion*, pages 229--237, 1989.

[159] M.E. Spetsakis and Y. Aloimonos. Structure from Motion Using Line Correspondences. *International Journal of Computer Vision*, 4:171--183, 1990.

[160] A. Stein and M. Werman. Robust Statistics in Shape Fitting. *Computer Vision, Graphics and Image Processing*, pages 540--546, 1992.

[161] C.V. Stewart. MINPRAN: A New Robust Estimator for Computer Vision. *IEEE Transactions on PAMI*, 17(10):925--938, 1995.

[162] Q.F. Stout. Mapping Vision Algorithms to Parallel Architectures. *Proceedings of the IEEE*, 76(8):982--995, August 1988.

[163] M. Subbarao. *Interpretation of Visual Motion*. PhD Dissertation, Center for Automation Research, Univ. of Maryland, College Park, MD, 1988.

[164] Swain and D. Ballard. Object Identification Using Color Cues. Technical report, University of Rochester, 1990.

[165] J. Swain and M.A. Stricker. Promising Directions in Active Vision. *International Journal of Computer Vision*, 11(2):109--126, 1993. Written by the attendees of the NSF Active Vision Works., Univ. of Chicago, August 5-7.

[166] R. Szeliski and S.B. Kang. Recovering 3D Shape and Motion from Image Streams using Non-Linear Least Squares. Technical Report CRL 93/3, Cambridge Research Laboratory, March 1993.

[167] S.L. Tanimoto. *Architectural Issues for Intermediate Level Vision*, chapter 1, pages 3--17. Academic Press, Inc, 1985.

[168] W.B. Thompson, P. Lechleider, and E.R. Stuck. Detecting Moving Objects Using the Rigidity Constraint. *IEEE Transactions on PAMI*, 15(2):162--166, February 1994.

[169] W.B. Thompson and T.C. Pong. Detecting Moving Objects. *International Journal of Computer Vision*, 4:39--57, 1990.

[170] M. Tistarelli and G. Sandini. Dynamic Aspects in Active Vision. *Computer Vision, Graphics and Image Processing*, 56(1):108--129, July 1992.

[171] J.T. Todd and F.D. Reichel. Ordinal Structure in the Visual Perception and Cognition of Smoothly Curved Surfaces. *Psychology Review*, 96:643--657, 1989.

[172] C. Tomasi and T. Kanade. Shape and Motion from Image Streams under Orthography: a Factorization Method. *International Journal of Computer Vision*, 9(2):137--154, 1992.

[173] P. H. S. Torr and D. W. Murray. Stochastic Motion Clustering. In J.-O. Eklundh, editor, *Proceedings of ECCV'94, LNCS, vol. 80*, pages 328--337, 1994.

[174] P.H.S. Torr and D.W. Murray. Statistical Detection of Independent Movement from a Moving Camera. *Image and Vision Computing*, 11:180--187, May 1993.

[175] H.L. Van Trees. *Detection, Estimation, and Modulation Theory*. Wiley, New York, 1968-71.

[176] A Treisman. Preattentive Processing in Vision. *Computer Vision, Graphics and Image Processing*, 31:156--177, 1985.

[177] R.Y. Tsai and T.S. Huang. Uniqueness and Estimation of Three-Dimensional Motion Parameters of Rigid Objects with Curved Surfaces. *IEEE Transactions on PAMI*, PAMI-6(1):13--26, January 1984.

[178] G. Tziritas. Recursive and/or Iterative Estimation of Two-Dimensional Velocity Field and Reconstruction of Three-Dimensional Motion. *Signal Processing*, 16:53--72, January 1989.

[179] G. Tziritas and C. Labit. *Motion Analysis for Image Sequence Coding*. Elsevier, New York, 1994.

[180] S. Ullman. The Interpretation of Structure from Motion. In *Royal Society, London, B*, volume 203, pages 405--426, 1979.

[181] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, Cambridge, MA, 1979.

[182] S. Uras, F. Girosi, and V. Torre. A Computational Approach to Motion Perception. *Biological Cybernetics*, 60:79--87, 1988.

[183] A.V. van den Berg and E. Brenner. Humans Combine the Optic Flow with Static Depth Cues for Robust Perception of Heading. *Vision Research*, 34(16):2153--2167, 1994.

[184] A. Verri and T. Poggio. Motion Field and Optical Flow: Qualitative Properties. *IEEE Transactions on PAMI*, PAMI-11(5):490--498, May 1989.

[185] R.C. Wallace, P.W. Ong, B.B. Bederson, and L.E. Schwartz. Space Variant Image Processing. *International Journal of Computer Vision*, 13(1):71--90, 1994.

[186] D. Walters. Selection of Image Primitives for General Purpose Visual Processing. *Computer Vision, Graphics and Image Processing*, 37:261--298, 1987.

[187] J.Y.A. Wang and E.H. Adelson. Representing Moving Images with Layers. *IEEE Transactions on Image Processing*, 3(5):625--638, September 1994.

[188] W. Wang and J. H. Duncan. Recovering the three-dimensional motion and structure of multiple moving objects from binocular image flows. *Computer Vision and Image Understanding*, 63(3):430--440, May 1996.

[189] E. Warrington and T. Shallice. Category Specific Semantic Impairments. *Brain*, 107:829--854, 1984.

[190] A.B. Watson and A.J. Ahumada. Model of Human Visual Motion Sensing. *Journal of the Optical Society of America A*, 2:322--342, 1985.

[191] A.M. Waxman and J.H. Duncan. Binocular Image Flows: Steps Toward Stereo-Motion Fusion. *IEEE Transactions on PAMI*, PAMI-8(6):715--729, November 1986.

[192] A.M. Waxman, B. Kamgar-Parsi, and M. Subbarao. Closed-form Solutions to Image Flow Equations for 3D Structure and Motion. *International Journal of Computer Vision*, 1:239--258, 1987.

[193] A.M. Waxman and K. Wohn. Contour Evolution, Neighborhood Deformation and Global Image Flow. *International Journal of Robotics Research*, 4:95--108, 1985.

[194] C.C. Weems. Architectural Requirements of Image Understanding with Respect to Parallel Processing. *Proceedings of the IEEE*, 79(4):537--547, April 1991.

[195] C. Weiman. Tracking Algorithms Using Log-Polar Mapped Image Coordinates. In *Intelligent Robots and Computer Vision III: Algoorithms and Techniques*, pages 843--853, Philadelphia, Pensylvania, November 1989. SPIE.

[196] C.F.R. Weiman. Polar Exponential Sensor Arrays Unify Iconic and Hough Space Representation. In *SPIE, Intel. Robots and Computer Vision VIII: algorithms and Tech.*, volume 1192, pages 832--842, 1989.

[197] J. Weng, N. Ahuja, and T.S. Huang. Optimal Motion and Structure Estimation. *IEEE Transactions on PAMI*, 15(9):864--884, September 1993.

[198] J. Weng, T.S. Huang, and N. Ahuja. Motion and Structure from Two Perspective Views:Algorithms, Error Analysis, and Error Estimation. *IEEE Transactions on PAMI*, 11(5):451--476, May 1989.

[199] S.D. Whitehead and D.H. Ballard. Learning to Perceive and Act by Trial and Error. *Machine Learning*, 7:45--83, 1991.

[200] R. Whitman. Visual Space Perception. In Carterette EC and Friedman MP, editors, *Handbook of Perception*, volume 5, pages 351--386. Academic Press, 1975.

[201] L.E. Wixson. Exploiting World Structure to Efficiently Search for Objects. Technical Report TR 434, University of Rochester, Dept. of Computer Science, 1992.

[202] K.Y. Wohn, J. Wu, and R.W. Brockett. A Contour-Based Recovery of Image Flow: Iterative Transformation Method. *IEEE Transactions on PAMI*, 13(8):746--760, 1991.

[203] R.Y. Wong and E.L. Hall. Sequential Hierarchical Scene Matching. *IEEE Transactions on Computers*, 27:359--366, 1978.

[204] Y. Yeshurun and E. L. Schwartz. Cepstral Filtering on a Columnar Image Architecture: A Fast Algorithm for Binocular Stereo Segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(7):759--767, July 1989.

[205] A. Zeki. The Visual Image in Mind and Brain. *Scientific American*, 267(3):69--76, September 1992.

[206] Z. Zhang. Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting. Technical Report 2676, INRIA, October 1995.

[207] X. Zhuang, T. Wang, and P. Zhang. A Highly Robust Estimator through Partially Likelihood Function Modelling and its Application in Computer Vision. *IEEE Transactions on PAMI*, 14:19--35, 1992.