UNIVERSITY OF CRETE
DEPARTMENT OF COMPUTER SCIENCE

Navigation of Autonomous Robotic Systems Based on the Analysis of Visual Motion

Ph.D Thesis

Manolis I.A. Lourakis

Heraklion, December 1998

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

Πλοήγηση Αυτόνομων Ρομποτικών Συστημάτων με Βάση την Ανάλυση Κίνησης

Διατριβή που υποβλήθηκε από τον

Μανόλη Ι.Α. Λουράκη

ως μερική απαίτηση για την απόκτηση του

ΔΙΔΑΚΤΟΡΙΚΟΥ ΔΙΠΛΩΜΑΤΟΣ

Ηράκλειο, Δεκέμβριος 1998

Συγγραφέας:

Τμήμα Επιστήμης Υπολογιστών

Επταμελής Εξεταστική Επιτροπή:

Στέλιος Ορφανουδάκης, Καθηγητής, Επόπτης

Jan-Olof Eklundh, Καθηγητής, Μέλος

Στέφανος Κόλλιας, Καθηγητής, Μέλος

Γιάννης Πίτας, Καθηγητής, Μέλος

Giulio Sandini, Αναπληρωτής Καθηγητής, Μέλος

Γιώργος Τζιρίτας, Αναπληρωτής Καθηγητής, Μέλος

Πάνος Τραχανιάς, Αναπληρωτής Καθηγητής, Μέλος

Δεκτή:

Πάνος Κωνσταντόπουλος, Καθηγητής

Πρόεδρος Επιτροπής Μεταπτυχιακών Σπουδών

Ευχαριστίες

Κατ'αρχήν θέλω να ευχαριστήσω τον πατέρα μου Γιάννη και την αδελφή μου Ευτυχία για την αγάπη, την εμπιστοσύνη τους και την υποστήριξη που μου παρείχαν σε όλες τις δύσκολες στιγμές κατά τη διάρκεια εκπόνησης της εργασίας αυτής. Θερμά επίσης ευχαριστώ την Φλώρα για την κατανόηση και την υπομονή της.

Η παρούσα διατριβή ολοκληρώθηκε κάτω από την εποπτεία του καθηγητή κ. Στέλιου Ορφανουδάκη, τον οποίο και ευχαριστώ για την συνεργασία του και τις ευκαιρίες που μου έδωσε.

Θα ήθελα επίσης να ευχαριστήσω τα μέλη της εξεταστικής επιτροπής της διατριβής μου, τους κυρίους Jan-Olof Eklundh (καθηγητή του Τμήματος Αριθμητικής Ανάλυσης και Επιστήμης Υπολογιστών [NADA] του Βασιλικού Ινστιτούτου Τεχνολογίας της Στοκχόλμης [KTH]), Στέφανο Κόλλια (καθηγητή του Τμήματος Ηλεκτρολόγων Μηχανικών του Εθνικού Μετσόβειου Πολυτεχνείου), Γιάννη Πίτα (καθηγητή του Τμήματος Πληροφορικής του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης), Giulio Sandini (αναπληρωτή καθηγητή του Τμήματος Πληροφορικής, Συστημάτων και Επικοινωνιών [DIST] του Πανεπιστημίου της Γένοβα), Γιώργο Τζιρίτα (αναπληρωτή καθηγητή του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης) και Πάνο Τραχανιά (αναπληρωτή καθηγητή του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης). Οι παρατηρήσεις και υποδείξεις τους συνέβαλλαν στην αρτιότερη συγκρότηση του κειμένου της διατριβής.

Ιδιαίτερη αναφορά αξίζει στο φίλο και συνάδελφό μου Αντώνη Αργυρό. Αν και μακριά από το Ηράκλειο κατά τη διάρκεια των δύο τελευταίων χρόνων, η συνεργασία μαζί του στα πλαίσια της δικής του διδακτορικής διατριβής, αποτέλεσε πολύτιμο εφόδιο για τις δικιές μου προσπάθειες.

Θέλω ακόμα να ευχαριστήσω από καρδιάς τους φίλους μου, οι οποίοι μου έδω-

σαν την αμέριστη συμπαράστασή τους, υπομένοντας αδιαμαρτύρητα τις παραξενιές που μου προκαλούσε η κούραση και το άγχος.

Τελειώνοντας, θέλω να ευχαριστήσω το διοικητικό και τεχνικό προσωπικό τόσο του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης όσο και του Ινστιτούτου Πληροφορικής του Ιδρύματος Τεχνολογίας και Έρευνας για την άμεση βοήθεια που μου παρείχε όποτε την χρειάστηκα. Ιδιαίτερα ευχαριστώ το Ινστιτούτο Πληροφορικής του ΙΤΕ, η οικονομική και υλικοτεχνική συνδρομή του οποίου υπήρξε σημαντικότατη για την ολοκλήρωση της εργασίας.

Στή μνήμη της μητέρας μου
Αικατερίνης Μαρκουτσάκη - Λουράκη

Πλοήγηση Αυτόνομων Ρομποτικών Συστημάτων με Βάση την Ανάλυση Κίνησης

Μανόλης Ι.Α. Λουράκης

Διδακτορική Διατριβή

Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

Περίληψη

Τα πλήρως αυτόνομα ή ακόμα και τα τηλεχειριζόμενα ημιαυτόνομα ρομποτικά συστήματα μπορούν να έχουν πληθώρα εφαρμογών σε τομείς όπως ο βιομηχανικός και οικιακός αυτοματισμός, η εξερεύνηση του διαστήματος, η ασφάλεια και η φύλαξη χώρων, η ανάπτυξη οχημάτων για χρήση σε επικίνδυνα για τον άνθρωπο περιβάλλοντα, η υποστήριξη ατόμων με ειδικές ανάγκες, κ.λπ. Για να μπορέσουν τα συστήματα αυτά να λειτουργήσουν σε άγνωστα ή μη δομημένα και μεταβαλλόμενα περιβάλλοντα, θα πρέπει να μπορούν να αντιλαμβάνονται το περιβάλλον τους και να δρούν ανάλογα. Μια από τις πιο σημαντικές ικανότητες αντίληψης ενός αυτόνομου συστήματος είναι αυτή της πλοήγησης, η δυνατότητα δηλαδή αυτόνομης κίνησης στο περιβάλλον με βάση τις μετρήσεις που παρέχουν διάφοροι αισθητήρες. Το κύριο θέμα της παρούσας εργασίας είναι η *οπτική πλοήγηση* (visual navigation), όπου η όραση αποτελεί την βασική αίσθηση και η πλοήγηση βασίζεται στην *ανάλυση κίνησης* (visual motion analysis). Πιο συγκεκριμένα, η διατριβή αυτή ασχολείται

με την μελέτη της διδιάστατης κίνησης σημείων τα οποία παρατηρούνται σε μια επίπεδη φωτοευαίσθητη επιφάνεια προβολής (π.χ. κάμερα), με στόχο την εξαγωγή περιγραφών αναφορικά με την κίνηση της επιφάνειας αυτής ως προς το περιβάλλον καθώς και την γεωμετρία της απεικονιζόμενης σκηνής. Τέτοιες περιγραφές μπορούν να χρησιμοποιηθούν για να υποστηρίξουν την επίτευξη των στόχων κινούμενων ρομποτικών συστημάτων. Επιπλέον, εκτός από την εξαγωγή πληροφοριών που μπορούν να χρησιμοποιηθούν για την καθοδήγηση μηχανικών συστημάτων, η μελέτη της κίνησης μπορεί να συμβάλλει στην αυτοματοποίηση περίπλοκων διαδικασιών σε ερευνητικούς τομείς όπως είναι ο ευρετηριασμός βίντεο (video indexing), η σύνθεση νέων απόψεων μιας σκηνής (novel view synthesis), η εικονική και επαυξημένη πραγματικότητα (virtual and augmented reality), η επεξεργασία ταινιών βίντεο (video post production), κ.λπ.

Η προσέγγιση που υιοθετείται στην εργασία αυτή ακολουθεί την θεωρία της *τελεολογικής* (purposive) ή αλλιώς *συμπεριφορικής* (behavioral) όρασης, σύμφωνα με την οποία ένα σύστημα όρασης πρέπει να οργανώνεται με βάση τους στόχους του και τις ικανότητες που απαιτείται να έχει και όχι με βάση την λειτουργική σημασία των δομικών του στοιχείων. Κάθε ικανότητα υλοποιείται από μια ξεχωριστή διεργασία (process), η οποία έχει έναν καλά ορισμένο στόχο και είναι προσαρμοσμένη στις ιδιαιτερότητες του περιβάλλοντος για το οποίο προοριζεται το σύστημα όρασης. Έτσι, η όραση επιτυγχάνεται από ένα σύνολο συνεργαζόμενων διεργασιών, οι οποίες επιδιώκουν τους στόχους του αντίστοιχου συστήματος με ένα συνεργατικό τρόπο. Η συμπεριφορική προσέγγιση στην όραση παρέχει σημαντικά μεθοδολογικά πλεονεκτήματα. Πρώτον, η τελεολογία των διεργασιών της όρασης επιτρέπει την διατύπωση απλούστερων και άρα ευκολότερων προβλημάτων. Δεύτερον, καθένα από αυτά τα προβλήματα επιδέχεται ένα μικρό αριθμό από πιθανές λύσεις, οι οποίες μπορούν να έχουν ποιοτικό χαρακτήρα. Επομένως, αν μπορούν να βρεθούν άμεσες λύσεις σε τέτοια προβλήματα, ένα σύστημα όρασης μπορεί να λειτουργήσει με βάση μερικές (partial) αναπαραστάσεις του περιβάλλοντος. Οι αναπαραστάσεις αυτές περιορίζονται σε πληροφορίες που αφορούν τα συγκεκριμένα προβλήματα

που πρέπει να αντιμετωπιστούν, καταργώντας την ανάγκη κατασκευής μιας πλήρους, γενικού σκοπού αναπαράστασης. Το χαρακτηριστικό αυτό είναι ιδιαιτέρως σημαντικό, μια και η εξαγωγή μιάς λεπτομερούς, γενικού σκοπού αναπαράστασης είναι εξαιρετικά δύσκολη. Τέλος, η φυσιολογία ενός συστήματος σε συνδιασμό με τα χαρακτηριστικά του περιβάλλοντός του, θέτουν περιορισμούς η εκμετάλλευση των οποίων μπορεί να απλοποιήσει προβλήματα τα οποία είναι πολύ δύσκολα στην γενικότητά τους.

Η παρούσα διατριβή περιγράφει τα αποτελέσματα των ερευνητικών προσπαθειών που αφορούν τέσσερεις οπτικές ικανότητες, συγκεκριμένα την ανίχνευση ανεξάρτητης κίνησης, την εκτίμηση ιδίας κίνησης, την ανίχνευση εμποδίων και την εκτίμηση του χρόνου πρόσκρουσης. Ιδιαίτερη έμφαση δόθηκε τόσο σε θεωρητικές όσο και σε πρακτικές πτυχές των παραπάνω προβλημάτων. Αρχικά, με βάση θεωρητικές μελέτες, αναπτύχθηκαν υπολογιστικά μοντέλα για κάθε οπτική ικανότητα. Δεδομένου ότι η υπολογιστική όραση είναι κυρίως ένας εμπειρικός τομέας, το επόμενο βήμα ήταν η πειραματική επαλήθευση των υπολογιστικών μοντέλων με χρήση πρωτότυπων υλοποιήσεων κατάλληλων αλγορίθμων. Ιδιαίτερη προσοχή δόθηκε στην ανάπτυξη τεχνικών που αποφεύγουν την διατύπωση πολύ περιοριστικών υποθέσεων αναφορικά με τον παρατηρητή ή/και το περιβάλλον, είναι ανεκτικές στην ύπαρξη θορύβου και βασίζονται σε απλές αναπαραστάσεις οι οποίες δεν απαιτούν την εξαγωγή περιττών πληροφοριών. Εφόσον κάθε μια από τις αναπτυχθείσες ικανότητες ασχολείται με την επίτευξη ενός καλά ορισμένου στόχου και δεν εξαρτάται καίρια από το περιβάλλον, μπορεί να αποτελέσει μια τεχνική γενικής χρήσης, κατάλληλη για διάφορες πρακτικές εφαρμογές. Συνολικά, αυτές οι οπτικές ικανότητες αποτελούν ένα σύνολο εργαλείων, ικανό να υποστηρίξει σύνθετες συμπεριφορές. Στη συνέχεια, περιγράφονται με συντομία η συμβολή και τα αποτελέσματα αυτής της εργασίας.

Η πρώτη από τις ικανότητες που μελετήθηκαν ασχολείται με την αναγνώριση αντικειμένων τα οποία κινούνται ανεξάρτητα από έναν κινούμενο παρατηρητή

μέσα στο οπτικό του πεδίο. Οι περισσότερες από τις τεχνικές που έχουν προταθεί για την ανίχνευση ανεξάρτητης κίνησης βασίζονται σε περιοριστικές υποθέσεις σχετικά με το περιβάλλον ή την κίνηση του παρατηρητή. Επιπλέον, βασίζονται στον υπολογισμό ενός πυκνού πεδίου οπτικής ροής, το οποίο αντιστοιχεί στην επίλυση του προβλήματος της αντιστοίχισης το οποίο είναι ασθενώς ορισμένο (ill-posed). Στα πλαίσια αυτής της εργασίας, η ανίχνευση ανεξάρτητης κίνησης ανάγεται σε ένα πρόβλημα εύρωστης εκτίμησης παραμέτρων κίνησης, το οποίο εφαρμόζεται στα οπτικά ερεθίσματα που δέχεται ένας συμπαγώς κινούμενος παρατηρητής. Η προτεινόμενη μέθοδος επιλέγει αυτόματα μια επίπεδη επιφάνεια στην σκηνή και υπολογίζει το πεδίο κάθετης υπόλοιπης ροής λόγω παράλλαξης (residual parallax normal flow field) σε δύο διαδοχικές χρονικές στιγμές. Στη συνέχεια, τα δυο πέδια κάθετης ροής που προκύπτουν συνδιάζονται με ένα γραμμικό μοντέλο. Οι παράμετροι του μοντέλου αυτού σχετίζονται με τις παραμέτρους της κίνησης του παρατηρητή και η εύρωστη εκτίμησή τους παρέχει μια τμηματοποίηση της σκηνής με βάση την τρισδιάστατη κίνηση. Η μέθοδος αποφεύγει μια πλήρη λύση στο πρόβλημα της αντιστοίχισης με το να αντιστοιχεί επιλεκτικά υποσύνολα των σημείων των εικόνων και να χρησιμοποιεί πεδία κάθετης ροής. Πειραματικά αποτελέσματα δείχνουν την αποτελεσματικότητα της προτεινόμενης μεθόδου στην ανίχνευση ανεξάρτητης κίνησης σε περιπτώσεις σκηνών με μεγάλες διακυμάνσεις βάθους και γενικών κινήσεων του παρατηρητή.

Η δεύτερη ικανότητα ασχολείται με το πρόβλημα εκτίμησης της ιδίας κίνησης (της ταχύτητας δηλαδή ενός κινούμενου παρατηρητή ως προς το περιβάλλον), με χρήση οπτικής πληροφορίας. Η γνώση της ιδίας κίνησης είναι πολύ χρήσιμη για διάφορες διαδικασίες βασισμένες σε οπτική ανάδραση. Πολλές από τις υπάρχουσες τεχνικές για την επίλυση του προβλήματος αυτού βασίζονται σε περιοριστικές υποθέσεις σχετικά με την κίνηση του παρατηρητή ή τη δομή της παρατηρούμενης σκηνής. Επιπλέον, συχνά καταφεύγουν σε αναζήτηση στον πολυδιάστατο χώρο των δυνατών λύσεων. Συχνά, τέτοιες τεχνικές αναζήτησης συνεπάγονται μεγάλο υπολογιστικό κόστος ή παρουσιάζουν προβλήματα σύγκλισης στη σωστή λύση.

Στην εργασία αυτή, εξάγεται ένας νέος γραμμικός περιορισμός που περιλαμβάνει ποσότητες εξαρτώμενες απο τις παραμέτρους της ιδίας κίνησης. Ο περιορισμός αυτός ορίζεται μέσω των διανυσμάτων οπτικής ροής που αντιστοιχούν σε τετράδες συνευθειακών σημείων των εικόνων και είναι εφαρμόσιμος ανεξάρτητα από το είδος της ιδίας κίνησης ή τη δομή της σκηνής. Επιπλέον, είναι ακριβής υπό την έννοια ότι δεν εξάγεται με χρήση κάποιων προσεγγίσεων. Σε συνδιασμό με τεχνικές εύρωστης εκτίμησης παραμέτρων, ο περιορισμός αυτός επιτρέπει την εκτίμηση της κατεύθυνσης της μεταφορικής κίνησης (δηλ. του FOE), διαχωρίζοντας έτσι τις μεταφορικές από τις περιστροφικές συνιστώσες της ιδίας κίνησης. Εκτενείς προσομειώσεις καθώς και πειράματα με πραγματικά πεδία οπτικής ροής, δείχνουν την ακρίβεια της μεθόδου με διάφορα επίπεδα θορύβου και κινήσεις του παρατηρητή.

Η ανίχνευση και αποφυγή εμποδίων αποτελούν δεξιότητες απαραίτητες για την ασφαλή μετακίνηση ενός αυτόνομου συστήματος στο περιβάλλον. Η τρίτη ικανότητα που μελετήθηκε επιτρέπει σε ένα κινούμενο ρομπότ να εντοπίσει εμπόδια στο οπτικό του πεδίο χρησιμοποιώντας δύο εικόνες του περιβάλλοντα χώρου. Η μέθοδος ταξινομεί τα σημεία μιάς εικόνας σε δύο κατηγορίες, χαρακτηρίζοντάς τα είτε σαν εμπόδια είτε σαν ελεύθερο χώρο. Υποθέτοντας ότι το σύστημα κινείται πάνω σε μια τοπικά επίπεδη επιφάνεια, η μέθοδος χρησιμοποιεί ένα σύνολο από σημεία τα οποία έχουν αντιστοιχιστεί μεταξύ των δύο όψεων, για να εκτιμήσει την ομογραφία (homography) που ορίζεται από το επίπεδο του πατώματος. Με βάση την ομογραφία αυτή, είναι δυνατή η αναίρεση της κίνησης του πατώματος μεταξύ των δύο εικόνων και στη συνέχεια η ανίχνευση εμποδίων στις περιοχές των εικόνων που παραμένουν κινούμενες μετά και την αναίρεση της κίνησης. Η μέθοδος που προκύπτει δεν απαιτεί βαθμονόμηση (calibration) της κάμερας, είναι εφαρμόσιμη τόσο σε στερεοσκοπικά ζεύγη όσο και σε ακολουθίες εικόνων, δεν χρησιμοποιεί πυκνά πεδία ταχυτήτων και παρακάμπτει το πρόβλημα της τρισδιάστατης ανακατασκευής της σκηνής. Πειραματικά αποτελέσματα από την εφαρμογή της μεθόδου σε πραγματικές εικόνες αποδεικνύουν την ευρωστία και την αποτελεσματικότητά της.

Η τέταρτη και τελευταία ικανότητα είναι συμπληρωματική στην ικανότητα ανίχνευσης εμποδίων. Πιο συγκεκριμένα, σχετίζεται με μια μέθοδο εκτίμησης του χρόνου πρόσκρουσης, του χρονικού διαστήματος δηλαδή που απομένει μέχρις ότου ένας κινούμενος παρατηρητής συγκρουστεί με αντικείμενα στο οπτικό του πεδίο. Ο χρόνος πρόσκρουσης παρέχει ένα μέτρο εκτίμησης της εγγύτητας των εμποδίων, το οποίο είναι χρήσιμο για την αποφυγή συγκρούσεων. Εδώ παρουσιάζεται μια νέα μέθοδος για την εκτίμηση του χρόνου πρόσκρουσης, η οποία βασίζεται στην υπόθεση ότι ο παρατηρητής κινείται πάνω σε μια επίπεδη επιφάνεια και κάνει χρήση του πεδίου οπτικής ροής που προκαλείται από την κίνηση αυτή. Αρχικά υπολογίζεται ο χρόνος πρόσκρουσης με σημεία του πατώματος και στη συνέχεια το φαινόμενο της παράλλαξης λόγω επιπέδου (planar parallax) επιτρέπει την εκτίμηση του χρόνου πρόσκρουσης με τα σημεία που αντιστοιχούν σε εμπόδια. Η μέθοδος αποφεύγει τον υπολογισμό παραγώγων υψηλής τάξης του πεδίου οπτικής ροής, μια και είναι γνωστό ότι ο υπολογισμός αυτός είναι ευαίσθητος στο θόρυβο. Επιπλέον, δεν απαιτείται καμία γνώση της ιδίας κίνησης. Παρατίθενται τέλος πειραματικά αποτελέσματα από την εφαρμογή της μεθόδου σε πραγματικά και συνθετικά πεδία ταχυτήτων.

Navigation of Autonomous Robotic Systems Based on the Analysis of Visual Motion

Manolis I.A. Lourakis

Doctoral Dissertation

Department of Computer Science
University of Crete

Abstract

Fully autonomous or even teleoperated, semi-autonomous robotic systems can have numerous practical applications in fields such as industrial automation, space exploration, home automation, space monitoring and security, automatic guided vehicles (AGV's) for use in hazardous environments, support of people with special needs, etc. In order to be able to function in unknown or unstructured and changing environments, such systems should possess effective perceptual capabilities for sensing their surroundings and acting accordingly. One of the most important perceptual capabilities of an autonomous system is that of navigation, that is the capability of autonomous motion in the environment, based on the measurements provided by various sensors. The main theme of this work is *visual navigation*, where vision is the primary sensing modality and navigation is based on the analysis of *visual motion*. More specifically, this dissertation is concerned with the interpretation of the 2D motion of points observed on a planar light-sensitive projection surface, in order to derive descriptions of the surface's self

motion and the geometry of the imaged scene. Such descriptions are intended to support autonomous robots for achieving their goals. However, apart from producing results that can be used for guiding mechanical systems, the study of visual motion is beneficial to research areas such as video browsing and indexing, novel view synthesis, virtual and augmented reality, animation, video post-production, etc, helping to automate tasks that are either tedious or too complicated to perform manually.

The research approach adopted here follows the *purposive* or *behavioral* vision paradigm, according to which a vision system should be organized on the basis of the different capabilities that it should possess and not according to the functional role of each component. Each capability is implemented by a separate process, having a distinct, well-defined goal and being tailored to the environment the vision system is expected to operate in. Thus, vision is realized by a set of cooperating processes, which pursue the system's goals in a synergistic manner. There are important methodological reasons in favor of the behavioral approach to vision. First, the purposiveness of the visual processes permits the formulation of simpler, therefore easier problems. Second, each of these problems can admit a small number of possible answers that can have a qualitative nature. Thus, if direct solutions to such problems can be found, a vision system can operate on the basis of partial environment representations that capture only those aspects relevant to particular tasks, hence alleviating the need for a complete, general purpose representation. This is of particular importance, since the construction of a detailed, general purpose representation is extremely difficult. Finally, the physiology of a system along with its environment, impose constraints that when exploited, can simplify problems which are very difficult in their general form.

This dissertation describes the results of investigations regarding four visual capabilities, namely independent motion detection, egomotion estimation, obstacle detection and time-to-contact estimation. Emphasis has been given both on theoretical and practical aspects of the above problems. First, a computational model regarding each visual capability has been derived from theoretical study. Since computer vision

is primarily an experimental discipline, the next step was to demonstrate the validity of relevant models through experiments involving prototype implementations of proper algorithms. Special care has been taken to develop techniques that avoid making restrictive assumptions which limit their applicability, are robust to noise and rely on simple representations that do not require the recovery of redundant information. Since each of the developed capabilities deals with a well defined goal and does not depend critically on the environment, it can function as a generic navigational tool for various practical applications. Collectively, these visual capabilities constitute a solid arsenal of primitive algorithms that is able to support complex behavioral repertoires. The specific contributions of this work are briefly discussed below.

The first of the capabilities studied deals with the identification of objects that move independently of a mobile observer within his field of view. Most of the existing techniques for detecting independent motion rely on restrictive assumptions about the environment, the observer's motion, or both. Moreover, they are based on the computation of a dense optical flow field, which amounts to solving the ill-posed correspondence problem. In this work, independent motion detection is formulated as a problem of robust parameter estimation applied to the visual input acquired by a rigidly moving observer. The proposed method automatically selects a planar surface in the scene and the *residual planar parallax* normal flow field with respect to the motion of this surface is computed at two successive time instances. The two resulting normal flow fields are then combined in a linear model. The parameters of this model are related to the parameters of egomotion and their robust estimation leads to a segmentation of the scene based on 3D motion. The method avoids a complete solution to the correspondence problem by selectively matching subsets of image points and by employing normal flow fields. Experimental results demonstrate the effectiveness of the proposed method in detecting independent motion in scenes with large depth variations and unrestricted observer motion.

The second capability is concerned with the problem of using visual input to

estimate egomotion, i.e. the velocity of a mobile system with respect to its environment. Knowledge of the egomotion is essential for various servoing tasks that are based on visual feedback. Many of the existing techniques for solving this problem rely on restrictive assumptions regarding the observer's motion or even the scene structure. Moreover, they often resort to searching the high dimensional space of possible solutions, which might be inefficient and exhibit convergence problems. In this work, a novel linear constraint that involves quantities that depend on the egomotion parameters is derived. This constraint is defined in terms of the optical flow vectors pertaining to four collinear image points and is applicable regardless of the egomotion or the scene structure. In addition, it is exact in the sense that no approximations are made for deriving it. Combined with robust linear regression techniques, the proposed constraint enables the recovery of the direction of translation (e.g. the FOE), thereby decoupling the 3D motion parameters. Extensive simulations as well as experiments with real optical flow fields provide evidence regarding the performance of the proposed method under varying noise levels and camera motions.

Obstacle detection and avoidance capabilities are essential to an autonomous robot for it to move safely in the environment. The third capability refers to a method that enables a mobile robot to locate obstacles in its field of view using two images of its surroundings. The method provides a binary labeling of image points, classifying them either as obstacles or as free space. Assuming that the robot is moving on a locally planar ground, the method uses a set of reference point features (corners), that have been matched between the two views, to compute a robust estimate of the homography of the ground. Knowledge of this homography permits us to compensate for the motion of the ground and to detect obstacles as areas in the image that appear nonstationary after the motion compensation. The resulting method does not require camera calibration, is applicable either to stereo pairs or to image motion sequences, does not rely on a dense disparity/flow field and circumvents the 3D reconstruction problem. Experimental results from the application of the method on real images indicate that it is both effective and robust.

The fourth and last capability is complementary to the obstacle detection capability. In particular, it involves a method for estimating the *time-to-contact*, i.e. the amount of time that remains before a mobile observer collides with objects in his field of view. The time-to-contact provides a measure for assessing the proximity of obstacles that is well suited to avoiding collisions. Here, a novel method for estimating the time-to-contact is presented. The method is based on the assumption that the robot is moving on a locally planar ground and employs the optical flow field induced by the robot's motion. First, the time-to-contact with points on the ground is estimated and then the concept of *planar parallax* is exploited to recover the time-to-contact with obstacle points. The computation of high order derivatives of image flow, which are known to be very difficult to estimate accurately, is completely avoided. Moreover, no knowledge of the egomotion is necessary. Experimental results from the application of the method on real and simulated flow fields are also reported.

Contents

List of Figures

Part I

Background

Chapter 1

Introduction

The comprehension of the principles and mechanisms of visual perception has been a long sought goal for science. As such, it has challenged many researchers from diverse disciplines ranging from philosophy, physiology, psychology, ethology and neurobiology to mathematics and engineering. The motivation behind their efforts has been twofold. On one hand, it relates to man's inquiry into his own nature and the understanding of consciousness. On the other hand, it has to do with the practical applications that can be devised for artificial vision systems. This second aspect has been reemphasized by the advent of digital hardware, which provides cheap and powerful computational engines for use in building intelligent robotic systems [126, 127, 70]. Based on representations of the environment delivered by artificial vision systems, autonomous robots have the potential of perceiving their surroundings and adapting their behavior to unforeseen changes. Vision is particularly attractive for this purpose, since in principle it does not require any modifications of the environment and is passive in the sense that it does not rely on the emission of any kind of signal. Thus, many theories, computational paradigms and models have been proposed in order to explain various aspects of vision [156, 107, 174, 121, 258]. The study of vision, however, is no simple matter. The stimuli reaching a light sensitive sensor encode a tremendous amount of information that is constantly changing with time. Using this information,

visible objects should be located, categorized and characterized in a timely manner. In other words, vision is responsible for relating a varying 2D light pattern to object descriptions that facilitate appropriate actions. In the effort toward understanding vision, the study of working solutions that nature has invented for biological vision systems can prove to be fruitful. Although artificial seeing systems should not necessarily imitate biological ones, a close look at the functioning of the latter provides invaluable insight and inspiration [103]. Nevertheless, despite the fact that vision appears to be effortless to living organisms, researchers in the field have realized that the processing of visual information is extremely difficult to perform computationally. The situation is further complicated by the fact that, apart from visual stimuli, biological vision also depends upon a priori knowledge and context specific information, both of which are hard to reveal and describe in detail. Consequently, computational issues have been the primary subject of research in vision. Although the results of relevant research during the last decades are by no means negligible, much work is still required both at the theoretical and the practical level.

This thesis is concerned with the interpretation of the 2D motion of points observed on a projection surface (retina, film or CCD array), in order to derive descriptions of the surface's 3D motion (*egomotion*) and the geometry of the imaged scene. Such descriptions are intended to be used by autonomous robotic systems for sensing their environment and acting accordingly. In the remainder of this chapter, the prevalent theories of computational vision are presented and discussed. Based on biological findings, the significance of the image motion cue to visual perception is then emphasized. Later sections discuss various aspects of the problem of visual navigation, which is the main theme of the work described in this thesis.

1.1 Theories of Computational Vision

Considering the vision problem primarily as a problem of deriving symbolic information from images, has been one of the major early developments in the area of computational vision [107]. Eyes focus images on the retina which through the optical nerve transmits stimuli to the brain where they are processed my specialized neurons. The results of this process comprise the symbolic descriptions of objects and their properties, which are used by the brain to determine the interaction of an organism with its environment.

Such a description, however, does not supply answers to vital questions such as which is the information encoded in images that is essential for perceiving the environment, how is this information extracted and how it is represented at the symbolic level. Such questions are crucial for the study of vision, since they define the problems that have to be solved in order to fully understand vision and at the same time to be able to construct useful artificial vision systems. During the sixties and seventies, researchers in the field strived to resolve these issues, facing serious difficulties when attempting the practical application of their theories. Therefore, the common practice was to formulate oversimplifying assumptions, which lead to vision systems that worked only in carefully controlled situations and did not contribute to the more general goal of understanding vision.

1.1.1 Marr's vision theory

David Marr was one of the first researchers to recognize that the main reason for the failure of the early efforts in the field of computational vision was the lack of a complete vision theory. Hence, he attempted to develop a theory for explaining vision, that could also be exploited to build practical vision systems [156]. More specifically, he proposed that most tasks that employ vision rely upon the solution of the following problem: Given a set of images depicting the same scene, the goal is to recover an accurate, general

purpose, 3D geometric description of the scene and a quantitative description of the properties of visible objects. Tasks such as obstacle avoidance, recognition, grasping, etc, are carried out by appropriate symbolic processing of the data provided by such a representation. In other words, Marr's theory calls for the reconstruction of a complete, quantitative representation of the environment, which is capable of supporting any task that relies upon the use of visual information. Marr concluded that the role of vision is to capture every detail of the environment in the above representation. Marr's theory reflects the established beliefs of the neurobiologists of his time, which claimed that the neurons assigned to visual processing analyze in detail all the information that is available to them. It has also been influenced by philosophical ideas that were prominent at the time. The roots of these ideas can be traced to Kant's theories, who advocated that the processes of ''seeing'' and ''thinking'' are separate [133].

Based on the principles of information theory, Marr attempted to systematize the study of vision by focusing at three levels:

- **The computational theory level**
  Based on a detailed mathematical analysis, the relation between the quantity that is to be computed and the data provided by the vision sensors, is derived. This analysis determines the cases in which the problem can be solved and whether it has a unique solution.

- **The level of algorithms and representations**
  Based on the derived computational theory, appropriate algorithms and data structures are designed. At this level, issues related to the efficiency and robustness of algorithms are addressed.

- **The implementation level**
  The algorithms that were developed at the previous step are implemented on serial or parallel hardware.

Marr's theory proved to be very influential to the study of vision and of artificial

intelligence in general. Most scientists were convinced that vision was responsible just for the first component of the sense - think - act cycle, whereas other subfields of artificial intelligence should deal with the remaining two. Following standard techniques for designing complex systems, scientists dealt with the task of reconstructing the environment by breaking it into functionally separate subsystems that could then be studied separately. After studying the various subsystems in detail, they could be implemented and integrated into a working practical system. Hence, research in the various subfields of artificial intelligence progressed separately, with each subfield ignoring the weaknesses of the remaining ones, but expecting from them results that were far beyond those that could actually be obtained. For instance, researchers working in motion planning expected that vision would provide them with accurate models of the environment, those working in vision assumed that machine learning would provide them with effective means for dealing with noise and uncertainty, etc. Until now, such expectations have not been realized, resulting in the inability of understanding the components comprising the phenomenon of intelligence. The vast majority of the studies that have appeared during the last 25 years are affected by Marr's theory, and despite their elegant theoretical contributions, they cannot be applied to yield useful artificial vision systems.

The weaknesses of these studies can be attributed to various reasons: First, the process of fully reconstructing the environment has enormous computational requirements. To appreciate this, it should be noted that more than half of the neurons of the human brain are devoted to the processing of visual information and yet there is strong evidence that the human visual system is confined to recovering partial representations of the environment. Moreover, even in the case that a detailed representation was indeed available, the computational burden for exploiting it would be prohibitive. Second, although Marr's theory addressed some aspects of perception (stimuli, sensors and processing of sensor data), it completely ignored others, such as the environment in which a vision system operates, the physiology of its body and the goals that it has to achieve [80]. Third, the ultimate goals of most research efforts have

been very ambitious, aiming to develop systems with advanced capabilities. Nature, however, provides many examples of biological vision systems that are very simple in terms of their capabilities but are nevertheless sufficient for accomplishing the goals of the organisms possessing them. Fourth, modeling the vision process as a closed system in order to study it with the aid of information theory, is unrealistic. Since it is impossible to describe the environment in detail, the assumption of a closed system is violated. Finally, many of the procedures for reconstructing the environment are very sensitive to the accuracy of their input [262]. Therefore, they are unstable and thus inapplicable in the presence of noise, uncertainty and missing data.

1.1.2 The purposive or behavioral vision paradigm

In the mid-eighties, a few researchers began to question the validity of Marr's theory by pointing out its weaknesses and putting forward alternative theories (see, for example, the theory proposed by Feldman in [77]). At about the same time, researchers from the fields of phychophysics and neurobiology [13] concluded that vision, as every other sense, is adapted to the environment where an organism lives, and is designed to serve the achievement of specific goals, while taking into account the organism's physiology and overall capabilities. Consequently, vision could be studied more effectively in terms of the behaviors that an organism is expected to exhibit. This approach was also found appealing by researchers in the field of general artificial intelligence, who questioned the validity of the traditional methods and focused on what has been termed ''*autonomous agent research*'' [282, 41, 155]. Based on these premises, a new vision theory has emerged, that is known by the names *active*, *behavioral* or *animate* vision [4, 23, 6, 189, 24, 241, 245]. According to this theory, a vision system should be constantly controlling the parameters of its optical apparatus, so as to acquire images that facilitate the tasks that it has to accomplish. In addition, a vision system should be decomposed on the basis of the different behaviors that it should support and not according to the functional role of each component (e.g. structure from motion, structure from shading,

etc). Each behavior is implemented by a separate process, having a distinct, well-defined goal and being tailored to the environment the vision system is expected to operate in. Vision is realized by a set of cooperating such processes, which achieve the system's goals in a synergistic manner. Thus, the role of vision is to transform visual stimuli to behaviors, i.e. sequences of actions, for interacting with the environment. In other words, the emphasis is put on the development of behaviors rather than the delivery of a representation suitable for visual tasks. Clearly, this is in complete contrast to Marr's theory, which separates vision from action.

Therefore, the key question that should be answered, before any study of vision, is related to the goals that the latter is required to support. An artificial vision system for example, may need some of the capabilities possessed by certain living organisms and lack others. This is because every seeing system, be it an artificial or a biological one, has a special relationship with its environment and its intelligence is an embodiment of that relationship. There is no self-existent intelligence, that is intelligence without goals to pursue and a body to interact with the environment. Each system has a set of visual capabilities that are realized by a set of specialized processes that encode the system's intelligence and make use of different, partial representations of the environment. There is no centralized control nor a single, coherent representation. This point of view is consistent with the theory of evolution [80]: It is very probable that the species which developed light sensitive sensors did not develop their visual capabilities at the same time, but through different time periods according to their needs. Hence, the existence of many different processes that perform different duties and are implemented by different groups of neurons should not come as a surprise. The existence of such processes in biological vision systems has been proven experimentally. For example, experiments over long periods of time have led Zeki to the conclusion that distinct areas in the cortex analyze different attributes such as color, shape and motion [288]. The functional specialization of the brain has been demonstrated based on clinical evidence of certain lesions causing the malfunction of specific visual capabilities, but not depriving the patient of his vision.

The purposiveness of the visual processes permits the formulation of simpler, therefore easier problems. Each of these problems can admit a small number of possible answers that can have a qualitative nature [247]. Thus, if direct solutions to such problems can be found, a general purpose representation of the environment becomes unnecessary. The physiology of a system along with its environment, impose constraints that enable the solution of problems which are in general very difficult. Therefore, the aim is to provide solutions that supply answers to specific problems under general assumptions. In contrast, the solutions advocated by the reconstructionist paradigm favor general solutions under very restrictive assumptions.

Apart from inhabiting a specific environment, an intelligent system is also situated in time. This enables it to develop itself so as to improve its capabilities by learning from experience. An intelligent system can benefit from unsupervised machine learning techniques in various ways. For example, learning can help the system cope with uncertainty and unpredictability. Learning can also be employed for deducing the mapping of sensor information to appropriate motor commands. Tedious preprogramming of all possible situations that the system might find itself in can be avoided by using learning to infer the control laws for motor control. Time also allows for the construction of incremental solutions to specific problems instead of the provision of complete solutions in one step. These solutions evolve as they are implemented by the execution of proper actions, resulting in a tight coupling between perception and action. This is more tolerant to outdated or faulty representations and therefore yields more flexible and robust intelligent systems. A related consequence is that the dynamics of interaction between the system and its environment can lead to emergent complexity [3, 39]. In other words, a simple, reflex-like interaction of a system with its environment can result in a sophisticated behavior imposed by the complexity of the latter.

A multiprocess system has serious computational advantages over a monolithic, centralized one. At the system level, a decentralized set of cooperating processes is easier to distribute to different processing modules, is more fault tolerant and can adapt more

easily to changing goals and interrupts. The lack of a unique, common representation permits the processes to communicate by exchanging short, inexpensive messages. At the process level, each process saves computational resources by avoiding irrelevant computations and using a limited number of information processing layers.

1.2 The Role of Visual Motion Processing in Perception

In this section, neurobiological evidence is used to emphasize the central role played by motion understanding in the way biological organisms perceive their environment.

Consider a seeing mobile observer who is moving relative to his environment. Assuming that the observer is continuously acquiring images, objects in the environment are projected on his retina and these projections appear to be moving in a manner that is dependent on the object's relative motion with respect to the observer and the geometry of the viewed scene [144]. The branch of computational vision that deals with the processing of time varying imagery is known as *motion analysis.* As demonstrated by a plethora of experiments studying the phychophysics and the physiology of motion, image motion analysis is a fundamental property of biological visual systems [173]. Color perception and stereopsis are also important visual cues, yet it is also clear that color processing is not present in all species and that binocular vision is restricted to animals with laterally placed eyes. Furthermore, stereo processing can be considered as a special case of motion analysis, since the underlying principles are the same. Although numerous animals lack color vision or significant binocular vision or both, none have been shown to lack mechanisms for motion processing. Horridge [110], a neurobiologist working on vision, argues that insects base their perception of the world on the detection of relative motion between objects and their immediate background. In doing so, insects employ brains with limited computational capabilities compared to those available to species that stand higher in the evolution pyramid. According to the well established belief that the principles underlying vision are shared among all living organisms, the

implication of Horridge's argument is that motion perception is of utmost importance to all biological organisms possessing the sense of vision. All living organisms exist in space and time and, since most of their objects of interest are moving, they should be able to interpret visual motion. The analysis of visual motion helps an organism maintain continuity of its perception of the constantly changing environment around it. It should also be noted that motion is a *precategorical* cue to visual perception, i.e. it is not necessary to recognize an object before its motion can be analyzed. Hildreth and Koch [102] provide an extensive review of a number of aspects of visual motion analysis in biological systems from a computational perspective.

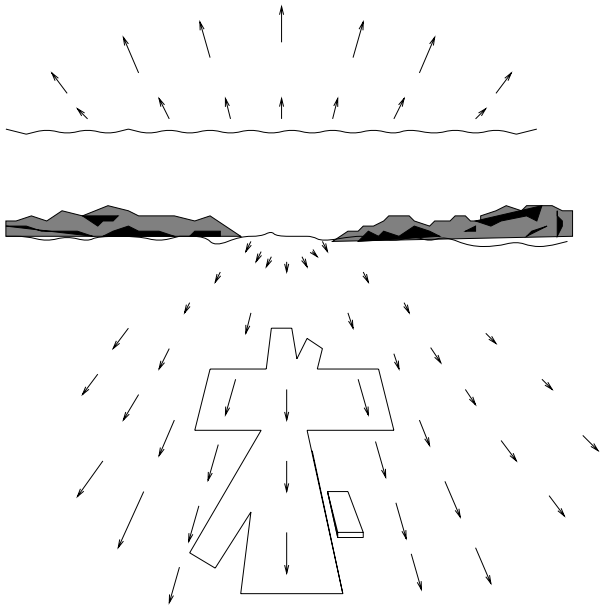


Figure 1.1: The optical flow field perceived by a pilot in level flight.

The observation that a sequence of images can be assigned a vector field that describes the velocity of the visual motion of each image point and contains crucial information regarding the environment, was put forward by Gibson [89]. Today, this field for which Gibson coined the term *optical flow*, is generally accepted as one of the most vital contributions to the research area of motion analysis. Figure 1.1 depicts the optical flow field perceived by a pilot who is looking straight ahead in a level flight. More details regarding the optical flow field can be found in chapter 2. Nakayama has identified the following different roles that are played in vision by optical flow [173]:

- **Encoding of the third dimension**

  Retinal images are inherently two dimensional and yet the visual system is capable of automatically extracting information about depth, i.e. the third dimension of space that is lost by the projection of the 3D world on the retina. Velocity fields that are extracted from moving images contain rich information regarding the slant of surfaces and the relative depth of surfaces from the observer.

- **Time-to-contact**

  Apart from providing information regarding the relative distances to environmental points, velocity fields also supply distance information in the form of time-to-contact. In other words, the time that remains before the observer collides with points in his field of view can be recovered from image flow, without any knowledge of absolute depths or relative velocity.

- **Image segmentation**

  Also known as ''figure-ground segregation'', this refers to the problem of partitioning a viewed scene into a set of different physical objects. Image points that belong to the same object move with similar retinal velocities and object boundaries give rise to motion discontinuities which form a strong cue for perceptual grouping.

- **Kinetic stabilization**

  Visual motion is one of the primary proprioceptive sources of information enabling a moving observer to estimate his own motion with respect to the environment.

- **Occulomotor control stimulus**

  Image motion provides velocity signals that drive the occulomotor pursuit system, through which an object of interest can be tracked through a scene.

- **Detection of moving objects**

  Visual motion analysis enables an observer to sense the independent motion of environmental objects. Such independently moving objects may be targets on which the observer's attention should be focused.

The existence of the above very diverse functions suggests that several motion analysis systems might exist simultaneously in living organisms [173]. Evidently, the analysis of visual motion can be employed by an artificial vision system for the purpose of autonomous movement in a dynamic environment. This is the task of vision-based navigation, which is examined in more detail in the following sections.

1.3 Vision Based Navigation

One of the most important capabilities of an intelligent system is that of navigation, that is the capability of autonomous motion in the environment, based on the measurements provided by various sensors [39, 126]. This definition is broad enough to include a plethora of biological and man-made systems. For instance, almost all living organisms have capabilities enabling them to search for prey, to avoid predators and to find their way to their nest (*homing*). To accomplish such tasks, nature has devised various navigation mechanisms and navigation sensors. Ants and bees, for example, employ inertial cues, bats use sonars, pigeons sense the earth's magnetic field and dogs are guided by smell and/or sound [49, 53, 217, 54, 111, 70]. Among the various sensing modalities, vision is the most widely used, complemented by other senses. Apart from the various biological organisms, there are a few man-made systems possessing some autonomous navigation capabilities. Planes and ships combine inertial information with beacons whose position is accurately known, military missiles are attracted by sources of electromagnetic radiation and spaceships navigate using the position of planets. It should be noted, however, that the man-made systems described above require limited navigational competences. This is due to the fact that they are designed to operate in simpler and more predictable environments compared to those of most biological organisms.

In order to support a variety of complex navigational tasks in fields like industrial automation, space exploration, home robotics, space monitoring and security, support

of people with special needs, etc, an autonomous system should possess advanced perceptual capabilities that will enable it to recognize the aspects of reality that are essential to achieve its goals [126, 127, 70]. Since such systems are intended to function in a 3D world, they should be capable of capturing the structure of their environment. As has been argued already, vision has the potential to provide most of the information necessary to navigate in unknown, unstructured and possibly changing environments. Hence, this dissertation focuses on *visual* navigation, where vision is the primary sense and navigation is based on the analysis of sequences of images. More specifically, due to its widespread use by biological seeing systems, visual motion will be the cue on which our investigation will be based. The study of visual motion will lead to a deeper understanding of the mechanisms for visual perception and will produce results that can be used to guide mechanical systems and can also be transferred to seemingly unrelated application domains. Research and development areas, such as video browsing, video indexing, view synthesis, virtual reality, augmented reality (i.e. merging of graphics and video), animation, CAD modeling, video post-production, etc, can benefit from the application of motion analysis techniques for the development of powerful tools that will automate tasks that are either tedious or too complicated to perform manually [293, 204]. A general overview of the state of the art in vision-based navigation can be found in [69]. The next section presents an overview of the traditional approach to vision-based navigation. Next, the purposive vision paradigm is considered and a hierarchy of well defined visual capabilities are identified for further study.

1.3.1 Structure from motion

Todate, most research work on vision-based navigation has been influenced by Marr's reconstruction paradigm [156]. Consequently, vision-based navigation has been considered as the solution to a more general problem, namely that of *structure from motion (SFM)*. Given a sequence of images, SFM can be defined as the problem of recovering the 3D structure of world objects and their 3D velocities relative to the

observer. Clearly, if SFM can be successfully solved, it provides sufficient information for dealing with various navigation tasks. However, despite the enormous research effort devoted to SFM, related issues remain unresolved for all practical purposes. The main reason for this difficulty is that SFM is ill-posed, i.e. it can have multiple solutions which do not depend continuously on the input.

Photosensitive surfaces such as an animals' retina, a camera film or a CCD array share in common the fact that they are all 2D. Since the natural world is 3D, a reduction of dimensionality takes place during the process of image acquisition. In other words, the depth of world points is lost when they are projected on a 2D surface. This fact accounts for many of the difficulties involved in SFM. SFM attempts to invert the projection process by taking advantage of the fact that the motion of the photosensitive surface causes every object projected on it to appear moving in a manner dependent on its relative motion with respect to the photosensitive surface and the structure of the viewed scene.

It is customary to study SFM as a two-step process. Initially, accurate displacements of image points due to the relative motion between the observer and the scene are computed using successive image frames. This amounts to solving the so-called *correspondence* problem, which involves the identification of image points in different views that are projections of the same three dimensional world point. In case of infinitesimal motions, the image sequence is regarded as a function of two temporal and one time variables. The spatiotemporal derivatives of this function, combined with some additional assumptions, permit the estimation of the *optical flow*, represented by a vector field describing the motion of image points. When the motion between successive images is significant, a solution to the correspondence problem is obtained in the form of *disparity* maps. In this last case, instead of having a single camera moving in space, the image sequence can be captured by a multioccular system composed of multiple fixed cameras that have different 3D positions and orientations. In a second step, the 3D motion and the structure of the scene are recovered from the equations

relating the 2D image velocity to the 3D motion. Techniques that can be used to establish correspondence and to recover motion and structure information, are examined below in greater detail.

Establishing correspondence

Based on the two-step decomposition of SFM that was described above, research has progressed in two directions. A large body of work has been devoted to the development of algorithms for accurately estimating the motion of image points. In the case of optical flow computation, three different approaches can be identified: *gradient based* approaches, which assume conservation of image intensity [284, 78, 108, 170, 101, 257, 172, 266, 30, 278], *correlation or area based* approaches, which assume conservation of the local intensity distribution [284, 151, 46, 10, 9, 87], and *frequency* or *filter based* approaches [274, 1, 83, 99], which consider the problem of motion estimation in the spatiotemporal frequency domain. Unifying theories of optical flow estimation techniques can be found in [171, 228, 223] and a performance comparison of various algorithms is carried out in [28, 88]. Mitiche and Bouthemy [165] provide an excellent review of the state of the art in optical flow estimation.

Despite the fact that an enormous amount of research effort has been devoted to the problem of estimating optical flow, estimates of image velocity are notoriously error-prone. As it will be made clear by a detailed discussion in section 2.1.6, current techniques cannot adequately cope with several difficult situations that occur frequently in natural scenes. For instance, a fundamental assumption made by almost every optical flow estimation algorithm is that the brightness variations observed in an image sequence are strictly due to motion. However, this assumption is violated in the case of moving shadows, change of illumination and specular reflections. Another problem has to do with the fact that the motion of constant intensity regions is underconstrained. The lack of sufficient texture results in more than a single motion being consistent with the observed image sequence. A related shortcoming is the well-known *aperture problem*,

which arises due to the fact that, based solely on local measurements, only the component of optical flow along the gradient direction (known as *normal flow*) can be computed. To overcome this indeterminacy, regularization techniques that are based on various smoothness assumptions impose additional constraints. Unfortunately, smoothness assumptions are very often violated due to depth discontinuities, independently moving objects or non-rigid and articulated motion. Finally, when image motion is large or temporal sampling is not adequate, the phenomenon of *temporal aliasing* further complicates the estimation of motion, since in this case the assumption that motion is infinitesimal is violated.

Regarding the computation of disparities, three different approaches have been studied: 1) *area correlation* approaches [27, 240], that assume locally frontoparallel surfaces; 2) *feature matching* approaches [160, 196, 184, 91, 138, 106, 290, 216, 145, 161, 201, 147], that correspond sparse image tokens such as lines, points and curves; and *filter based* methods [125, 203], that prefilter images with banks of linear filters tuned to different orientations and scales and use their responses to describe the local structure of image patches. The problem of image matching is complicated by occlusions and changes in object appearance observed in the images. Correlation approaches suffer from sensitivity to brightness in the case of non-lambertian surfaces. More serious problems can be caused by different amounts of foreshortening in the images and surface boundaries that run through the region of the image that is used for correlation. Feature matching approaches are in general more accurate, but have the drawback of yielding sparse information. Typically, the interpolation of depth across surfaces in the scene is employed as an additional step. Dhond and Aggarwal [67] provide a detailed review of techniques for estimating disparities in the case of binocular stereo.

Recovery of 3D motion and structure from image correspondences

The second aspect research in SFM has focused on, is that of estimating the motion and structure given correspondences between image points. Initially, issues regarding the

existence and uniqueness of the solution were addressed. Ullman, in his well-known SFM theorem [265], showed that a minimum of three distinct orthographic views of four non-planar points in a rigid configuration allow the motion and structure to be completely determined. In the case of perspective projection, two views are, in principle, sufficient. Longuet-Higgins later showed that two perspective views of eight points allow SFM to be solved with linear methods [142]. Faugeras and Maybank proved that five points in two views yield ten possible solutions [74]. Early research also addressed the case of simple parametric surfaces in motion. For example, Tsai and Huang showed that given four coplanar points in two views, there exist two possible solutions for the plane normal and its 3D motion [260]. When three views are available, a unique solution can be found [261]. During the last few years, the application of tools from projective and algebraic geometry has led to significant results, relevant to SFM [167, 72]. Recent theoretical contributions to SFM include the *fundamental matrix* [153], which expresses the epipolar constraint for an image pair and the *trifocal tensor* [221], which relates corresponding points and lines in three views.

In recent years, many researchers have attempted to exploit the results of theoretical analysis to develop computer algorithms for solving SFM. An assumption commonly made by these algorithms is that 3D motion to be recovered is at least piecewise rigid, since otherwise the observed 2D motion could have resulted from any arbitrary 3D motion. The algorithms can be classified to various categories according to criteria, such as the choice of the image projection model, the use of optical flow or disparities, the use of closed form formulas or iterative methods for estimating the solution, the use of long or short image sequences, the combination of monocular imagery with stereo cues (*motion stereo*) and the use of calibrated or uncalibrated cameras. Representative examples of the various approaches can be found in [104, 275, 276, 281, 280, 264, 251, 232, 233, 273, 140, 134, 130, 129, 279, 157, 287, 164, 268, 270]. Huang and Netravali [113] provide an overview of techniques for motion and structure estimation from sparse feature correspondences.

A shortcoming of most current approaches to solving the SFM problem, revealed by experimenting with the implemented algorithms, is that they are extremely sensitive to noise [262, 232]. Thus, very few algorithms have proved to be successful in realistic scenarios. The sensitivity to noise stems from the errors in image motion measurements and the nonlinearity of the equations involved in SFM. In addition, image motion fields can be ambiguous, giving rise to more than one 3D motion: An upward movement of a camera, for example, induces an optical flow field that is locally indistinguishable from the flow generated by a tilt rotation. Consequently, quantitative studies of robustness, through statistical analyses such as those presented [232, 64, 65, 92], are of utmost importance for building practical vision systems.

1.3.2 Visual capabilities

Sensitivity to noise is not the only reason for the failure of SFM algorithms. Many of these algorithms have overambitious goals. The strategy of extracting and storing as much information as possible wastes computational resources. As dictated by the purposive vision paradigm [6, 247], vision should not be considered as a goal by itself, but should be studied in conjunction with the visual tasks the system is engaged in. This approach reduces the general SFM problem to a set of more tractable processes. The purposive paradigm can make an efficient use of the available computational resources, facilitating the supply of a continuous stream of environmental information that will enable a system to perceive unforeseen changes in the environment as they happen. Each visual process implements a specific visual capability and operates on the basis of incomplete, but robust qualitative representations, instead of relying on precise measurements of motion and structure that are susceptible to noise. For example, instead of employing a complete, quantitative representation, an obstacle avoidance process can depend on the answers to a set of simple questions. Such questions would be concerned with whether the observer is approaching an obstacle, if the observer is in a collision course with an obstacle, how much time with respect to the observer's reaction time

remains until collision, etc. These questions admit a small number of possible answers and are thus of a qualitative nature.

According to their complexity, navigational capabilities fit into a natural hierarchy. This hierarchy bears resemblance to the subsumption architecture proposed by Brooks [40]. According to the latter, intelligence is achieved by a hierarchy of concurrent processes pursuing a common goal, with processes higher in the hierarchy having more general subgoals and lower level processes inhibiting higher level ones whenever dynamic changes occur in the environment. Brooks's point of view, however, is too extreme, since it imposes a strict priority scheme among processes and refuses to use any representation of the environment, arguing that the latter is the best representation of itself [42]. Some of the most important visual capabilities are listed below in order of increasing complexity. A few of them were also mentioned in section 1.2 and are repeated here for completeness. These visual capabilities are the building blocks from which more complex behaviors can be synthesized.

- **Detection of independent motion** [14, 2, 246, 220, 182, 16, 17, 18, 146, 19]
  An intelligent observer should be able to perceive dynamic changes in his field of view, since such events indicate areas of interest where his attention should be focused and possibly maintained. One such dynamic change is due to the existence of objects that move independently of the observer. The identification of these objects is particularly challenging, since every point in the field of view appears to be moving in a manner dependent on its relative motion with regard to the observer and the (unknown) structure of the viewed scene.

- **Egomotion estimation** [197, 44, 100, 115, 8, 141, 175, 120, 62, 148]
  This capability deals with the estimation of the velocity of a mobile observer with respect to his environment. This information is essential for various servoing tasks that are based on visual feedback.

- **Obstacle detection and avoidance** [180, 48, 68, 123, 214, 56, 295, 212, 86, 150]
  In order to avoid collisions, a mobile observer should have a means for detecting

and avoiding objects obstructing his route. The observer should differentiate between the surface on which he is moving and the objects lying on it.

- **Time-to-contact estimation** [238, 51, 250, 163, 21, 149]
  A very useful visual measure for avoiding obstacles is time-to-contact, i.e. the amount of time that remains before the observer comes into contact with the objects appearing in his field of view. This measure is more useful compared to some form of distance, since it is defined with respect to the velocity of the observer, giving an estimate of the time period during which proper action to avoid collision should be taken.

- **Object interception (docking)** [57, 55, 213, 181]
  A mobile observer often needs to maneuver, so that he can approach certain objects of interest. This task is known as *docking*.

- **Wall and corridor following** [219, 29, 105, 15]
  In certain situations, a mobile observer needs to remain at a fixed, close distance to a wall while following it, or keep himself centered with respect to the walls of a corridor. These are the tasks of *wall* and *corridor following* respectively, and can be considered as simplified versions of the problem of controlling the heading direction in a cluttered environment.

- **Gaze maintenance (tracking)** [7, 185, 188, 169, 190, 12, 43, 38]
  Tasks such as visual surveillance require the observer to be capable of properly controlling the motion of his eyes, so that a moving object of interest remains approximately in the center of his field of view. Thus, the observer should be able to estimate the object's relative retinal velocity and then compensate for it by rotating his eyes appropriately.

- **Hand-eye coordination** [191, 192, 26, 116, 63, 93, 285]
  Assuming an observer equipped with a gripper, hand-eye coordination refers to the task of generating appropriate motor commands for positioning the gripper in 3D space and grasping objects. One of the main difficulties of this task is that all

measurements derived from images are expressed in the camera coordinate frame, which does not coincide with the coordinate frame attached to the gripper.

- **Visual homing** [139, 178, 59]

  Visual homing is at the zenith of the visual navigation capabilities hierarchy. It constitutes the ability of an autonomous system to use visual information for reaching a target location outside its field of view, starting from an arbitrary location. To achieve this, the system should extract, memorize and later recognize distinct locations in the environment (*landmarks*). Landmarks can be used either for determining the position of the system in the environment (*localization*) or as intermediate destinations when planning an appropriate route towards a final destination.

1.4 Overview of the Thesis

It is our belief that the behavioral approach to vision has the potential to provide in real time visual information adequate for closing the motor control loop. Furthermore, it allows the specification of a set of visual processes that permit the incremental development of a navigation system, so that there is no need for the whole system to be constructed before experiments can be conducted. This thesis describes the results of our investigations regarding the first four of the visual capabilities presented in the previous section, namely independent motion detection, egomotion estimation, obstacle detection and time-to-contact estimation. Since each capability deals with a well defined goal and does not depend critically on the environment, it can function as a generic navigational tool for various practical applications. Collectively, these visual capabilities constitute a solid arsenal of primitive algorithms that is able to support complex behavioral repertoires. Below, the adopted research framework and the specific contributions of this work are discussed.

1.4.1 Research approach

Our research effort has focused both on theoretical and practical aspects. First, a computational model regarding each visual capability has been derived from theoretical study. Since computer vision is primarily an experimental discipline, the next step was to demonstrate the validity of relevant models through experiments involving prototype implementations of proper algorithms. Work on both aspects has progressed along the following lines:

- **Avoid complete solutions to SFM.**

  We have already argued that complete solutions to SFM are an overkill for the application of vision-based navigation. As dictated by the purposive vision paradigm, more accurate and efficient techniques can be developed by addressing only the aspects relevant to the task at hand.

- **Avoid assumptions regarding the egomotion and the scene structure that are not always valid.**

  Restrictive assumptions regarding the egomotion and/or the structure of the viewed scene are not uncommon in the literature. For example, the cases of purely translational or rotational egomotion have been addressed and visible surfaces have been assumed to have special properties such as planarity or smoothness. With the exception of environmental constraints that are discussed below, such assumptions cannot be guaranteed to always hold and should be avoided when possible. This is because the violation of the assumptions on which the visual processes are based can be devastating for the autonomous system employing them. Imagine, for example, what could happen if a system moving with general 3D motion attempted to estimate its egomotion using an algorithm that assumes pure translation.

- **Exploit environmental constraints.**

  When designing the vision system of a mobile robot that will operate in a certain

setting, generality can be traded for simplicity and reliability by exploiting special constraints that might be satisfied in the specific environment. An indoor mobile robot, for instance, rovers on a planar floor and therefore, this constraint can be taken into account when designing the robot's visual processes. Although the visual processes that will result from such an assumption will be useless in the case of an outdoor robot, they will hopefully be more efficient and accurate compared to any visual processes that do not make a planarity assumption regarding the floor.

- **Develop algorithms that are robust with respect to noise.**

  Errors in the measurements of 2D image motion are unavoidable. Since the problem of inferring 3D information from 2D data is inherently nonlinear, special care must be taken to prevent corrupted, noisy measurements from yielding completely erroneous 3D estimates. Some of the possible ways for achieving robustness are indicated below.

  -- **Exploit redundancy.**

  Constraints derived from local data tend to be erroneous and should be avoided when possible. Instead, overdetermined systems formed by multiple constraints arising from global data should be used. Techniques borrowed from the field of robust statistics can then be used to identify and mask out the minority of invalid constraints [162, 252]. When estimating rigid egomotion for example, methods that employ flow measurements from the whole visual field perform better than methods relying on flow estimates in small image windows.

  -- **Use qualitative, nonmetric information.**

  It is often the case that the task of answering questions that admit a limited number of answers is easier and less sensitive to noise than the task of estimating quantitative (metric) information. For example, deducing which of a pair of objects is closer to the observer is easier than measuring their exact distances. Therefore, obstacle avoidance could be implemented by a reactive

scheme based on the sign of the rate of approach for different obstacles. In a similar spirit, to detect independent motion, one does not necessarily have to estimate the exact 3D motions of the observer and other objects in the scene. The quest for qualitative information is a powerful concept that can be particularly beneficial to the stability of vision systems.

-- **Rely on as simple representations as possible.**
Since noise in image measurements is inevitable, methods that are direct in the sense that they use the least amount of information possible, might be more robust than the algorithms that compute more than they need. An algorithm, for example, that solves a specific task using normal flow only is expected to be more robust compared to an algorithm that employs optical flow. Also sparse optical flows and disparity maps which supply no measurements over underconstrained image regions are usually more accurate, and thus more preferable compared to dense optical flow fields and disparity maps.

- **Develop computationally efficient algorithms.**
Our ultimate goal in studying vision-based navigation is to construct autonomous seeing systems. Such systems will have to process huge amounts of data in real time, i.e. at a rate comparable to the rate that the environment they operate in changes. A mobile system poses restrictions regarding available onboard space, power consumption, cost, etc. Although computers get faster all the time, the computational requirements of visual processes should be of serious concern. Issues such as time and memory space requirements, ease of parallelization, etc, should be addressed and thoroughly understood.

1.4.2 Outline and contributions

The present section outlines the organization of the thesis and provides a synopsis relevant results.

Chapter 2 is devoted to the presentation of some background material that is essential for the comprehension of the work presented in the subsequent chapters. Specifically, this chapter discusses issues related to motion representation and derives the equations relating 2D retinal velocities to the 3D motion and structure of the environment. Moreover, a brief introduction to the field of robust statistics is provided, with emphasis on the *Least Median of Squares* robust estimator.

The following four chapters describe in detail the developed and implemented visual capabilities. Since each capability is independent of the others, the corresponding chapter is self-contained and can be read independently. Each consists of a detailed review of the relevant literature, the theoretical development of the corresponding method, an experimental evaluation and a discussion of the advantages and weaknesses of the method.

Chapter 3 considers a fundamental problem of visual perception of motion, namely the problem of visual detection of independent 3D motion. This problem concerns the identification of objects that move independently of a mobile observer within his field of view. Most of the existing techniques for detecting independent motion rely on restrictive assumptions about the environment, the observer's motion, or both. Moreover, they are based on the computation of a dense optical flow field, which amounts to solving the ill-posed correspondence problem. In this work, independent motion detection is formulated as a problem of robust parameter estimation applied to the visual input acquired by a rigidly moving observer. The proposed method automatically selects a planar surface in the scene and the *residual planar parallax* normal flow field with respect to the motion of this surface is computed at two successive time instances. The two resulting normal flow fields are then combined in a linear model. The parameters of this model are related to the parameters of self-motion (egomotion) and their robust estimation leads to a segmentation of the scene based on 3D motion. The method avoids a complete solution to the correspondence problem by selectively matching subsets of image points and by employing normal flow fields. Experimental results demonstrate

the effectiveness of the proposed method in detecting independent motion in scenes with large depth variations and unrestricted observer motion.

Chapter 4 deals with the problem of using visual input to estimate egomotion, i.e. the velocity of a mobile system with respect to its environment. Knowledge of the egomotion is essential for various servoing tasks that are based on visual feedback. Many of the existing techniques for solving this problem rely on restrictive assumptions regarding the observer's motion or even the scene structure. Moreover, they often resort to searching the high dimensional space of possible solutions, which might be inefficient and exhibit convergence problems. In this work, a novel linear constraint that involves quantities that depend on the egomotion parameters is derived. This constraint is defined in terms of the optical flow vectors pertaining to four collinear image points and is applicable regardless of the egomotion or the scene structure. In addition, it is exact in the sense that no approximations are made for deriving it. Combined with robust linear regression techniques, the proposed constraint enables the recovery of the FOE, thereby decoupling the 3D motion parameters. Extensive simulations as well as experiments with real optical flow fields provide evidence regarding the performance of the proposed method under varying noise levels and camera motions.

Obstacle detection and avoidance capabilities are essential to an autonomous robot for it to move safely in the environment. Chapter 5 presents a method that enables a mobile robot to locate obstacles in its field of view using two images of its surroundings. The method provides a binary labeling of image points, classifying them either as obstacles or as free space. Assuming that the robot is moving on a locally planar ground, the method uses a set of reference point features (corners), that have been matched between the two views, to compute a robust estimate of the homography of the ground. Knowledge of this homography permits us to compensate for the motion of the ground and to detect obstacles as areas in the image that appear nonstationary after the motion compensation. The resulting method does not require camera calibration, is applicable either to stereo pairs or to image motion sequences, does not rely on a

dense disparity/flow field and circumvents the 3D reconstruction problem. Experimental results from the application of the method on real images indicate that it is both effective and robust.

Chapter 6 is complementary to chapter 5. It presents a method for estimating the *time-to-contact*, i.e. the time that remains before a mobile observer collides with objects in his field of view. The time-to-contact provides a measure for assessing the proximity of obstacles that is well suited to avoiding collisions. In chapter 6, a novel method for estimating the time-to-contact is presented. The method is based on the assumption that the robot is moving on a locally planar ground and employs the optical flow field induced by the robot's motion. First, the time-to-contact with points on the ground is estimated and then the concept of *planar parallax* is exploited to recover the time-to-contact with obstacle points. The computation of high order derivatives of image flow, which are known to be very difficult to estimate accurately, is completely avoided. Moreover, no knowledge of the egomotion is necessary. Experimental results from the application of the method on real and simulated flow fields are also reported.

Finally, chapter 7 concludes the thesis by summarizing its contributions. In addition, a brief discussion concerning possible directions of further research is provided.

Chapter 2

Mathematical Preliminaries

Before proceeding to the description of the visual capabilities that have been studied, issues related to visual motion representation are discussed. Models for representing image motion describe image measurables such as the 2D displacement of image points in terms of the scene depth and the relative 3D motion between the scene and the camera. These models form the basis on which the theoretical developments of the next chapters will be grounded. Depending on the frequency of the temporal sampling used when capturing an image sequence, image motion can be modeled in two ways, infinitesimally in time and over discrete time intervals. In the first case, it is termed as the *2D instantaneous motion field* or *optical flow* whereas in the second it is referred to as the *disparity map*. In the following sections, the mathematical expressions for both models are obtained. Furthermore, since the developed techniques involve the solution of overdetermined systems of linear equations, a brief introduction to the field of robust regression is provided. Special emphasis is given to the *Least Median of Squares (LMedS)* estimator, since its characteristics make it particularly attractive for the purposes of our work.

2.1 Image Motion Equations in the Continuous Case

Relative motion between an image acquisition system and the environment it is observing gives rise to *image motion*, that is motion of the projections of environmental points on the imaging system's retina. This is illustrated in Fig. 2.1, which shows a side view of a camera (see also Fig. 2.2). Each point on the retina can be associated with a 2D *velocity vector* and the set of such vectors defines a *velocity field*. Assuming infinitesimal motion between successive images, the following subsections derive the equations describing velocity fields and discuss other related issues.

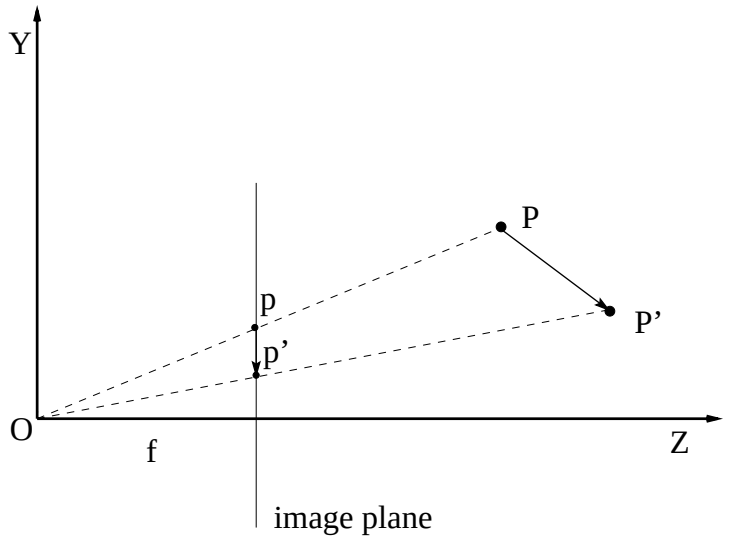


Figure 2.1: The displacement of a 3D point P to a new location P', creates a projected motion vector in the image.

2.1.1 The camera coordinate system

Consider a coordinate system $OXYZ$, the origin of which is at the optical center (nodal point) of a pinhole camera, and such that the OZ axis coincides with the optical axis (see Fig. 2.2). Under perspective projection, the 3D point $P(X, Y, Z)$ projects to the point

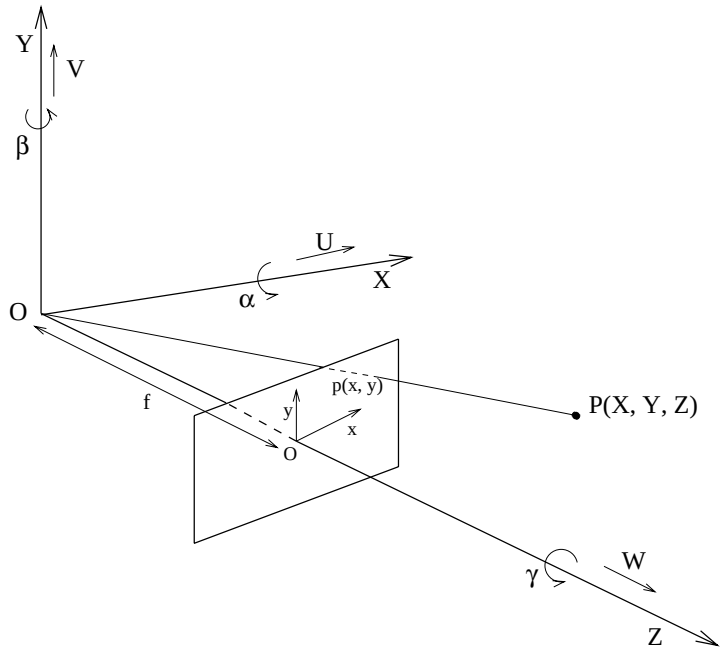


Figure 2.2: The camera coordinate system.

$p(x, y)$ on the image plane, according to the relations:

$$\begin{aligned} x &= \frac{Xf}{Z} \\ y &= \frac{Yf}{Z}, \end{aligned} \tag{2.1}$$

where f denotes the focal length of the imaging system. Note that the image coordinates (x, y) are defined with respect to the coordinate system that is centered on the *principal point*, i.e. the intersection of the optical axis with the image plane. This coordinate system is often referred to as the *normalized camera frame*. However, actual image acquisition systems provide image pixels in coordinate frames that have different origins and axis units from the normalized frame. Here, the change of coordinate frames will be studied and the camera intrinsic parameters will be defined.

Figure 2.3 depicts the normalized coordinate frame $(o, \mathbf{i}, \mathbf{j})$ and the new coordinate frame $(o_p, \mathbf{I}, \mathbf{J})$, which we call the *pixel coordinate frame*. Usually, the origin of the pixel coordinate frame is located at one of the image corners and, due to the electronics of acquisition, the unit vectors on the two axes are scaled with respect to those of the

normalized frame [71]. Therefore, we have that $\mathbf{i} = f_x\mathbf{I}$ and $\mathbf{j} = f_y\mathbf{J}$, where f_x and f_y is the focal length expressed in horizontal and vertical pixels respectively[1]. Let $\mathbf{m}$ be a point in the image and (x, y) its coordinates in the normalized coordinate frame. Let also (c_x, c_y) be the coordinates of the principal point expressed in the pixel coordinate frame. The quantities f_x, f_y, c_x and c_y depend on the camera only and are known as the camera *intrinsic calibration parameters*. The process of determining them, often with the aid of a calibration grid, is known as *camera calibration* [259, 159]. With reference to Fig. 2.3, the coordinates (x_p, y_p) of point $\mathbf{m}$ in the pixel coordinate frame can be expressed in terms of the normalized coordinates and the intrinsic calibration parameters by noting that

$$\overrightarrow{\mathbf{o_p m}} = \overrightarrow{\mathbf{o_p o}} + \overrightarrow{\mathbf{o m}} \tag{2.2}$$

Vector $\overrightarrow{\mathbf{o m}}$ is equal to (x, y) in the normalized coordinate system or to $(f_x\ x,\ f_y\ y)$ in the pixel coordinate frame. Thus, Eq. (2.2) yields

$$(x_p,\ y_p) = (c_x,\ c_y) + (f_x\ x,\ f_y\ y)$$

From this equation, it is clear that given the pixel coordinates, the normalized ones can be computed as

$$\begin{aligned} x &= \frac{x_p - c_x}{f_x} \\ y &= \frac{y_p - c_y}{f_y} \end{aligned} \tag{2.3}$$

In the following, we will assume that the intrinsic parameters are available and Eq. (2.3) has been employed to convert pixel coordinates to the normalized coordinate frame with $f = 1$. Therefore, the adjective ‘‘normalized’’ is omitted when we refer to normalized image coordinates.

[1] In practice, image pixels are not square.

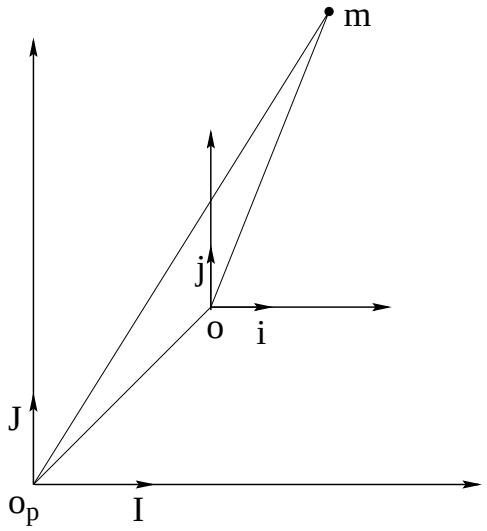


Figure 2.3: Changing coordinate systems in the image plane.

2.1.2 The equations of image point velocities

Suppose that the camera of Fig. 2.2 is moving rigidly with respect to its 3D static environment, with instantaneous translational and rotational velocity $\vec{t} = (U, V, W)$ and $\vec{\omega} = (\alpha, \beta, \gamma)$ respectively. This situation is geometrically equivalent to the case of a static camera, imaging a scene that is moving relative to it with translational velocity $-\vec{t}$ and rotational velocity $-\vec{\omega}$. The resultant velocity $\vec{V}$ of a point $P = (X, Y, Z)$ with respect to the $OXYZ$ coordinate system can be computed as follows. With reference to Fig. 2.4, let L be the axis of rotation defined by the vector $-\vec{\omega}$ and d be the distance of P from this axis. Point P rotates around L in a counterclockwise direction. The velocity $\vec{V}$ is the vector sum of the tangent velocity due to rotation $\vec{q}$ and the translational motion $-\vec{t}$. The tangential component $\vec{q}$ 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 L, that is $||\vec{q}|| = || - \vec{\omega}||\ d = || - \vec{\omega}||\ ||P||sin\theta = || - \vec{\omega} \times P||$, where "$\times$" denotes vector cross product. Thus, $\vec{V}$ is equal to

$$\vec{V} = -\vec{t} - \vec{\omega} \times \vec{r} \tag{2.4}$$

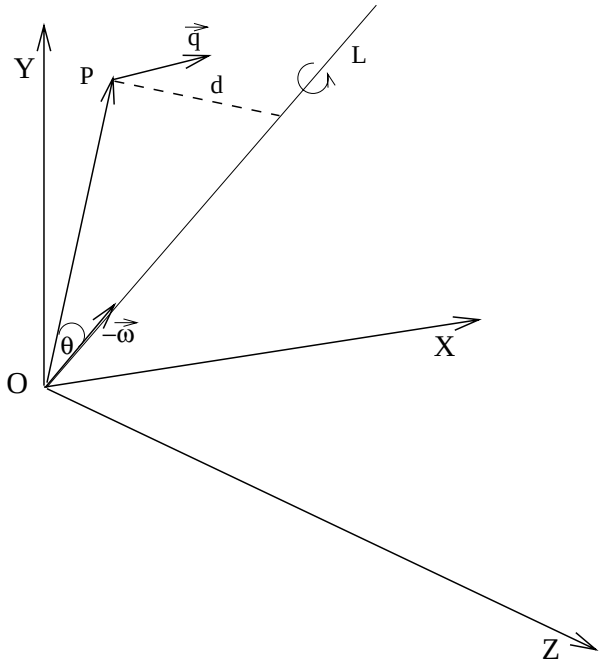


Figure 2.4: Rotation of a point P around the axis L.

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

$$\begin{aligned} X' &= -U - \beta Z + \gamma Y \\ Y' &= -V - \gamma X + \alpha Z \\ Z' &= -W - \alpha Y + \beta X, \end{aligned} \tag{2.5}$$

where $'$ denotes differentiation with respect to time.

Equations (2.5) can be used to relate image motion to the 3D velocity and structure, as follows. The relative motion between point $P(X, Y, Z)$ and the camera, results in a motion of P's projection $p(x, y)$ in the image. More specifically, the retinal velocity (u, v) at point (x, y) is given by:

$$\begin{aligned} u &= \frac{dx}{dt} \\ v &= \frac{dy}{dt} \end{aligned} \tag{2.6}$$

After substituting Eq. (2.1) into Eq. (2.6), differentiating and employing the expressions for the 3D velocity given by Eq. (2.5), 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 [144]:

$$\begin{aligned} u &= \frac{(-Uf + xW)}{Z} + \alpha\frac{xy}{f} - \beta\left(\frac{x^2}{f} + f\right) + \gamma y \\ v &= \frac{(-Vf + yW)}{Z} + \alpha\left(\frac{y^2}{f} + f\right) - \beta\frac{xy}{f} - \gamma x \end{aligned} \tag{2.7}$$

Note that (u, v) in the above equations are expressed in the normalized coordinate frame. By differentiating Eqs.(2.3) with respect to time, it is easy to see that the normalized velocities are related to the velocities (u_p, v_p) expressed in the pixel coordinate frame by

$$\begin{aligned} u &= \frac{u_p}{f_x} \\ v &= \frac{v_p}{f_y} \end{aligned} \tag{2.8}$$

Equations (2.7) can be written as

$$\begin{aligned} u &= u_{trans} + u_{rot} \\ v &= v_{trans} + v_{rot}, \end{aligned}$$

where

$$u_{trans} = \frac{(-Uf + xW)}{Z} \quad , \quad v_{trans} = \frac{(-Vf + yW)}{Z}$$

and

$$u_{rot} = \alpha\frac{xy}{f} - \beta\left(\frac{x^2}{f} + f\right) + \gamma y \quad , \quad v_{rot} = \alpha\left(\frac{y^2}{f} + f\right) - \beta\frac{xy}{f} - \gamma x$$

Clearly, the image velocity equations are separable, i.e. they consist of two parts, one due to translation and one due to rotation. It is interesting to note that only the translational part depends on the structure (Z) of the viewed scene. Another observation regarding the above equations is that the motion field is invariant under scaling of the depth Z and the

translational velocity (U, V, W). In other words, as explained in Fig. 2.5, absolute depths and translation magnitudes cannot be recovered from retinal velocities. Consequently, only the direction of translation $(\frac{Uf}{W}, \frac{Vf}{W})$ and the depth up to an unknown scale factor (i.e. shape) can be derived from an image velocity field. The point $(x_0, y_0) \equiv (\frac{Uf}{W}, \frac{Vf}{W})$ is of central importance to motion understanding [202] and has the property that the translational components of velocity vectors at each image point lie on lines going through it (see Fig. 2.6). This point is known as the *Focus of Expansion (FOE)*, since if the observer approaches the scene, the translational components radiate outward from it. In the case that the observer moves away from the scene, the translational components move towards the FOE, which is then referred to as the *Focus of Contraction (FOC)*.

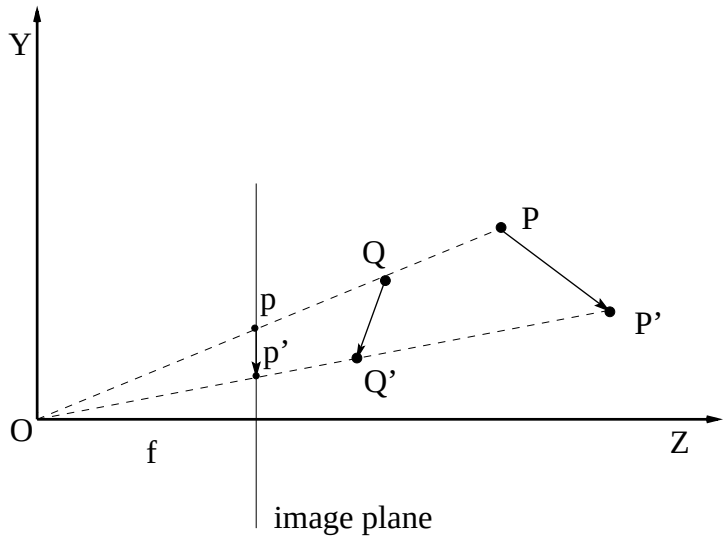


Figure 2.5: Based on the image velocity alone, we cannot differentiate between a distant point moving fast (i.e. point P moving to P') and a close point moving slowly (point Q moving to Q').

2.1.3 Optical flow field - Motion field

Equations (2.7) 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

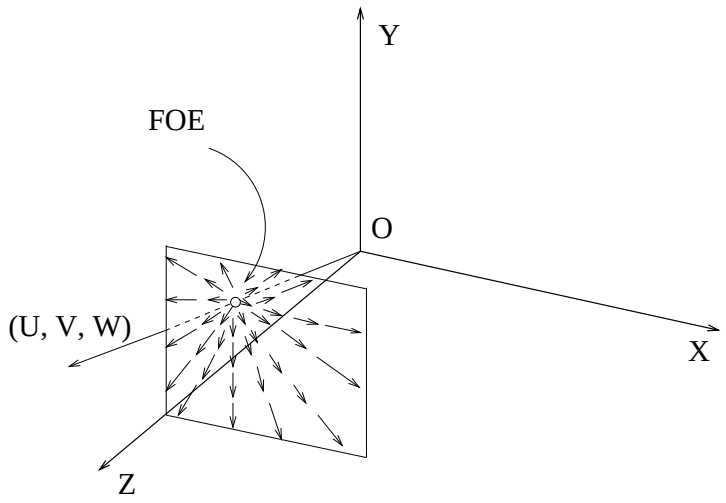


Figure 2.6: The FOE is the point on the image plane where the vector defined by the 3D translational velocity pierces the image plane. The translational components of image velocity vectors emanate from the FOE.

geometrical concept and, perhaps surprisingly, it is not necessarily identical to the optical flow field, which describes the motion of brightness patterns observed because of the relative motion between an imaging system and its environment. Horn [107], provides a classical example which demonstrates that the the optical flow field and the motion field are not equal. Consider a perfectly uniform sphere rotating in front of an imaging system. The curvature of the sphere's surface will give rise to spatial variations of brightness and shadows in an image of the sphere. This shading, however, does not move with the surface and therefore, the image does not change over time. This implies that the optical flow is zero everywhere, although the motion field is not. Conversely, consider a stationary sphere illuminated by a moving light source. The motion of the light source will result in a change of the shadings in the image over time. Clearly, in this case, the optical flow field is nonzero while the motion field is zero everywhere. Since the optical flow is the only source of information regarding motion that can be made available from the processing of images, the above inequality is an inherent problem for motion analysis. Fortunately, except for special cases such as the ones described previously, the optical flow field does not differ much from the motion field. In practice,

it is assumed that the optical flow field is equal to the motion field and the two terms are used interchangeably to refer to image velocity fields. Verri and Poggio [267] have quantified the difference between the motion and the optical flow fields and concluded that the two are identical only in the case of purely translating and uniformly illuminated Lambertian surfaces. Even in the cases that these two fields are indeed identical, the computation of the optical flow field is a difficult problem for reasons that will be explained in section 2.1.6.

2.1.4 The optical flow constraint equation

Assuming dense time sampling during image acquisition, 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. Thus, $I(x, y, t)$ denotes 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:

$$I(x + u\delta t, y + v\delta t, t + \delta t) = I(x, y, t) \tag{2.9}$$

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 $\delta t \to 0$, yields

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

Using the abbreviations

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

and

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

we obtain

$$I_x u + I_y v + I_t = 0 \tag{2.10}$$

Equation (2.10), which was originally developed by Horn and Schunck [108], is known as the *optical flow constraint equation* (or the *brightness constancy constraint equation*) because it gives one constraint for the components u and v of optical flow. Writing Eq. (2.10) in the form of a dot product

$$(I_x, I_y) \cdot (u, v) = -I_t \tag{2.11}$$

facilitates its geometrical interpretation as permitting the computation of the projection of an optical flow vector along the intensity gradient direction (i.e. the perpendicular to the edge at that point). This projection is also known as *normal flow*. Equation (2.11) can be viewed as the mathematical expression of the *aperture problem* which is explained schematically in Fig. 2.7. Assuming that an image feature such as a line is being watched through a small aperture, it is impossible to determine where each point of the feature has moved between two successive image frames. The only information that is readily available from local measurements is the component of the actual velocity along the direction that is perpendicular to the feature. On the contrary, the component of the velocity that is parallel to the feature cannot be determined. The aperture problem manifests itself regardless of the technique that is employed to estimate the optical flow. It is also directly related to the normal flow field, which is further examined below.

2.1.5 Normal flow field - normal motion field

The algebraic value of normal flow can be computed from Eq. (2.11), by dividing both sides with the gradient magnitude, and is equal to:

$$-\frac{I_t}{\sqrt{I_x^2 + I_y^2}} \tag{2.12}$$

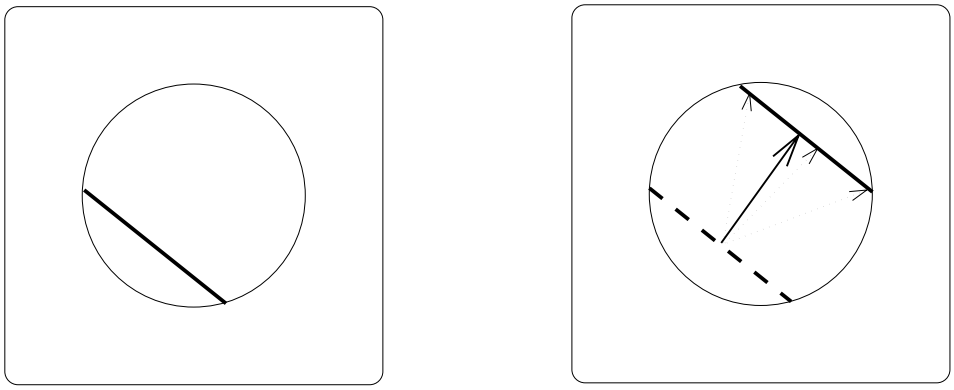


Figure 2.7: A schematic view of the aperture problem. The solid line in the left image has moved to a new position in the right image. Based on the information that is visible through the aperture, it is not possible to decide which of the dotted vectors corresponds to the motion of the line. However, whatever the motion might be, its projection on the direction perpendicular to the line is unique and is denoted by the solid vector.

As defined previously, the normal flow is the component of the optical flow along the image gradient and is 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 intensity gradient magnitude. Recalling that normal flow depends on the image gradient direction, it is obvious that there are infinitely many normal flow fields that can result from a particular optical flow field, depending on the actual gradient directions present in a scene. This is shown in Fig. 2.8, where the thin vectors denote some of the possible normal flow vectors that can be generated from the optical flow vector designated by the thick vector.

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 identical to the motion field [267]. The algebraic value of the normal

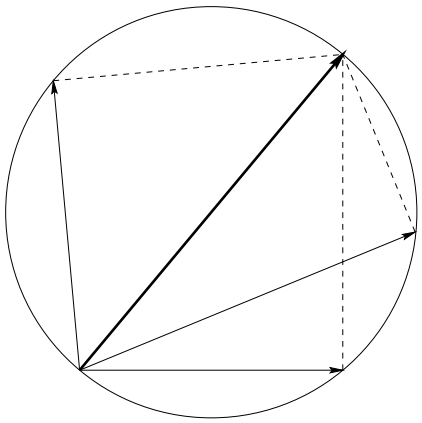


Figure 2.8: The set of normal flow vectors that can possibly originate from a single optical flow vector defines a circle having the given optical flow vector as a diameter.

motion field is equal to

$$\begin{aligned} \left(\frac{dx}{dt}\ ,\ \frac{dy}{dt}\right) \cdot \frac{(I_x\ ,\ I_y)}{\sqrt{I_x^2 + I_y^2}} &= \\ \left(\frac{dx}{dt}\ ,\ \frac{dy}{dt}\right) \cdot \frac{\nabla I}{||\nabla I||} &= \\ \frac{1}{||\nabla I||}\left(I_x\frac{dx}{dt} + I_y\frac{dy}{dt}\right) & \end{aligned} \tag{2.13}$$

The difference between the magnitudes of a normal flow vector and the corresponding normal motion vector, is given by the difference of Eqs. (2.12) and (2.13) and is equal to [267]:

$$\frac{1}{||\nabla I||}\frac{dI}{dt} \tag{2.14}$$

Equation (2.14) shows that normal flow is a good approximation to normal motion flow at points where $||\nabla I||$ assumes large values. Consequently, normal flow vectors at points corresponding to image edges provide reliable information that can be used for recovering the 3D motion and structure.

The expression relating the normal motion flow to the sought 3D quantities is derived as follows. Let (n_x, n_y) be the unit vector in the gradient direction. The magnitude un of a normal motion flow vector is given by

$$un \ = \ un_x + vn_y \tag{2.15}$$

which, by substitution from Eq. (2.7), yields:

$$\begin{aligned} un &= (-n_x f)\frac{U}{Z} \\ &+ (-n_y f)\frac{V}{Z} \\ &+ (xn_x + yn_y)\frac{W}{Z} \\ &+ \left\{\frac{xy}{f}n_x + \left(\frac{y^2}{f} + f\right)n_y\right\}\alpha \\ &- \left\{\left(\frac{x^2}{f} + f\right)n_x + \frac{xy}{f}n_y\right\}\beta \\ &+ (yn_x - xn_y)\gamma \end{aligned} \tag{2.16}$$

Equation (2.16) highlights some of the difficulties of motion analysis based on the normal flow. 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. In other words, N estimates of normal motion flows supply N equations in $N + 6M$ unknowns, where M is the number of different rigid motions within the camera field of view.

2.1.6 Optical flow estimation

This section gives a brief overview of the existing approaches to optical flow estimation. For more extensive treatments, the reader is referred to [228, 28, 165].

Optical flow is to be estimated from the time varying brightness patterns that are recorded by an imaging system. Estimation is accomplished through the exploitation

of a set of constraints regarding the observed patterns and the nature of the optical flow. Singh [228] stresses the fact that all optical flow estimation algorithms progress in two phases. First, based on the assumption that some image property is conserved, local velocity information is derived. Image properties that are typically used for this purpose include image intensity, spatiotemporal derivatives of intensity and intensity distribution. Then, the local velocity information computed in the previous step is combined with some additional constraints to recover the full velocity field. The constraints that are employed solve the aperture problem by fusing velocity information in small image neighborhoods. According to the image property that is exploited, optical flow estimation techniques can be classified into the following three categories.

- **Gradient based techniques**

  Techniques falling in this category assume the conservation of image intensity and use Eq. (2.10) to constrain the optical flow [284, 78, 108, 170, 101, 257, 172, 266, 30, 278]. This yields an underconstrained system composed of a single equation in two unknowns. Therefore, in order to estimate retinal velocities, such techniques rely on additional assumptions such as smoothness of the optical flow field. In more concrete terms, it is postulated that nearby points move in a similar manner and normal velocities are integrated over image regions [108] or along image contours [277, 172]. Next, the full velocities are recovered by minimizing an appropriate function of these integrals.

- **Correlation based techniques**

  These techniques rely on the conservation of local intensity distribution [284, 151, 46, 10, 9, 87]. Using two images of a time-varying scene, for each pixel in the first image they search for the best matching pixel in the second image. This best matching pixel is selected from a set of candidate matches that lie in an image region called the search window. The selection of the best candidate match is based on the maximization of a match measure, such as the correlation between a small window around the pixel under consideration and a corresponding window

centered on the candidate match. Obviously, correlation based methods suffer from the aperture problem in areas of strongly oriented intensity gradients such as edges, since in this case the correlation operation yields multiple local maxima.

- **Spatiotemporal energy-based techniques**

  These techniques are analogous to gradient based techniques in the spatiotemporal frequency domain [274, 1, 83, 99]. It has been proven that the spatiotemporal frequencies of a moving stimulus are related to the velocity of the stimulus by

$$\omega_x u + \omega_y v + \omega_t = 0, \tag{2.17}$$

  where ω_x, ω_y are the spatial frequencies and ω_t is the temporal one. This implies that the above three frequencies lie on a plane in the spatiotemporal frequency space, whose orientation depends upon the velocity of the visual stimulus. Spatiotemporal energy filters that are tuned to certain frequencies are used to derive optical flow. In the case of a textureless stimulus, there exists a single pair of spatial frequencies and the corresponding power spectrum is concentrated along a line in the frequency space. In other words, the motion plane cannot be fully determined, as dictated by the aperture problem. However, in the case of a textured moving pattern, the aperture problem does not exist. A textured pattern gives rise to a number of spatial frequency components in the neighborhood of each pixel and therefore the full optical flow can be derived.

Despite the considerable progress that has been made regarding optical flow estimation, the problem remains challenging and a generally applicable solution is still lacking. As pointed out by Mitiche and Bouthemy [165], the following difficulties pertain to all formulations of the optical flow estimation problem:

i. The image motion explaining an image brightness change is not unique, i.e. the problem of estimating optical flow is ill-posed [31]. Regularization techniques which are often used to alleviate this problem, may yield optical flow fields that

have no physical meaning. For example, the flow recovered around a motion discontinuity might be the average of the underlying actual flows.

Regularization is a mathematical tool that permits the recovery of an unknown function though the use of smoothness constraints [195, 244]. However, the optical flow field is not smooth for a number of reasons. First, as Eqs. (2.7) show, the smoothness of optical flow depends on the continuity of the depth variable Z. In most scenes depth discontinuities (e.g. occlusions) do exist and, therefore, the optical flow cannot be regarded as smooth. Second, the smoothness of the optical flow also depends on the constancy of 3D motion parameters. Thus, independently moving objects induce discontinuities in the optical flow field. Black and Anandan [34] provide a framework for estimating piecewise continuous optical flow. This framework is applicable to standard formulations of optical flow estimation, helping to reduce their sensitivity in the presence of phenomena such as transparency, depth discontinuities, independently moving objects, shadows and specular reflections.

ii. Changes in the illumination, surface reflections and nonuniformities, sensor noise and distortions, contribute to changes in the illumination that are not due to the relative motion between the imaging system and the environment. This is the case of unequal optical flow and motion field that was discussed in section 2.1.3.

iii. Physical models that are explicitly or implicitly embedded in the estimation process, are very often inadequate for accounting accurately for the image brightness formation. For example, the projections of a moving point may not have constant intensity, thus violating Eq. (2.10). Del Bimbo et al [33], consider various forms of the optical flow constraint equation that have been proposed in the literature and quantify their adequacy in modeling optical flow under different conditions. employed for optical flow estimation and analyze

2.1.7 Optical flow vs. normal flow

During the last decade, some researches have proposed the so-called *direct methods*, which refute the need for estimating the optical flow by advocating the use of normal flow as input to motion analysis algorithms [220, 16, 17, 18, 146, 8, 109, 176, 175, 222, 79, 94, 179, 5, 76, 85, 180, 249, 21]. Unquestionably, there exist some important reasons that favor direct approaches. To begin with, the problem of computing normal flow is well-posed, involving local measurements only and avoiding restrictive assumptions regarding the flow. In addition, it has low computational requirements compared to expensive optimizations that are often carried out by optical flow estimation algorithms. Hence, normal flow can be computed in real time with the aid of special image processing hardware. There is also no need for integrating measurements in image neighborhoods for solving the aperture problem. On the other hand, there are several disadvantages associated with the use of normal flow. Normal flow estimation involves the computation of intensity derivatives, an operation that is known to be sensitive to noise. In contrast to optical flow estimation, normal flow cannot be estimated using multiresolution techniques for coping with temporal aliasing caused by large time sampling periods [208, 46, 30]. The use of point calculations precludes the infliction of any kind of local spatial coherence on normal flow estimates. Even if the normal flow can be accurately estimated, it should be noted that it provides partial information regarding image motion, i.e. is less informative compared to the optical flow. Thus, although it might prove adequate for tasks that involve redundant measurements, such as the recovery of 3D motion, the lack of relevant publications probably implies that it is insufficient for direct recovery of scene structure.

For the development of the visual capabilities that are described in this thesis we have adopted a moderate approach for representing image motion. Instead of taking part in the religious war in favor of abolishing optical flow, we choose to employ whatever representation seems the most suitable for a particular problem. Our aim is to use the least amount of information that is adequate for supporting a specific capability.

2.2 Image Motion Equations in the Discrete Case

This section makes use of projective geometry to derive the equations relating discrete image displacements (i.e. disparities) between two uncalibrated images. The pinhole camera model that has been adopted in section 2.1 is also employed here. In contrast to section 2.1, however, the 3D motion between the viewpoints from which the two images have been captured can be arbitrarily large. More details can be found in [167, 71, 289].

2.2.1 Projective geometry and the projection equation

In the following, projective (homogeneous) coordinates are employed to represent image and scene points by 3×1 and 4×1 column vectors respectively. The symbol $\simeq$ will be used to denote equality of vectors up to a scale factor. Vectors and arrays will be written in boldface. For more in depth treatments of the application of projective geometry to computer vision, the interested reader is referred to [167, 168, 166, 131, 132].

Consider the $n+1$ dimensional space $\mathcal{R}^{n+1} - \{(0,\ldots,0)\}$ with the following equivalence relation:

$$(x_1,\ldots,x_{n+1})^T \simeq (x_1^{'},\ldots,x_{n+1}^{'})^T \;\; if \;\; and \;\; only \;\; if$$
$$\exists \;\; \lambda \neq 0, \;\; such \;\; that \;\; (x_1,\ldots,x_{n+1})^T = \lambda (x_1^{'},\ldots,x_{n+1}^{'})^T$$

This space is the projective space $\mathcal{P}^n$, in which the $(n+1)$-tuples of coordinates $(x_1,\ldots,x_{n+1})^T$ and $(x_1^{'},\ldots,x_{n+1}^{'})^T$ represent the same point. A projective transformation from $\mathcal{P}^n$ into $\mathcal{P}^k$ is defined by a $(k+1) \times (n+1)$ full rank matrix. The usual Euclidian space $\mathcal{R}^n$ is embedded in the n-dimensional projective space $\mathcal{P}^n$ through the mapping

$$\mathcal{R}^n \ni (x_1,\ldots,x_n)^T \mapsto (x_1,\ldots,x_n,1)^T \in \mathcal{P}^n$$

Projective geometry can elegantly model a pinhole camera imaging a three dimensional scene. The scene geometry is captured by the Euclidian space $\mathcal{R}^3$ embedded

in the 3-dimensional projective space $\mathcal{P}^3$, and the image plane is described by the 2-dimensional Euclidian space $\mathcal{R}^2$, embedded in $\mathcal{P}^2$. Then, perspective projection of a 3D point on the 2D image plane can be modeled as a linear projective mapping from $\mathcal{P}^3$ to $\mathcal{P}^2$ [71]. The projection matrix defining this mapping is given by

$$\underbrace{\begin{pmatrix} f_x & \kappa & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}}_{\mathbf{K}}$$

where $\mathbf{A}$ is the matrix of intrinsic camera parameters and $\mathbf{K}$ is the perspective projection matrix for the normalized camera frame [289]. More specifically, the elements f_x and f_y represent the focal length of the camera expressed in horizontal and vertical pixel units respectively, (c_x, c_y) is the principal point and κ, often assumed to be zero, is a skew parameter which is related to the angle between the horizontal and vertical axes of the sensor array. The ratio $\frac{f_y}{f_x}$ is referred to as the aspect ratio. Matrix $\mathbf{A}$ can be determined either with the aid of a calibration grid [259] or through a *self-calibration* process by observing an unknown scene [159, 296]. Thus, a point $\mathbf{M}^T = [X, Y, Z, 1]^T$ not on the image plane, projects to the image point $\mathbf{m}^T = [x, y, 1]^T$ according to the relation

$$Z\,\mathbf{m} = \mathbf{A}\,\mathbf{K}\,\mathbf{M} \tag{2.18}$$

2.2.2 The general disparity equation

Suppose now that two views of the same scene are available. The pair of views might have been captured either by a stereo rig or by a single moving camera. Our aim is to develop the equations relating the projections of a scene point in the two views. The two camera coordinate systems can be aligned by a rotation followed by a translation. As illustrated in Fig. 2.9, this change of coordinate systems is expressed by a 4×4 matrix composed of a 3×3 rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$. More specifically, a 3D point $\mathbf{M}^T$ in the first view is transformed to point $\mathbf{M'}^T = [X', Y', Z', 1]^T$ in the second

view, according to

$$\mathbf{M}' = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_3^T & 1 \end{pmatrix} \mathbf{M}$$

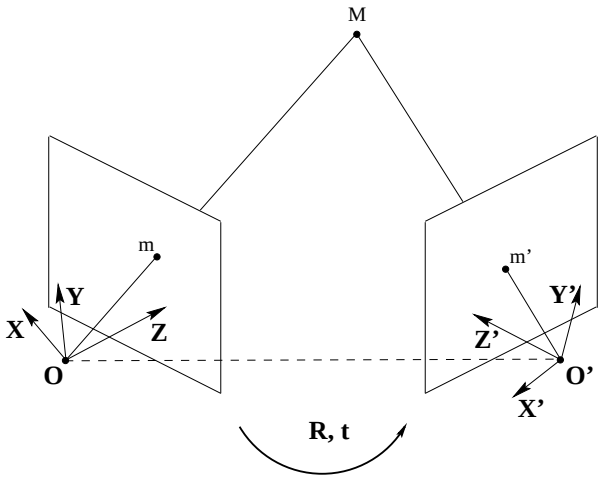


Figure 2.9: The $OXYZ$ coordinate system is transformed to $O'X'Y'Z'$ through a rotation $\mathbf{R}$ followed by a translation $\mathbf{t}$.

Assuming that point $\mathbf{M}^T$ projects to image points $\mathbf{m}$ and $\mathbf{m}'^T = [x', y', 1]^T$ in the first and second view respectively, the disparity equation relating $\mathbf{m}$ and $\mathbf{m}'$ is given by

$$Z'\,\mathbf{m}' = \mathbf{A}'\,\mathbf{K}\,\mathbf{M}' = \mathbf{A}'\,[\mathbf{R}\ \ \mathbf{t}]\,\mathbf{M} = \mathbf{Z}\,\mathbf{A}'\,\mathbf{R}\,\mathbf{A}^{-1}\,\mathbf{m} \ + \ \mathbf{A}'\,\mathbf{t} = Z\,\mathbf{H}_\infty\,\mathbf{m} \ + \ \mathbf{e}'$$

where $\mathbf{A}'$ is the intrinsic parameters matrix for the second view, $\mathbf{H}_\infty = \mathbf{A}'\,\mathbf{R}\,\mathbf{A}^{-1}$ and $\mathbf{e}' = \mathbf{A}'\,\mathbf{t}$. Matrix $\mathbf{H}_\infty$ is known as the *homography of the plane at infinity* and $\mathbf{e}'$ is the *epipole*, i.e. the projection of the focal center of the first view to the second. The above relation implies that point $\mathbf{m}'$ lies on the line going through $\mathbf{e}'$ and the point $\mathbf{H}_\infty\,\mathbf{m}$ (see Fig. 2.10). This line is the *epipolar line* of point $\mathbf{m}$ and is given by the vector $\mathbf{F}\,\mathbf{m}$, where $\mathbf{F}$ is the singular 3×3 matrix given by $\mathbf{F} = [\mathbf{e}']_\times\,\mathbf{H}_\infty$; $[\mathbf{e}']_\times$ is the 3×3 skew-symmetric matrix of rank 2 representing the vector product, i.e.

$$\begin{pmatrix} 0 & -e'_z & e'_y \\ e'_z & 0 & -e'_x \\ -e'_y & e'_x & 0 \end{pmatrix}$$

and $[\mathbf{e}']_\times \mathbf{x} = \mathbf{e}' \times \mathbf{x}$. The matrix $\mathbf{F}$ is the *fundamental matrix* [96, 95, 207, 153, 291, 253] and expresses mathematically the epipolar constraint in the case of uncalibrated cameras. This constraint can be compactly written as

$$\mathbf{m}'^T \mathbf{F} \mathbf{m} = 0 \tag{2.19}$$

Matrix $\mathbf{F}$ depends on the relative position of the two views and their corresponding intrinsic parameters, but not on the structure of the viewed scene. The fundamental matrix is the equivalent of the *essential matrix* [142, 262] in the uncalibrated case and plays a central role in applications involving the recovery of motion and structure information from uncalibrated images [268]. Since $det\mathbf{F} = 0$, $\mathbf{F}$ has seven degrees of freedom and at least seven corresponding points in two views are required for estimating it.

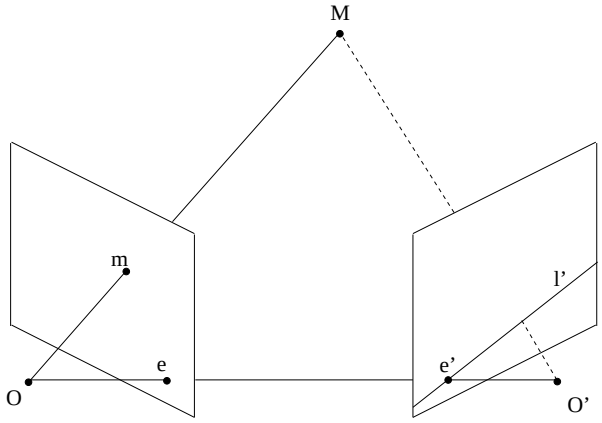


Figure 2.10: The epipolar constraint. The point corresponding to point m in the second view, must lie on the epipolar line $\mathbf{l}'$ which also contains the epipole $\mathbf{e}'$. The plane OMO' is the epipolar plane corresponding to point $\mathbf{M}$ and its image in the second view is the line $\mathbf{l}'$.

2.2.3 The case of a planar surface

Assume that the two views contain the image of a plane lying in the 3D space. Let the plane equation in the first view be $[\mathbf{n}^T \; - \; d]\, \mathbf{M} = 0$, where $\mathbf{n}$ is the unit vector normal

to the plane and d is the plane's distance from the optical center. Using Eq. (2.18), the plane equation can be written as

$$\mathbf{n}^T\ \mathbf{K}\ \mathbf{M}\ \ -\ \ d = Z\ \mathbf{n}^T\ \mathbf{A}^{-1}\ \mathbf{m}\ \ -\ \ d = 0$$

If the plane does not go through the optical center of the first view, then $d \neq 0$ and the disparity equation becomes

$$Z'\ \mathbf{m}' = Z\ \mathbf{H}\ \mathbf{m}$$

or

$$\mathbf{m}^{'} \simeq \mathbf{Hm}. \tag{2.20}$$

where $\mathbf{H}$ is the 3×3 nonsingular matrix that is equal to $\mathbf{H}_\infty\ +\ \mathbf{e}'\ \frac{\mathbf{n}^T}{d}\ \mathbf{A}^{-1}$. Matrix $\mathbf{H}$ is known as the *plane homography* (also also known as *plane projectivity* or *plane collineation*) and relates plane points in the first view to their corresponding points in the second view, without any knowledge of the camera calibration. Because of the fact that $\mathbf{H}$ is defined up to an unknown scale factor, it has 8 degrees of freedom. Therefore, noting that each pair of corresponding coplanar points provides 2 constraints, 4 pairs of corresponding coplanar points in general position (no three points are collinear) suffice for estimating it. This implies that a plane projectivity represents a plane to plane projective transformation which transforms any four points in general position to any other four points also in general position.

Matrices $\mathbf{F}$ and $\mathbf{H}$ are related by the fact that the matrix $\mathbf{F}^T\mathbf{H}$ is skew-symmetric [95, 207, 227], that is

$$\mathbf{F}^T\mathbf{H} + \mathbf{H}^T\mathbf{F} = \mathbf{0}, \tag{2.21}$$

where $\mathbf{F}^T$ and $\mathbf{H}^T$ are the transposes of $\mathbf{F}$ and $\mathbf{H}$ respectively.

2.3 Robust Regression

Regression analysis, in other words fitting a model to noisy data, is a very important subfield of statistics. In the general case of a linear model given by the relation

$$y_i = x_{i1}\theta_1 + \ldots + x_{ip}\theta_p + e_i, \tag{2.22}$$

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} [210]. 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:

$$\hat{y_i} = x_{i1}\hat{\theta_1} + \ldots + x_{ip}\hat{\theta_p} \tag{2.23}$$

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

$$r_i = y_i - \hat{y_i} \tag{2.24}$$

Traditionally, $\hat{\theta}$ is estimated by the least squares (LS) method, which is popular due to its low computational complexity [90]. LS involves the solution of a minimization problem, namely:

$$Minimize \sum_{i=1}^{n} {r_i}^2 \tag{2.25}$$

The LS estimator owes its popularity to the fact that a linear, closed-form solution to Eq. (2.25) can be found employing matrix pseudoinverses and Singular Value Decomposition (SVD) [200]. In addition, it can be proved that LS estimation 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 [37, 82, 117, 124, 297, 235]. From those, the RANSCAC method [82] is probably the most popular one. Other methods have been borrowed from statistics [32, 136, 162, 210, 229, 234]. Meer et al [162] and Zhang [292] provide excellent reviews of the use of robust regression methods in computer vision.

A very popular class of robust estimators consists of the M-estimators. M-estimators are based on the idea of replacing the squared residuals r_i^2 by another symmetric function of the residuals. The interested reader is referred to [114] for more details. M-estimators have not been employed in the context of this work because of the following two major drawbacks. First, it can be shown that although M-estimators behave better than least squares in practical situations, their breakdown point is equal to $1/n$ [210], where n is the number of observations. Clearly, this approaches zero as n increases[2]. Second, it can be shown that they require a reliable initial estimate of the model parameters, otherwise they can end up trapped in local minima.

In an effort to provide robust estimators with a higher breakdown point, Rousseeuw [210] introduced the so-called S-estimators which are defined by minimizing a robust measure of the scatter of the residuals. The most widely used 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 visual

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

navigation capabilities described in subsequent chapters. It can be argued that 50% is the highest possible breakdown point of an estimator, because for larger amounts of outlier contamination it is impossible to distinguish the ‘‘good’’ from the ‘‘bad’’ data. Recently, a new robust regression method, namely MINPRAN, has been proposed [235]. MINPRAN 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.

2.3.1 The Least Median of Squares robust estimator

The LMedS estimator, which was originally proposed by Rousseeuw [209], is able to handle data sets containing large amounts of outliers. LMedS involves the solution of a non-linear minimization problem, namely:

$$Minimize\ \{median_{i=1,\ldots,n}\ {r_i}^2\} \tag{2.26}$$

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%, a characteristic which makes it particularly attractive for the purposes of this work. Figure 2.11 demonstrates a representative example of the performance of LMedS relative to LS.

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

$$\hat{\sigma} = C\sqrt{med\ {r_i}^2} \tag{2.27}$$

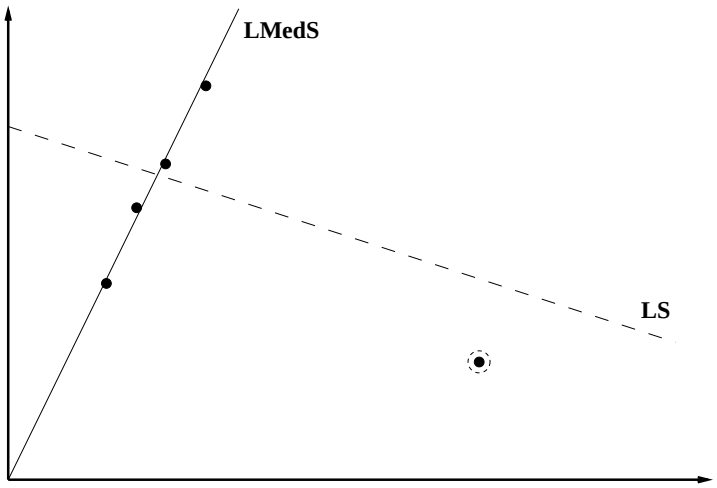


Figure 2.11: The performance of LMedS vs. LS: Given five noisy observations of a straight line, the solid line is the one fitted by LMedS while the dashed one is the one estimated by LS. Notice how a single highly erroneous observation (the one indicated with the dashed circle) can have a significant impact on the accuracy of the line estimated by LS.

where C is an application dependent constant. Rousseeuw and Leroy [210] suggest a value of

$$C = 1.4826\left(1 + \frac{5}{n-p}\right) \tag{2.28}$$

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

$$w_i = \begin{cases} 1, & \text{if } \frac{|r_i|}{\hat{\sigma}} \leq 2.5 \\ 0, & \text{if } \frac{|r_i|}{\hat{\sigma}} > 2.5 \end{cases} \tag{2.29}$$

All points with weight $w_i = 1$ correspond to model inliers, while points with weight $w_i = 0$ correspond to outliers. The threshold in Eq. (2.29) controls the sensitivity to outliers and its value reflects the fact that assuming a Gaussian distribution, very few residuals should be larger than $2.5\hat{\sigma}$. Note that since the criterion according to which observations are assigned the inlier/outlier binary label involves calculations over the mean of the residuals, it is itself robust. This implies that 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.

LMedS minimization is solved by a search in the space of possible estimates generated by the data. Since this search space is usually very large, a Monte-Carlo type of speedup technique is employed, in which a certain probability of error is tolerated [162]. More specifically, let p denote the number of parameters to be estimated; then there are $O(n^p)$ different p-tuples that can be formed by the available observations. Given that this search space grows exponentially with the number of observations, it is obvious that exhaustive search is prohibitively expensive in terms of computational complexity. In practical situations, this problem can be overcomed by a random, iterative scheme which guarantees that if a p-tuple of uncorrupted observations exists in the set of observations, it will be selected with high probability. Assuming that e is the fraction of data contaminated by outliers, then the probability Q that at least one out of m randomly selected p-tuples has only uncorrupted observations, is equal to:

$$Q = 1 - [1 - (1 - e)^p]^m \tag{2.30}$$

Thus, the solution of Eq. (2.30) for m, gives a higher bound for the number of p-tuples that should be tried. Note that Eq. (2.30) is independent of the number of available observations. Each of the m trials, requires the estimation of p candidate parameter values and the computation of the squared residuals between the observations and the predictions of the model. The set of parameter values that yields the minimum median residual, is declared as the solution of the regression problem. This solution is refined by a least squares estimation on the set of inliers [210]. At this point it is worth mentioning that LMedS supplies a general framework for dealing with multiple populations of data and does not impose any constraint on how is a candidate solution to be obtained. For example, the latter can be computed from a random sample of data using least squares, orthogonal regression or even a nonlinear, iterative scheme. Thus, LMedS can be employed even in cases of nonlinear regression.

As far as algorithmic improvements to reduce the execution time of LMedS are

concerned, it should be noted that the computation of the median of squared residuals that takes place at each iteration, can be attained without resorting to sorting the residuals. Instead, an algorithm that finds the kth largest number out of n numbers can be employed [218]. This algorithm has a time complexity of $O(n)$, which compares favorably to the $O(n \log n)$ complexity of the best serial sorting algorithm [58]. It should also be noted that the data dependency among the calculations involved in LMedS estimation is very small. This is because the computation of the median of the residuals for one candidate solution is independent of the computation of the residuals associated with another candidate solution derived from a different set of observations. Therefore, it is clear that LMedS estimation exhibits fine grain parallelism and thus its execution time would benefit greatly from a parallel implementation.

Part II

Development of Visual Capabilities

Chapter 3

Independent Motion Detection

3.1 Introduction

Independent 3D motion detection (IMD) is a fundamental motion perception capability of a seeing system. In a world where changes of state are often more important than the states themselves, the perception of independent motion provides a rich input to attention, informing a seeing system about dynamic changes in the environment [61, 14].

In the case of a static observer, the problem of independent motion detection can be treated as a problem of *change detection* [112, 230, 193]. The situation is much more complicated when the observer moves relative to the environment. In this case, even the static parts of the scene appear to be moving in a way that depends on the motion of the observer and on the structure of the viewed scene. The case of a moving observer, is also of great interest because biological and most man-made visual systems are usually in continuous motion.

In the case of a moving observer, IMD has often been approached as a problem of segmenting the 2D motion field that is computed from a temporal sequence of images. Wang and Adelson [272] for example, estimate affine models for optical flow in image

patches. Patches are then combined in larger motion segments based on a k-means clustering scheme that merges two patches if the distance of their motion parameters is sufficiently small. Bouthemy and co-workers [36, 183] also address the problem of segmenting the 2D velocity field and rely upon MRF models. Nordlund and Uhlin [182] estimate the parameters of an affine model of 2D motion, assuming that the estimation of the model parameters will not be considerably affected by the presence of small independently moving objects. IMD is then achieved by determining the points where the deviation of the measured from the predicted flow is large. Similar approaches have been pursued by Torr and Murray [254], Ayer et al [22], Bober and Kittler [35] and Irani et al [119] who combine normal flow with 2D parametric models for image velocities. Independent motion is then detected at the discontinuities of the parameters estimated for the adopted image motion model. The basic problem of the methods that employ 2D models is that they assume scenes where depth variations are small compared to the distance from the observer. However, in real scenes depth variations can be quite large and, therefore, 2D methods may detect discontinuities that are not only due to motion, but also due to the structure of the scene (see for example [14], p. 134).

Solutions to the problem of IMD have also been provided using 3D models. Employing 3D models makes the problem more difficult because extra variables regarding the depths of scene points are introduced. This in turn requires certain assumptions to be made in order to provide additional constraints to the problem. Most of the methods depend on the accurate computation of a dense optical flow field or on the computation of a sparse map of feature correspondences. Wang and Duncan [273], for instance, present an iterative method for recovering the 3D motion and structure of independently moving objects from a sparse set of velocities obtained from a pair of calibrated, parallel cameras. Da Vitoria Lobo and Tsotsos [141] use a constraint defined with respect to three collinear image points to estimate the egomotion from a dense optical flow field and then detect independently moving objects having small spatial extent. Other assumptions that are commonly made by existing methods are related to the motion of the observer, to the structure of the viewed scene, or both. Jain [122] and

Clarke and Zisserman [52] have followed the former approach and consider the IMD problem for an observer pursuing restricted translational motion. On the other hand, Adiv [2] performs segmentation by assuming planar surfaces undergoing rigid motion, thus introducing an environmental assumption. Torr and Murray [255] detect independent motion by recovering the set of fundamental matrices which optimally describes the epipolar constraint for the set of the observed point correspondences. Thompson and Pong [246] derive various principles for detecting independent motion when certain aspects of the egomotion or of the scene structure are known. However, the practical exploitation of the underlying principles is limited because of the assumptions they are based on and other open implementation issues. Sinclair [224] assumes that surfaces are locally planar and describes the motion of 3D points in terms of their angular velocity relative to the camera. His method detects independent motion that violates the epipolar constraint and recovers the orientations of the normals of planar patches. Sharma and Aloimonos [220] assume known egomotion and propose a direct method which detects independent motion at image points whose normal flow violates the epipolar constraint. Nelson [177] also develops two direct methods for IMD. The first is based on geometric constraints that are derived from a priori knowledge of the egomotion and upper bounds on the depths of the viewed scene. The second method is designed to detect rapidly accelerating objects rather than independent motion itself.

Argyros et al [16] present a method that uses stereoscopic information to segment an image into depth layers, in an effort to decompose the 3D problem into a set of 2D ones. The method provides reliable results at each depth layer, but there are certain limitations regarding the integration of results from the various depth layers. In Argyros et al [17], qualitative functions of depth estimated from stereo and motion are extracted in image patches. Comparison of these functions leads to conclusions regarding the number of 3D motions in a patch. The method is reliable and computationally efficient, but the resulting map of independently moving objects is coarse. In Argyros et al [18], the combination of depth and motion information extracted by a binocular observer permits the elimination of depth from the motion equations. This leads to a linear

model involving the 3D motion parameters and the problem of IMD is then solved by estimating the linear model with robust regression methods. Although [16, 17, 18] avoid any assumptions related to the egomotion or the scene structure and do not require the correspondence problem to be solved, their main disadvantage is that they assume that normal flow can be computed from a pair of stereo images, an assumption that is valid in special cases only.

In order to overcome the limitations of existing methods, a novel method for IMD is proposed in this chapter. This method is based on two key observations. The first is that, although an accurate solution to the correspondence problem by recovering the optical flow field is in the general case very difficult, the problem can be solved with satisfactory accuracy in special cases. Such a case involves the estimation of the optical flow field for points belonging to a planar surface, since once a planar surface in the scene has been identified, the problem of estimating its optical flow is a well-posed problem. The second observation is that the *residual parallax field* that remains after the registration of the images of a planar surface in two frames is an epipolar field. The proposed method exploits the information contained in the *normal residual field*, the component of residual motion in the direction of the image gradient. This field is less informative compared to the full residual flow, but can be more accurately computed from a temporal sequence of images. The combination of two such residual normal flow fields allows the elimination of the depth variables from the 3D motion equations, which in turn leads to the derivation of a model that is linear in the 3D motion parameters. IMD is then handled by applying a robust estimator to solve for the parameters of the linear model. Points that conform to the estimated model are labeled as moving due to the motion of the observer, while points that are characterized as outliers during the estimation process are labeled as independently moving. The proposed method assumes an observer that moves rigidly with unrestricted translational and rotational egomotion. Independent motion can be rigid or non-rigid and no calibration information is necessary.

The rest of this chapter is organized as follows. Section 3.2 develops a technique for identifying the dominant planar surface in a scene. Section 3.3 turns to the estimation of the dominant plane's motion. Section 3.4 outlines the decomposition of a normal flow field in terms of the normal flow field induced by the motion of a planar surface and a residual parallax normal flow field. The proposed method for IMD is detailed in section 3.5. Experimental results from the application of the method on real-world image sequences are presented in section 3.6. Finally, the chapter is concluded with a short discussion in section 3.7.

3.2 Dominant Plane Extraction

The traditional approach for identifying planar regions using two images of a scene has been to recover the depth of each point in the field of view and then segment the resulting depth map into planes. This process however, involves computations that are numerically unstable and requires difficult problems such as point correspondence and camera calibration to be solved. To avoid these difficulties, Sinclair et al [225] have proposed a method for identifying coplanar sets of corresponding points, using simple results from projective geometry. Based on [225], the dominant plane in a scene is extracted using a method which is briefly outlined in the following subsections.

3.2.1 The invariants of five coplanar points

A well known result from projective geometry is that groups of five coplanar points give rise to two *projective invariants* [168]. More specifically, two functions of five coplanar points can be constructed, whose values do not change under perspective viewing (i.e. under the application of arbitrary plane homographies). In other words, the invariants

are identical for every quintuple of corresponding points and are expressed by

$$I_1 = \frac{|\mathbf{M_{124}}||\mathbf{M_{135}}|}{|\mathbf{M_{134}}||\mathbf{M_{125}}|} \quad , \quad I_2 = \frac{|\mathbf{M_{241}}||\mathbf{M_{235}}|}{|\mathbf{M_{234}}||\mathbf{M_{215}}|}, \tag{3.1}$$

where $|\mathbf{M_{ijk}}|$ denotes the determinant of the matrix whose columns are the vectors $\mathbf{m_i}, \mathbf{m_j}, \mathbf{m_k}$ formed by the homogeneous coordinates of three image points, i.e. $\mathbf{M_{ijk}} = (\mathbf{m_i}, \mathbf{m_j}, \mathbf{m_k})$. Note that both I_1 and I_2 degenerate when any three of the five points are collinear. Each of the two invariants defined above, corresponds to the cross-ratio [168, 128] of a pencil of four lines, which is constructed by connecting one of the five points with each of the remaining four. To prove that the quantities defined by Eq. (3.1) are indeed invariant under any plane homography $\mathbf{H}$, assume that $\mathbf{H}$ transforms point $\mathbf{m_i}$ to point $\mathbf{m_i'}$, i.e. $m_i' = \lambda_i \mathbf{H m_i}$, where λ_i is an unknown scale factor. The quantity corresponding to I_1 for the transformed points is equal to

$$\begin{aligned} I_1' = \frac{|\mathbf{M_{124}}'||\mathbf{M_{135}}'|}{|\mathbf{M_{134}}'||\mathbf{M_{125}}'|} &= \frac{|(\lambda_1\mathbf{Hm_1}, \lambda_2\mathbf{Hm_2}, \lambda_4\mathbf{Hm_4})||(\lambda_1\mathbf{Hm_1}, \lambda_3\mathbf{Hm_3}, \lambda_5\mathbf{Hm_5})|}{|(\lambda_1\mathbf{Hm_1}, \lambda_3\mathbf{Hm_3}, \lambda_4\mathbf{Hm_4})||(\lambda_1\mathbf{Hm_1}, \lambda_2\mathbf{Hm_2}, \lambda_5\mathbf{Hm_5})|} \\ &= \frac{\lambda_1^2\lambda_2\lambda_3\lambda_4\lambda_5}{\lambda_1^2\lambda_2\lambda_3\lambda_4\lambda_5}\frac{|\mathbf{H}||\mathbf{M_{124}}||\mathbf{M_{135}}|}{|\mathbf{H}||\mathbf{M_{134}}||\mathbf{M_{125}}|}, \end{aligned}$$

and thus is equal to I_1. The proof for the invariance of I_2 can be obtained in a similar manner.

To test whether a set of five points imaged in two views satisfy the above invariants, a statistical test based on the variance in the values of the invariants is employed. This variance is estimated from the variances in the positions of the points in an image. The reader is referred to [225] for more details.

3.2.2 Estimation of the plane homography

Assuming a set of N pairs of corresponding coplanar points, the plane homography $\mathbf{H}$ that they define can be estimated as follows: Equation (2.20) provides $2 * N$ constraints regarding the elements of the homography matrix, which can be written more compactly

as $\mathbf{Ah} = 0$, where $\mathbf{A}$ is a $(2 * N) \times 9$ matrix and $\mathbf{h}$ a 9×1 vector. The plane homography is then estimated from the solution of the following minimization problem:

$$min_{\mathbf{h}} ||\mathbf{Ah}||^2 \quad subject\ to \quad ||\mathbf{h}||^2 = 1, \tag{3.2}$$

where $||\ ||$ denotes the vector 2-norm. As shown in appendix B, the solution to this problem is the eigenvector of the matrix $\mathbf{A}^T\mathbf{A}$ that corresponds to the smallest eigenvalue, where $\mathbf{A}^T$ denotes the transpose of $\mathbf{A}$. Similar to what noted in [97, 291], $\mathbf{A}^T\mathbf{A}$ is inhomogeneous in image coordinates, and, therefore, ill-conditioned. To improve its condition number and to derive a more stable linear system, the coordinates of the set of corresponding points are normalized by a pair of linear transformations $\mathbf{L}$ and $\mathbf{L}'$ as follows: $\mathbf{L}$ defines a translation of the points in the first image, such that their centroid is brought to the origin of the coordinate system, followed by an isotropic scaling that maps the average point coordinates to $(1, 1, 1)$. $\mathbf{L}'$ is defined similarly for points in the second image. These transforms result in a more stable system, from which a homography matrix $\hat{\mathbf{H}}$ can be estimated. $\mathbf{H}$ is then computed from $\hat{\mathbf{H}}$ as $\mathbf{L}'^{-1}\hat{\mathbf{H}}\mathbf{L}$.

Since the set of normalized matching pairs that is given as input to the estimation process is very likely to contain errors, care must be taken so that these errors do not corrupt the computed estimate. Thus, instead of using all N points to estimate $\mathbf{H}$, the LMedS estimator is employed to find an estimate that is consistent with the majority of the matched points. Using a predetermined number of iterations, LMedS picks random samples of matching pairs and computes an estimate of $\mathbf{H}$ from each of them. The estimate that yields the smallest median error is returned as the plane homography which best fits the set of matched points.

3.2.3 Iterative algorithm for the extraction of planes

Based on the above discussion, an iterative method for extracting the dominant plane can now be described. First, the SUSAN corner detector [231] is used to extract a set of corners from a pair of images. Corners are distinct image features that can be

accurately localized and correspond to 3D scene elements appearing in consecutive images. Here it is assumed that the two images have been acquired from considerably different locations in the 3D space. Such an image pair can be captured either by the two cameras of a binocular system, or by the single camera of a monocular system at two instants that are far apart in time. The extracted corners are then matched using a similarity criterion based on normalized cross-correlation. The matching algorithm is based on that proposed in [290, 294]. A random sample consisting of five pairs from the set of matched corners is then formed. If the selected corners satisfy the invariants in Eq. (3.1), they are likely to belong to the same plane. To ensure that the selected corners are sufficiently far apart so that the invariants and the corresponding plane homography are not swamped by noise, a bucket-based sampling technique similar to that discussed in [291] is employed. Next, the plane homography corresponding to the selected corners is computed as described previously. To verify that the five selected points lie on the same plane, the estimated plane homography is used to find more coplanar points. For every corner in one image, the plane homography can predict the location of the corresponding corner in the second image. If this location is sufficiently close to the true location of the matching corner, the corner in question is assumed to be coplanar with the corners in the selected sample. If the number of coplanar points identified during this step is above a threshold, the method concludes that a plane has indeed been found. The corresponding plane homography is then refined using the LMedS estimator over the whole set of matched coplanar points and this set is removed from further consideration. The sampling process iterates until either the number of corners that have not been assigned to a plane drops below a threshold or a predetermined number of iterations is completed.

When the iterative algorithm terminates, a set of planes along with their homographies have been computed. The application of Eq. (2.20) to each point in the first view warps the second view with respect to the first and registers the image of the corresponding plane in the two views. Change detection between the first and the warped second view can label image points as changing in the two views or not.

Points that remain unchanged belong to the plane under consideration. To account for the fact that typical change detection algorithms fail in uniform, textureless areas, a pixel is assumed to belong to a plane when it is labeled as not changing by the change detection algorithm and the magnitude of its intensity gradient is above some threshold. The plane having the largest number of points is declared to be the dominant one. As it will be clear from the following sections, the result of change detection does not have to be very accurate, since the part of the proposed method for IMD that makes use of the location of the dominant plane is tolerant to errors. In our implementation, the change detection algorithm described in [230] is employed. This algorithm is based on a test regarding the variance of the intensity ratios in small neighborhoods of two images.

3.3 Robust Parametric Estimation of Optical Flow

The problem of estimating 2D image velocity, or optical flow, from image sequences is generally very difficult. This difficulty mainly stems from the fact that transparencies, specular reflections, shadows, occlusions, depth boundaries and independent motions give rise to discontinuities in the optical flow field [165]. This in turn implies that an optical flow field is typically only piecewise smooth [34]. Since the estimation of optical flow involves the combination of constraints arising from an image region, no guarantee is given that the selected region will contain only a single motion. In other words, the primary difficulty of most optical flow estimation techniques is that they lack any information regarding the region of support of a particular motion. This problem is referred to in [34] as the *generalized aperture problem*.

In the case that an image region is known to correspond to a plane in the scene, the optical flow within the region can be accurately modeled as a parametric function of the image coordinates [2]. More specifically, assuming that the equation of the imaged plane in image coordinates is $\frac{1}{Z} = px + qy + r$, substitution into Eq. (2.7) yields an eight parameter model for optical flow. This model is known as the *quadratic* model since it

contains terms that are of degree two in the image coordinates:

$$u = a + bx + cy + gx^2 + hxy$$
$$v = d + ex + fy + gxy + hy^2 \tag{3.3}$$

At this point, it should be noted that, if the camera is not calibrated, the unknown intrinsic parameters (i.e. focal lengths and location of principal point) are absorbed in the eight parameters $a, \ldots, h$. By employing the quadratic model, the estimation of optical flow amounts to the estimation of the eight parameters involved. Substitution of Eq. (3.3) into Eq. (2.10), permits the derivation of a model relating the planar flow parameters to the spatiotemporal intensity derivatives. This model is linear in the parameters to be estimated and is overdetermined, since each point of the plane contributes one constraint regarding the eight unknown parameters. To account for errors in the computation of derivatives, violations of the intensity conservation assumption, errors in the determination of the region corresponding to the image of the plane, etc, the LMedS estimator is again employed to give a robust estimate of the parameters satisfying the majority of the constraints. This ''robustification'' of the optical flow estimation problem has already been suggested by Black and Anandan [34], with the major difference being that they employed M-estimators which are less robust compared to LMedS that is employed in this work.

3.4 The Residual Normal Flow Field with Respect to a Plane

Let (u, v) be the displacement field between two images $\mathcal{I}_t$ and $\mathcal{I}_{t+dt}$ acquired at time instants t and $t + dt$ respectively. Let also Π be a 3D plane in the viewed scene and let (u^π, v^π) be the 2D motion vector of a single point belonging to Π. As shown in section 3.3, (u^π, v^π) is defined by a linear model with eight parameters. As explained in appendix A, warping $\mathcal{I}_t$ towards $\mathcal{I}_{t+dt}$ according to (u^π, v^π) will register $\mathcal{I}_t$ and $\mathcal{I}_{t+dt}$ over regions of Π, while regions not belonging to Π will be unregistered. According to

Eq. (A.1), the residual flow (u^r, v^r) between the warped $\mathcal{I}_t$ and $\mathcal{I}_{t+dt}$ is equal to

$$\begin{aligned} u^r = \quad u - u^\pi \quad &= (xW - Uf)(\frac{1}{Z} - \frac{1}{Z^\pi}) \\ v^r = \quad v - v^\pi \quad &= (yW - Vf)(\frac{1}{Z} - \frac{1}{Z^\pi}) \end{aligned} \tag{3.4}$$

where $\frac{1}{Z^\pi}$ is the depth of the 3D plane at pixel (x, y). As can been seen from Eq. (3.4), the residual flow field is purely translational, since the rotational components have been canceled out by the warping step. It is also straightforward to show that the *residual normal flow field* between the warped $\mathcal{I}_t$ and $\mathcal{I}_{t+dt}$ is given by:

$$u_{nr} = \quad u^r n_x + v^r n_y \quad = \{(xW - Uf)n_x + (yW - Vf)n_y\}\,(\frac{1}{Z} - \frac{1}{Z^\pi}) \tag{3.5}$$

where (n_x, n_y) is the unit vector in the direction of the intensity gradient.

3.5 Using Residual Parallax Normal Flows to Detect Independent Motion

Consider a rigid observer that is moving with unrestricted egomotion in 3D space. Due to this motion, a reliable normal flow vector can be computed at each point where the image intensity gradient is sufficiently large. Let (n_x, n_y) be the unit vector in the gradient direction. The magnitude u_n of the normal flow vector is given by $u_n = un_x + vn_y$, which, by substitution from Eq. (2.7), yields Eq. (2.16) which is repeated here for convenience:

$$\begin{aligned} u_n \quad = \quad & -n_x f\frac{U}{Z} - n_y f\frac{V}{Z} + (xn_x + yn_y)\,\frac{W}{Z} \\ + \quad & \left\{\frac{xy}{f}n_x + \left(\frac{y^2}{f} + f\right)n_y\right\}\alpha \\ - \quad & \left\{\left(\frac{x^2}{f} + f\right)n_x + \frac{xy}{f}n_y\right\}\beta + (yn_x - xn_y)\gamma \end{aligned} \tag{3.6}$$

As it has been discussed in section 2.1.5, Eq. (3.6) shows that the problem of recovering 3D motion from a single normal flow field is underconstrained. This is because each

normal flow provides one constraint on the 3D motion parameters but also introduces one unknown variable corresponding to the depth Z. To overcome this difficulty, the remainder of this section employs a pair of residual parallax normal flow fields that provide additional constraints on the 3D motion parameters.

Let us begin by supposing that at least one of the surfaces in the scene is planar or can be well approximated by a plane. This assumption is often satisfied in practice, especially in scenes containing man-made objects [270]. Using the technique described in section 3.2, the dominant plane in the scene can be extracted. Following this, the parametric model describing the motion of this plane can be estimated as described in section 3.3. The residual planar parallax flow can then be computed from Eq. (3.4). Irani and Anandan [118] have recently described a method for IMD that computes the *relative projective 3D structure* from this residual parallax flow. Their method, however, requires the computation of a dense optical flow field, a difficult problem in its own right. Noting that the residual flow field is translational, another approach to detect independent motion is to locate the FOE and then, similar to [220], label points that violate the epipolar constraint as independently moving. The major drawback of this approach is that it depends critically on the correctness of the estimated FOE. To avoid this problem, the proposed method for IMD does not attempt to estimate the FOE. Instead, it combines the information from two residual normal flow fields computed at consecutive time instants.

Assume that three consecutive images $\mathcal{I}_{t-dt}$, $\mathcal{I}_t$ and $\mathcal{I}_{t+dt}$ are captured at time instants $t-dt$, t and $t+dt$ respectively. Let $\mathcal{I}_0$ be a fourth image that along with $\mathcal{I}_t$ permits the extraction of the dominant plane. Also, let u_{nr} be the residual normal flow computed by warping $\mathcal{I}_t$ towards $\mathcal{I}_{t+dt}$ using the motion of the dominant plane. Similarly, let u'_{nr} be the residual normal flow computed by warping $\mathcal{I}_t$ towards $\mathcal{I}_{t-dt}$ using the dominant plane. According to Eq. (3.4), u_{nr} and u'_{nr} are given by

$$u_{nr} = \{(xW - Uf)n_x + (yW - Vf)n_y\}\,(\frac{1}{Z} - \frac{1}{Z^\pi}) \tag{3.7}$$

$$u'_{nr} = \{(xW' - U'f)n_x + (yW' - V'f)n_y\}(\frac{1}{Z} - \frac{1}{Z^\pi})$$

where (U, V, W) and (U', V', W') are the 3D translational velocity vectors for the displacement between t and $t + dt$ and t and $t - dt$ respectively. Notice that here it is implicitly assumed that the translational component of the observer's velocity is nonzero.

Both residual normal flow fields given by Eqs. (3.7) are defined in the the same reference frame, namely $\mathcal{I}_t$. This implies that at each point (x, y) of $\mathcal{I}_t$ having considerable gradient magnitude, two normal flow vectors along the same direction (n_x, n_y) can be computed. Solving the first of Eqs. (3.7) for $\frac{1}{Z} - \frac{1}{Z^\pi}$ and substituting into the second results into the following equation

$$\begin{aligned} W(xn_x + yn_y)u'_{nr} - {} & Ufn_x u'_{nr} - Vfn_y u'_{nr} - \\ & W'(xn_x + yn_y)u_{nr} + U'fn_x u_{nr} + V'fn_y u_{nr} = 0, \end{aligned} \tag{3.8}$$

in which the terms related to depth have been eliminated. The above equation is linear in the variables $\phi_1 = W$, $\phi_2 = Uf$, $\phi_3 = Vf$, $\phi_4 = W'$, $\phi_5 = U'f$, $\phi_6 = V'f$. These variables involve the 3D motion parameters and the camera focal length. Assuming that the dominant plane is not independently moving, violations of Eq. (3.8) signal the presence of independently moving objects. LMedS estimation can be applied to a set of observations of the model of Eq. (3.8) as a means to estimate the parameters ϕ_i, $i = 1, \ldots, 6$. To avoid the trivial solution $\phi_i = 0$, the solutions tried by LMedS are computed with an eigenvector technique that imposes the constraint $||(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6)||^2 = 1$ (see appendix B). LMedS will provide estimates $\hat{\phi}_i$ of the parameters ϕ_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 at least one of the following holds:

1. The quantities u_{nr} and/or u'_{nr} 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 vectors 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. 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 with 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 inliers will correspond to the dominant motion (egomotion) and the set of outliers will contain all secondary (independent) 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 be terminated when the remaining points become fewer than a certain threshold. There are two reasons for this [14]. First, if the number of points becomes too small, then the number of constraints provided by Eq. (3.8) 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 flows have not been computed accurately. The proposed algorithm for IMD is summarized in the block diagram of Fig. 3.1. The postprocessing step is described in the following subsection.

When implementing the method presented in the preceding paragraphs, the residual normal flow can be computed without actually warping the first image towards the second according to the estimated planar flow. Knowledge of the eight parameters in Eq. (3.3) enables the prediction of the normal flow that would result if the dominant

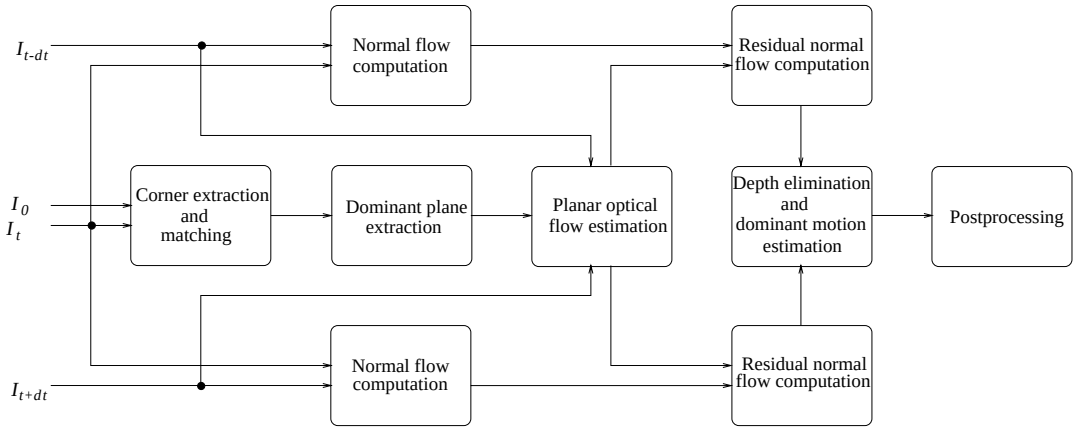


Figure 3.1: Block diagram of the proposed method (see text for explanation).

plane covered the whole visual field. The residual normal flow can then simply be estimated as the difference between the normal flow computed directly from the pair of input images and the predicted planar normal flow. Normal flow between a pair of input images is computed from the spatiotemporal derivatives I_x, I_y and I_t of the image intensity function. To reduce the effects of noise, images are smoothed by convolution with a 3×3 Gaussian prior to the computation of derivatives.

3.5.1 Postprocessing

According to the proposed method for independent motion detection, points are characterized as being independently moving or not based on their conformance to a general rigid 3D model of egomotion. 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, it is often considered preferable to refer to connected, independently moving areas rather than to isolated points. There are three reasons why the points of a motion segment may not form connected regions [14]. First, the normal flow field is usually a sparse field, because normal flow values are considered unreliable in certain cases (e.g. at points with a small gradient value). Second, there is always

the possibility of errors in measurements of normal flow and, therefore, some points may become model inliers (or outliers) because of these errors and not due to their 3D motion parameters. Finally, normal flow is a projection of the optical flow onto a certain direction. Infinitely many other optical flow vectors have the same projection onto this direction. Consequently, a normal flow vector may be 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. Therefore, 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.

3.6 Experimental Results

The proposed method has been evaluated experimentally with the aid of several real-world image sequences. During the course of all experiments, quantitative information regarding camera motion and calibration parameters was not available. This section reports two of the conducted experiments.

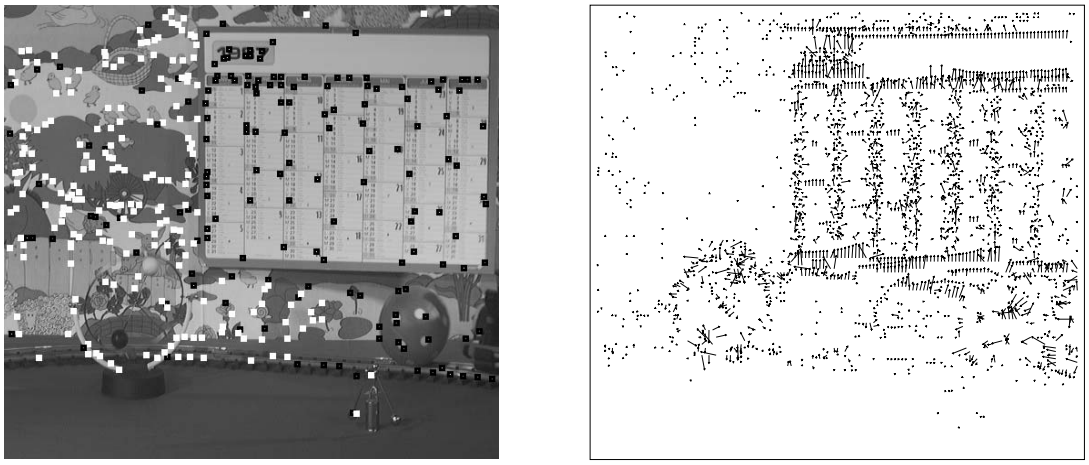
The first experiment is based on the well known ‘‘calendar’’ image sequence. Frames 2 and 30 of this sequence are shown in Fig. 3.2. In this sequence, the camera is panning with a right to left direction and the viewed scene consists of a planar background and a nonplanar foreground. The background contains a stationary wall and a calendar that is independently moving upwards. The foreground contains three

(a) (b)

Figure 3.2: Frames (a) 2, and (b) 30 of the ‘‘calendar’’ sequence.

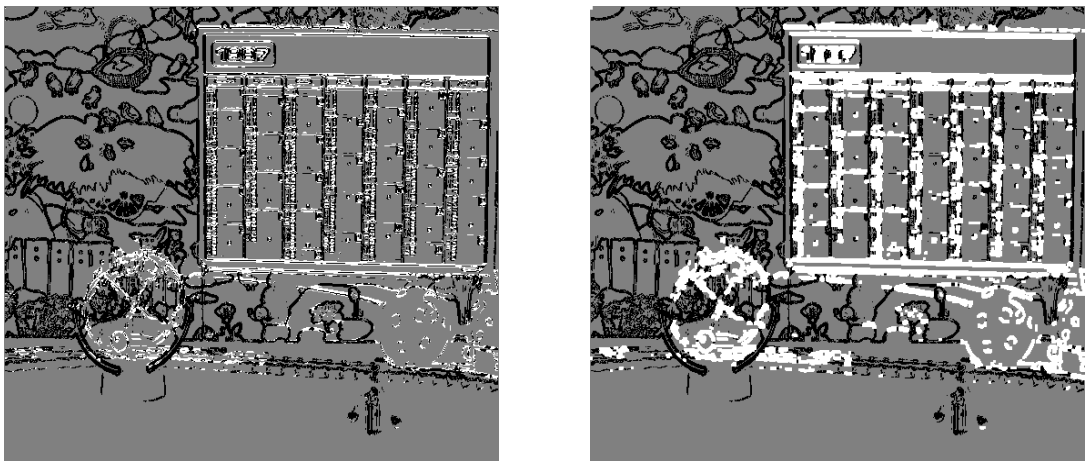
independently moving objects. A pair of spheres is rotating in the left side of the scene, while a ball followed by a toy train are moving in a right to left direction. The dominant plane was extracted using frames 2 and 30. Corners belonging to the dominant plane are marked with white rectangles in Fig. 3.3(a), while all other corners are black.

The pair of residual parallax normal flow fields is computed between frames 2 - 3 and 2 - 1. The residual parallax normal flow for frames 2 - 3 is shown in Figure 3.3(b). As can be seen from this figure, the residual flow field is zero over the area corresponding to the dominant plane, indicating that the dominant plane has been successfully registered. Figure 3.4 illustrates the results of motion segmentation on the ‘‘calendar’’ sequence. Figure 3.4(a) shows the intermediate segmentation results. Black color corresponds to egomotion and white color 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 indeed over the regions of the independently moving objects. Note that independent motion was not detected along the vertical edges of the calendar. This is because the intensity gradient is perpendicular to the direction of motion on these edges, which results in the corresponding normal flow vectors being equal to zero. The elongated areas below the calendar that are marked as independently



(a) (b)

Figure 3.3: (a) Corners belonging to the dominant plane for the ‘‘calendar’’ sequence, (b) residual normal flow field for frames 2-3.



(a) (b)

Figure 3.4: Motion segmentation for the ‘‘calendar’’ sequence (a) before and, (b) after postprocessing.

moving are actually shadows, cast by the calendar and the rotating spheres, that are moving during time. Figure 3.4(b) presents the same result after postprocessing, which eliminates isolated outliers (inliers) in large populations of inliers (outliers) and, in the resulting map, dilates the label of remaining outliers in a small neighborhood. It is clear that after this step, the bodies of the four independently moving objects have been successfully identified as such. An MPEG video demonstrating the results of applying the proposed method on the first 10 frames of the ‘‘calendar’’ sequence can be found at http://www.ics.forth.gr/proj/cvrl/demos/lourakis/IMD/calendar.mpg

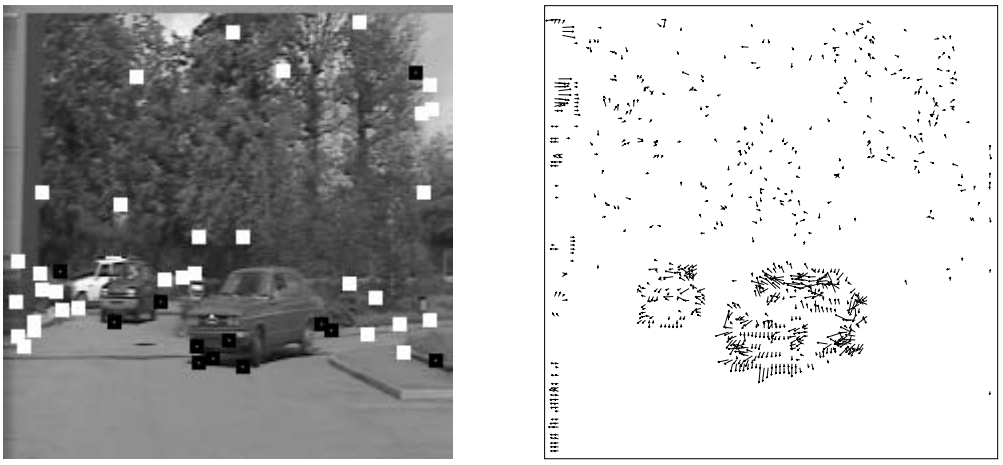
The second experiment concerns the ‘‘cars’’ image sequence. Frames 5 and 20 of this sequence are shown in Fig. 3.5.

(a) (b)

Figure 3.5: Frames (a) 5, and (b) 20 of the ‘‘cars’’ sequence.

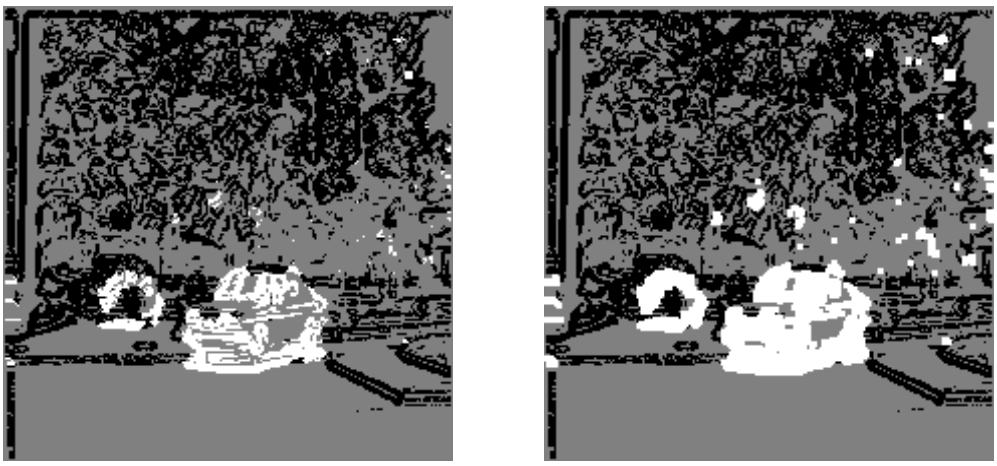
In this sequence, the camera is again panning with a right to left direction. The two dark gray cars in the foreground move independently while the white car on the far left is stationary. A few trees in the background form an approximately planar surface. Frames 5 and 20 were used to extract the dominant plane. Figure 3.6(a) shows corners belonging to the dominant plane marked with white rectangles, while all other corners are black.

Frames 5 - 6 and 5 - 4 are used to compute the pair of residual parallax normal



(a) (b)

Figure 3.6: (a) Corners belonging to the dominant plane for the ‘‘cars’’ sequence, (b) residual normal flow field for frames 5-6.



(a) (b)

Figure 3.7: Motion segmentation for the ‘‘cars’’ sequence (a) before and, (b) after postprocessing.

flow fields. Figure 3.6(b) shows the residual parallax normal flow computed from frames 5 - 6. The results of motion segmentation on the ‘‘cars’’ sequence before and after postprocessing are illustrated in Figures 3.4(a) and 3.7(b) respectively. Black color corresponds to egomotion and white color corresponds to independent motion. Gray color corresponds to points with low intensity gradient, and thus without normal flow vectors. As can be seen from Fig. 3.7, the two cars are correctly identified as independently moving. Moreover, the independent motions of small parts of the tree foliage are also detected.

3.7 Summary

Artificial seeing systems should operate in dynamic environments that consist of both stationary as well as moving objects. The perception of independent 3D motion is crucial because it provides useful information on where attention should be focused and, possibly, maintained. In this work, independent 3D motion detection was based on a pair of residual parallax normal flow fields that are computed by an observer that moves freely in the 3D space. The proposed method employs 3D motion models and is able to perform satisfactorily even in scenes with considerable depth variations. Both rigid and non-rigid independent motion can be detected. Moreover, apart from the requirement for the existence of a stationary planar surface in the viewed scene, no further assumptions regarding the structure of the external world are made. The method avoids a complete solution to the ill-posed correspondence problem by matching only carefully selected sets of image points. To guard against errors caused by false matches, robust estimation techniques are employed. Experimental results from the application of the proposed method on real image sequences were also presented.

Chapter 4

Egomotion Estimation

4.1 Introduction

Knowledge of the velocity of a mobile system with respect to its environment is essential for various servoing tasks that are based on visual feedback. Such tasks include collision avoidance, docking, gaze maintenance, etc. Given a sequence of images acquired by a monocular observer pursuing unrestricted rigid motion, the problem of egomotion estimation can be stated as the problem of recovering the translation and rotation comprising the motion of the observer. Although simply stated, the problem of estimating egomotion using visual input is particularly difficult. This difficulty primarily stems from the fact that the only information available from images is related to the 2D motion of image points, while the sought egomotion is a 3D quantity. The observed 2D motion depends not only on the egomotion, but also on the unknown structure of the viewed scene. Since the dependence of the 2D motion on the scene structure is nonlinear, small errors in the estimates of 2D motion can have a significant impact on the accuracy of the recovered 3D motion [65, 248]. In addition, the confounding of translation and rotation makes the problem of estimating unrestricted egomotion much harder compared to the problem of estimating pure translation or rotation.

Depending on the frequency of time sampling during the acquisition of an image sequence, egomotion estimation algorithms can be subdivided into two broad categories. The algorithms that belong to the first category assume an infinitesimal time sampling period and employ vector fields to model the 2D motion of image points. The second category includes algorithms that assume coarse time sampling and use sparse displacement maps to describe the 2D motion of isolated features extracted from the images. Although algorithms in the second category make less assumptions regarding the image acquisition process, algorithms in the first category are more popular. This is mainly due to the separability of the translational and rotational components in the 2D motion equations and the fact that, in principle at least, the problem of correspondence establishment is easier in the case of small motions. A typical assumption that is implicitly made by most algorithms is that the viewed scene is static, i.e. there are no objects moving independently from the observer. Since the egomotion estimation algorithm proposed here assumes fine time sampling, we focus our review on algorithms in the first of the categories defined above. Owing to the inherent scale ambiguity that characterizes visual motion (see section 2.1.2), the information regarding the translational component of egomotion that can be recovered by all these algorithms is at most the direction of 3D translation, i.e. the FOE.

We start by reviewing algorithms that rely on the availability of a dense optical flow field to describe 2D motion. Longuet-Higgins and Prazdny [144] and Reiger and Lawton [205] solve for translation by exploiting a phenomenon known as *motion parallax*: A pair of 3D points projecting to nearby retinal locations but having different depths, have almost the same rotational motion. This approximation is exact when the two 3D points project to the same retinal location, as is the case with a transparent surface. Reiger and Lawton showed that the approximation is still valid when the depth difference between the 3D points is large. Thus, subtracting the optical flow vectors at two appropriate image locations yields a flow vector that is approximately pointing towards the FOE. After recovering the translation, rotation can be estimated with linear regression on projections of the optical flow that are perpendicular to lines through the FOE. The main

drawback of these approaches stems from the fact that most optical flow algorithms cannot give accurate estimates of optical flow in areas with large depth variations. Recently, Irani et al [120] alleviated some of the difficulties related to the estimation of motion parallax by decomposing image motion into the sum of the motion of a planar surface and a residual parallax field that is purely translational. Prazdny [199] showed that the difference between any pair of flow vectors gives a constraint on translation but did not develop an algorithm exploiting this constraint.

In earlier work, Prazdny [197] assumes that surfaces in the viewed scene are smooth and solves for rotation using a set of nonlinear equations that are independent of translation. This nonlinear system is solved by numerical optimization techniques. Apart from the smoothness assumption, this method suffers from high computational costs. Prazdny [198] and later Burger and Bhanu [45] also suggested solving for rotation first and employed a search in the space of rotational parameters. For each hypothesized rotation, the corresponding rotational field was subtracted from the optical flow and the remaining field was tested for how well it approximated a purely translational flow field. Ballard and Kimball [25] assumed that the depth of the viewed scene is known and employed a generalized Hough transform to solve for the 3D motion parameters. An advantage of this approach is that multiple moving objects give rise to multiple peaks in the Hough space. This method, however, is difficult to apply in practice since the depth of the scene is usually unknown and the search through the multidimensional solution space is very expensive computationally. Bruss and Horn [44] combine information from the whole visual field to determine the 3D motion that is the best least squares fit to the observed velocity field. They developed three different algorithms. The first two algorithms give closed form solutions for translation and rotation, when the motion is purely translational or purely rotational respectively. The third algorithm applies in the case of general motion and provides a residual function that involves the unknown translation only. Translation is then found by minimizing this residual function using iterative numerical procedures. Adiv [2] utilized the same residual function developed by Bruss and Horn, but suggested an alternative scheme for minimizing it. He subdivides

the flow field in patches and estimates the 3D motion of each patch independently. Iterative minimization of the residual function is achieved by sampling the solution space of all possible candidate translations and declaring the direction with the smallest residual function as the correct one. Patches that share the same 3D motion are then merged, since they belong to objects undergoing the same rigid motion. Thus, Adiv's algorithm is capable of handling independently moving objects. Waxman and Subbarao [237, 276] employ local estimates of flow velocities and their derivatives up to second order for determining motion parameters and local surface structure. Apart from using only local constraints, this approach has the drawback that the optical flow derivatives it utilizes are extremely sensitive to noise. Heeger and Jepson [100] also make use of the residual function used in [44, 2] and propose an efficient technique for locating its minimum. The space of all possible translation directions is again sampled and the residual function is evaluated as the linear sum of the flow vectors weighted by a set of coefficients that have been computed off-line.

Hummel and Sundareswaran [115] present an algorithm for finding the rotational motion and one for locating the FOE. The first algorithm, known as the *flow circulation* algorithm, computes the curl of the optical flow field. It is based on the observation that curl is approximately a linear function whose coefficients are proportional to the desired rotational parameters of motion. The algorithm for locating the FOE extends the work of Heeger and Jepson [100]. For each candidate FOE, the *circular component* field, defined as the projection of the optical flow along vectors emanating from the hypothesized FOE, is computed. The circular component field that corresponds to the true FOE is a quadratic function of a special form. Three different techniques that can be formulated as quadratic functionals of the observed circular component data are proposed for determining the point having the appropriate quadratic form. MacLean et al [154] combine subspace methods with a finite mixture model and apply the EM algorithm to cluster constraints on the 3D velocity. Then, the results of clustering provide an initial guess for solving for the parameters of the different 3D motions present in the scene. Da Vitoria Lobo and Tsotsos [141] develop a constraint (the *Collinear Point Constraint - CPC*) involving

three collinear image points, which provides a means for canceling rotation and at the same time constraining the FOE to lie on the line defined by the collinear points. The FOE is defined as the intersection of lines defined by triplets of collinear points that satisfy the constraint. CPC is discussed in more detail in section 4.2. Rousso et al [211] show that rotation can be computed from any three homography matrices. To compute the homographies, they employ the trilinear tensor defined by three frames. Daniilidis [62] employs fixation on a scene point to reduce the number of motion parameters to be estimated from five to four. The spherical motion field is projected on two latitudinal directions, effectively decoupling the motion parameter space. The motion parameters are then found by two one-dimensional searches along meridians of the image sphere. Fejes and Davis [76] also deal with the egomotion estimation problem by employing projections of the flow field in various directions. These projections exhibit simple geometric properties, independent of the scene structure, and are combined with the aid of a recursive filter to yield the motion parameters.

To avoid problems related with the computation of optical flow [165], the so-called *direct* paradigm to egomotion estimation has emerged. Instead of employing the full optical flow field, its projections in various directions are used. These projections, given by the spatiotemporal derivatives of the image intensity function, are easier to compute than the full flow. Direct methods were first introduced by Horn and his associates [109, 176, 175] and solve the egomotion problem in the case of translation only. Aloimonos and Brown [5] address the case of an observer pursuing purely rotational motion. They estimate rotation by exploiting the fact that in this case the motion equations are linear in the rotational parameters of egomotion and do not involve the scene structure. Nelson and Aloimonos [179] assume a spherical retina and show that the spherical motion field has a focus of expansion and a focus of contraction separated by 180 degrees if and only if the rotational component of motion is zero. They also show how the problem of determining the motion parameters can be separated into three two dimensional problems, which allow the rotation parameters to be determined independently. The direction of translation can then be found from the

vector joining the foci of expansion and contraction. Hanna [94] assumes that the viewed scene can be locally approximated by planes and presents a direct iterative method for recovering egomotion and scene structure at multiple resolutions. Taalebizhaad [243] exploits the fact that fixation on a scene point reduces by one the dimensionality of the 3D motion recovery problem, and suggests a method for direct recovery of unrestricted egomotion. Aloimonos and Duric [8] assume a translating observer and present a qualitative algorithm that uses a voting scheme based on the signs of optical flow projections to locate the FOE. A similar method was independently developed by Sinclair et al [226]. Fermüller [79] addresses the case of unrestricted egomotion and bases egomotion estimation on the geometrical properties of the normal flow field. The signs of optical flow projections give rise to simple patterns on the image plane, which depend on the egomotion parameters. However, although her method can be employed to verify the correctness of a given set of motion parameters, it cannot be used for making a hypothesis regarding them. In addition, the extraction of appropriate patterns is made difficult by the fact that she employs sparse motion fields. Silva and Santos-Victor [222] assume unrestricted 3D motion and locate the FOE as the intersection of two constraint lines. Despite the fact that their method depends critically on the accuracy of the recovered constraint lines, they do not present any results indicating the behavior of their algorithm in the presence of different amounts of noise.

In this work, a new method for egomotion estimation is presented. The motivation behind our effort is twofold. First, we are interested in estimating egomotion by means of linear constraints. Second, we want to avoid making any restrictive assumptions regarding the egomotion or the scene structure. Hence, we have developed a novel linear constraint regarding the motion parameters, defined in terms of four collinear image points. The constraint is applicable regardless of the egomotion or the scene structure and combined with robust linear regression techniques, permits the recovery of the direction of translation, thereby decoupling the 3D motion parameters.

The rest of this chapter is organized as follows. Section 4.2 develops the proposed

constraint and shows how it can be employed to recover egomotion. Experimental results from an implementation of the method are presented in section 4.3. The chapter is concluded with a brief discussion in section 4.4.

4.2 Using Quadruples of Collinear Points to Constrain the FOE

Before proceeding to the description of the proposed method, we state two lemmas which are essential for its derivation. The symbolic calculations that are reported in the following have been carried out with the aid of the MATHEMATICA symbolic mathematics package [283].

4.2.1 Two precursory lemmas

Lemma 4.1 *Suppose that two image points lie on a line that goes through the origin of the image coordinate system (i.e. the principal point). The difference of the projections of their corresponding optical flow vectors along the direction that is normal to the line does not depend on the α and β components of rotation.*

Proof. Let $\mathbf{p_1} = (x_1, y_1)$ and $\mathbf{p_2} = (x_2, y_2)$ be two points in the image and $\vec{\mathbf{n}} = (n_x, n_y)$ be the unit vector that is normal to the line $\mathcal{L}$ defined by $\mathbf{p_1}$ and $\mathbf{p_2}$. Since $\mathcal{L}$ goes through the image principal point, its equation is $y = -\frac{n_x}{ny}x$, and therefore Theorem C.1 from appendix C yields for $\nu = 0$

$$un_1 - un_2 = [(x_1 - x_0)n_x + (y_1 - y_0)n_y]W(\frac{1}{Z_1} - \frac{1}{Z_2}) + \frac{\gamma}{n_y}(x_2 - x_1) \tag{4.1}$$

□

Lemma 4.2 *Let* $\mathbf{p_1} = (x_1, y_1)$, $\mathbf{p_2} = (x_2, y_2)$ *and* $\mathbf{p_3} = (x_3, y_3)$ *be three collinear image points lying on a line whose equation is* $y = \kappa x + \nu$. *Let also* (x_0, y_0) *be the FOE and assume that* $\mathbf{p_2}$ *divides the vector* $\overrightarrow{\mathbf{p_1}\,\mathbf{p_3}}$ *in ratio* λ. *For the projections* $un_i, i = 1 \ldots 3$ *of the optical flow vectors at points* $\mathbf{p_1}$, $\mathbf{p_2}$ *and* $\mathbf{p_3}$ *along a direction* (n_x, n_y), *the following equation holds*

$$un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3 = D_2 W(\frac{1}{Z_2} - \frac{1}{1+\lambda}\frac{1}{Z_1} - \frac{\lambda}{1+\lambda}\frac{1}{Z_3}) + \frac{d_{21}}{1+\lambda}W(\frac{1}{Z_1} - \frac{1}{Z_3}) + \frac{\kappa d_{21}(x_2 - x_3)}{f}\alpha - \frac{d_{21}(x_2 - x_3)}{f}\beta \quad (4.2)$$

In the above equation, $D_2 = (x_2 - x_0)n_x + (y_2 - y_0)n_y$ *and* $d_{21} = (x_2 - x_1)n_x + (y_2 - y_1)n_y$.

Proof. The desired result follows directly from Theorem C.2 in appendix C.

□

By inspecting Eq. (4.2), it can easily be seen that in the case that the direction of projection (n_x, n_y) is perpendicular to the line defined by the points $\mathbf{p_i}$, the term d_{21} is zero, thus the sum of the rotational components vanishes. The remaining terms express the Collinear Point Constraint (CPC), which has been previously derived in [141]. CPC states that when an appropriate linear combination of the projections of optical flow vectors in the direction perpendicular to the line joining them is zero, there exist two possible situations. Either the three 3D points whose projections form the collinear triplet are also collinear in the scene (i.e. $\frac{1}{Z_2} - \frac{1}{1+\lambda}\frac{1}{Z_1} - \frac{\lambda}{1+\lambda}\frac{1}{Z_3} = 0$), or the line defined by the collinear triplet passes through the FOE (i.e. $D_2 = 0$). By employing a voting scheme to differentiate between these two cases, Da Vitoria Lobo and Tsotsos have exploited the CPC for locating the FOE [141].

4.2.2 The proposed constraint on egomotion

Assume now a mobile observer undergoing rigid motion in a static environment. Let $\mathbf{p_1} = (x_1, y_1)$, $\mathbf{p_2} = (x_2, y_2)$ and $\mathbf{p_3} = (x_3, y_3)$ be three collinear image points lying on a line $\mathcal{L}$ through the image principal point. Let also $y = \kappa x$ be the equation of line $\mathcal{L}$, (n_x, n_y) be the direction normal to it and $(n_x^{'}, n_y^{'})$ and $(n_x^{''}, n_y^{''})$ two directions that are not perpendicular to $\mathcal{L}$. According to Lemma 4.2, for the projections of the optical flow vectors along the direction $(n_x^{'}, n_y^{'})$ the following holds

$$un_2^{'} - \frac{1}{1+\lambda}un_1^{'} - \frac{\lambda}{1+\lambda}un_3^{'} = D_2^{'}W(\frac{1}{Z_2} - \frac{1}{1+\lambda}\frac{1}{Z_1} - \frac{\lambda}{1+\lambda}\frac{1}{Z_3}) + \frac{d_{21}^{'}}{1+\lambda}W(\frac{1}{Z_1} - \frac{1}{Z_3}) + (\kappa\alpha - \beta)\frac{d_{21}^{'}(x_2 - x_3)}{f}, \quad (4.3)$$

where the primed terms are defined analogously to the unprimed ones in Eq. (4.2). Similarly, for the projections along the normal direction (n_x, n_y), Eq. (4.2) gives

$$un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3 = D_2W(\frac{1}{Z_2} - \frac{1}{1+\lambda}\frac{1}{Z_1} - \frac{\lambda}{1+\lambda}\frac{1}{Z_3}) \quad (4.4)$$

Dividing Eq. (4.3) with Eq. (4.4) yields

$$\frac{un_2^{'} - \frac{1}{1+\lambda}un_1^{'} - \frac{\lambda}{1+\lambda}un_3^{'}}{un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3} = \frac{D_2^{'}}{D_2} + \frac{d_{21}^{'}}{1+\lambda}\frac{\frac{1}{Z_1} - \frac{1}{Z_3}}{D_2(\frac{1}{Z_2} - \frac{1}{1+\lambda}\frac{1}{Z_1} - \frac{\lambda}{1+\lambda}\frac{1}{Z_3})} + (\kappa\alpha - \beta)\frac{d_{21}^{'}(x_2 - x_3)}{f}\frac{1}{un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3} \quad (4.5)$$

Applying Eq. (4.1) for points $\mathbf{p_1}$ and $\mathbf{p_3}$ results in

$$un_1 - un_3 = D_2W(\frac{1}{Z_1} - \frac{1}{Z_3}) + \frac{\gamma}{n_y}(x_3 - x_1) \quad (4.6)$$

Solving Eq. (4.6) for $\frac{1}{Z_1} - \frac{1}{Z_3}$ and dividing in terms by Eq. (4.4) gives

$$\frac{\frac{1}{Z_1} - \frac{1}{Z_3}}{D_2(\frac{1}{Z_2} - \frac{1}{1+\lambda}\frac{1}{Z_1} - \frac{\lambda}{1+\lambda}\frac{1}{Z_3})} = \frac{un_1 - un_3 - \frac{x_3 - x_1}{n_y}\gamma}{D_2(un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3)} \quad (4.7)$$

Substituting Eq. (4.7) into Eq. (4.5) yields

$$\frac{un_2^{'} - \frac{1}{1+\lambda}un_1^{'} - \frac{\lambda}{1+\lambda}un_3^{'}}{un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3}\frac{1}{d_{21}^{'}} = \frac{D_2^{'}/d_{21}^{'}}{D_2} + \frac{1}{1+\lambda}\frac{un_1 - un_3 - \frac{x_3 - x_1}{n_y}\gamma}{D_2(un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3)} + (\kappa\alpha - \beta)\frac{(x_2 - x_3)}{f}\frac{1}{un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3} \quad (4.8)$$

Let now $\mathbf{p_4} = (x_4, y_4)$ be a fourth point collinear with the triplet $\mathbf{p_1}$, $\mathbf{p_2}$ and $\mathbf{p_3}$ and such that point $\mathbf{p_2}$ divides the vector $\overrightarrow{\mathbf{p_1}\,\mathbf{p_4}}$ in ratio μ. Eq. (4.8) gives for the projections along the direction $(n_x^{''}, n_y^{''})$

$$\frac{un_2^{''} - \frac{1}{1+\mu}un_1^{''} - \frac{\mu}{1+\mu}un_4^{''}}{un_2 - \frac{1}{1+\mu}un_1 - \frac{\mu}{1+\mu}un_4}\frac{1}{d_{21}^{''}} = \frac{D_2^{''}/d_{21}^{''}}{D_2} + \frac{1}{1+\mu}\frac{un_1 - un_4 - \frac{x_4 - x_1}{n_y}\gamma}{D_2(un_2 - \frac{1}{1+\mu}un_1 - \frac{\mu}{1+\mu}un_4)} + (\kappa\alpha - \beta)\frac{(x_2 - x_4)}{f}\frac{1}{un_2 - \frac{1}{1+\mu}un_1 - \frac{\mu}{1+\mu}un_4} \quad (4.9)$$

Subtracting Eq. (4.9) from Eq. (4.8) results in

$$\frac{un_2^{'} - \frac{1}{1+\lambda}un_1^{'} - \frac{\lambda}{1+\lambda}un_3^{'}}{un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3}\frac{1}{d_{21}^{'}} - \frac{un_2^{''} - \frac{1}{1+\mu}un_1^{''} - \frac{\mu}{1+\mu}un_4^{''}}{un_2 - \frac{1}{1+\mu}un_1 - \frac{\mu}{1+\mu}un_4}\frac{1}{d_{21}^{''}} = \frac{D_2^{'}/d_{21}^{'} - D_2^{''}/d_{21}^{''}}{D_2} +$$

$$\frac{1}{D_2}(\frac{1}{1+\lambda}\frac{un_1 - un_3}{un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3} - \frac{1}{1+\mu}\frac{un_1 - un_4}{un_2 - \frac{1}{1+\mu}un_1 - \frac{\mu}{1+\mu}un_4}) +$$

$$\frac{\gamma}{D_2}(-\frac{1}{1+\lambda}\frac{x_3 - x_1}{n_y}\frac{1}{un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3} + \frac{1}{1+\mu}\frac{x_4 - x_1}{n_y}\frac{1}{un_2 - \frac{1}{1+\mu}un_1 - \frac{\mu}{1+\mu}un_4}) +$$

$$(\kappa\alpha - \beta)(\frac{x_2 - x_3}{f(un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3)} - \frac{x_2 - x_4}{f(un_2 - \frac{1}{1+\mu}un_1 - \frac{\mu}{1+\mu}un_4)}) \quad (4.10)$$

Noting that $\frac{x_1 - x_3}{1+\lambda} = x_2 - x_3$ and $\frac{x_1 - x_4}{1+\mu} = x_2 - x_4$, Eq. (4.10) can be rewritten as

$$\frac{un_2^{'} - \frac{1}{1+\lambda}un_1^{'} - \frac{\lambda}{1+\lambda}un_3^{'}}{un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3}\frac{1}{d_{21}^{'}} - \frac{un_2^{''} - \frac{1}{1+\mu}un_1^{''} - \frac{\mu}{1+\mu}un_4^{''}}{un_2 - \frac{1}{1+\mu}un_1 - \frac{\mu}{1+\mu}un_4}\frac{1}{d_{21}^{''}} = \frac{D_2^{'}/d_{21}^{'} - D_2^{''}/d_{21}^{''}}{D_2} +$$

$$\frac{1}{D_2}(\frac{1}{1+\lambda}\frac{un_1 - un_3}{un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3} - \frac{1}{1+\mu}\frac{un_1 - un_4}{un_2 - \frac{1}{1+\mu}un_1 - \frac{\mu}{1+\mu}un_4}) +$$

$$(\frac{\gamma f}{D_2 n_y} + \kappa\alpha - \beta)(\frac{x_2 - x_3}{f(un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3)} - \frac{x_2 - x_4}{f(un_2 - \frac{1}{1+\mu}un_1 - \frac{\mu}{1+\mu}un_4)}) \quad (4.11)$$

The term $\frac{D_2^{'}/d_{21}^{'} - D_2^{''}/d_{21}^{''}}{D_2}$ in Eq. (4.11) is independent of the FOE and can be computed using the point retinal coordinates only. Indeed, it can be shown that

$$\frac{D_2^{'}/d_{21}^{'} - D_2^{''}/d_{21}^{''}}{D_2} = \frac{(n_x^{''}n_y{}' - n_x^{'}n_y^{''})n_y}{(n_x n_y^{'} - n_x^{'}n_y)(n_x n_y^{''} - n_x^{''}n_y)(x_2 - x_1)} \quad (4.12)$$

Thus, Eq. (4.11) is linear in the two unknowns $\frac{1}{D_2}$ and $\frac{\gamma f}{D_2 n_y} + \kappa\alpha - \beta$.

Given a line $\mathcal{L}$ through the image principal point, the proposed method relies on Eq. (4.11) for estimating the term $\frac{1}{D_2}$ corresponding to $\mathcal{L}$. In theory, two quadruples of image points lying on $\mathcal{L}$ suffice to provide estimates of the unknown parameters $\frac{1}{D_2}$ and $\frac{\gamma f}{D_2 n_y} + \kappa\alpha - \beta$. However, to enhance noise immunity, multiple quadruples of points on $\mathcal{L}$ are selected at random and robust estimates of the two unknowns are computed using the Least Median of Squares (LMedS) robust estimator [209]. Knowledge of the term $D_2^{\mathcal{L}}$ for a line $\mathcal{L}$ provides one constraint on the location of the FOE, namely

$$x_0 n_x^{\mathcal{L}} + y_0 n_y^{\mathcal{L}} = x^{\mathcal{L}} n_x^{\mathcal{L}} + y^{\mathcal{L}} n_y^{\mathcal{L}} - D_2^{\mathcal{L}}, \tag{4.13}$$

where (x_0, y_0) is the sought FOE, $(n_x^{\mathcal{L}}, n_y^{\mathcal{L}})$ is the unit normal for line $\mathcal{L}$ and $(x^{\mathcal{L}}, y^{\mathcal{L}})$ is a point on $\mathcal{L}$. Noting that each line $\mathcal{L}$ through the image principal point supplies one constraint of the form of Eq. (4.13) regarding the FOE, the constraints arising from multiple such lines can be combined to yield the FOE. More specifically, using many lines through the image principal point, robust estimates of the corresponding distances $\frac{1}{D_2^{\mathcal{L}}}$ are obtained as previously outlined. For each of the obtained distance estimates, Eq. (4.13) gives rise to a linear constraint regarding the FOE. LMedS is then applied once again on these constraints to give a robust estimate of the FOE. If required, estimates of the rotational velocity can be obtained in a similar manner by employing robust regression on the constraints derived from the terms $\frac{\gamma f}{D_2^{\mathcal{L}} n_y} + \kappa\alpha - \beta$ computed for each line through the image principal point.

4.3 Experimental Results

The proposed method has been extensively tested with the aid of simulated and real flow fields. Representative results from these experiments are given in this section. In all the experiments reported here, at most 180 lines through the image principal point and 200 quadruples of points along each line have been employed.

4.3.1 Synthetic flow fields

The use of simulated data is justified by the fact that knowledge of the ground truth facilitates a quantitative assessment of the accuracy of the results. Besides, simulation enables us to vary in a controlled manner subsets of the parameters involved in the problem of egomotion estimation and then study their effect on the recovered motion.

Therefore, a simulator has been constructed, which given appropriate values for the intrinsic parameters of the simulated camera (focal length and principal point), the translational and rotational motion parameters, the dimensions of the retina and the depth corresponding to each image point, employs Eqs. (2.7) to synthesize an optical flow field. Depths of image points are generated by random variables following various distributions. For the experiments reported here, a uniform distribution in the range $[Z_{min}, Z_{max}]$ and a Gaussian distribution with nonzero mean have been employed. All distances and sizes used by the simulator are specified in units of pixels. To account for the fact that optical flow fields might be sparse, a percentage specifying the fraction of image points having flow vectors can be specified. This percentage is termed the *density* of the optical flow field. To make the simulated optical flow fields more realistic, noise is added to the synthetic optical flows. The noise we employ is generated according to the model suggested in [141]:

$$u_{noisy} = u + sign_1 * N(a, b) * 0.01 * u$$
$$v_{noisy} = v + sign_2 * N(a, b) * 0.01 * v$$

where $sign_1$ and $sign_2$ are binary values that are randomly chosen with equal probability and $N(a, b)$ is a Gaussian random variable with mean a and standard deviation b. This noise model is referred to as ‘‘Gaussian noise with mean $a\%$ and $\sigma = b\%$’’. As noted in [141], 8% and 2% are realistic values for the noise mean and the standard deviation respectively, accounting for most of the errors observed in actual flow fields.

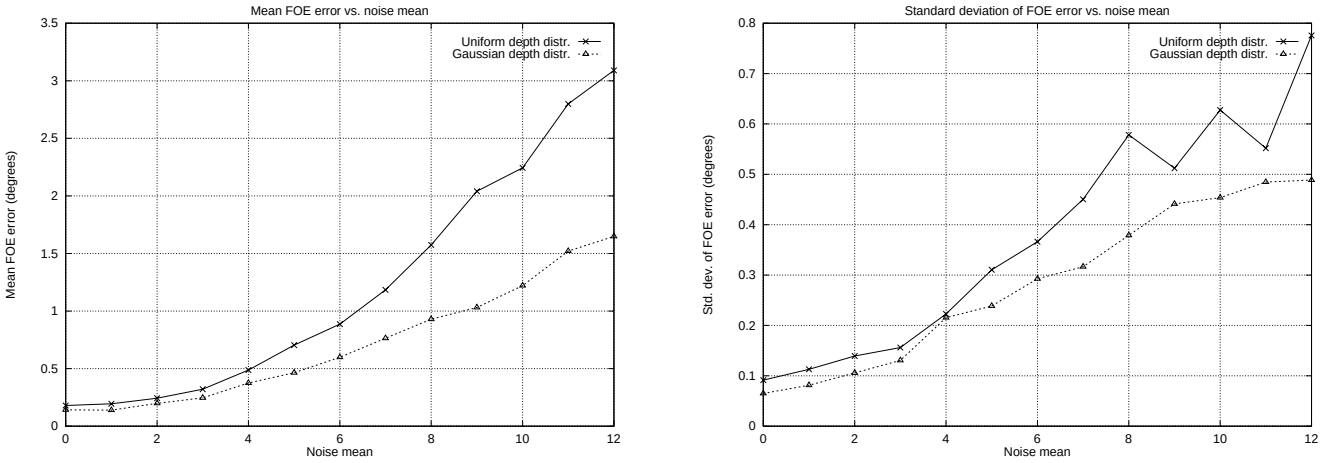
Throughout all experiments, image size was 512×512 pixels, the principal point was assumed to be in the center of the image and the focal length was 256 pixels, amounting

to a field of view of 90 degrees. The density of the optical flow fields was 70%. Two different scenarios for the scene depth were simulated. The first uses a random variable that is uniformly distributed in the range $[10000, 50000]$ pixels to model the depth of a scene with large depth variations. The second scenario employs a Gaussian distribution with mean 15000 pixels and standard deviation 3000, to emulate a scene with less depth variation, in which the majority of the points lie at a dominant depth rather close to the camera. To ensure that the results are independent of the exact depth values used to synthesize the optical flow field, each experiment was run 100 times, each time using a different depth population drawn from the distributions described above.

In the first set of experiments, the effect of noise on the accuracy of the estimated FOE is examined. Employing increasing noise levels, Figures (4.1)(a) and (b) illustrate the mean and the standard deviation respectively of the FOE error for both depth distributions. Each point in the plots summarizes error statistics computed from 100 runs. If f is the focal length and the true FOE is at $(x_0,\ y_0)$ while the estimated is at $(\hat{x_0},\ \hat{y_0})$, the error in the FOE is defined as the angle between the vectors $(x_0,\ y_0, f)$ and $(\hat{x_0},\ \hat{y_0}, f)$, given by

$$cos^{-1}\left(\frac{(x_0,\ y_0,\ f)\cdot(\hat{x_0},\ \hat{y_0},\ f)}{||(x_0,\ y_0,\ f)||\ ||(\hat{x_0},\ \hat{y_0},\ f)||}\right) \tag{4.14}$$

The 3D motion parameters used to synthesize flow were $(U, V, W) = (-120, 100, 150)$ (measured in pixels per frame) and $(\alpha, \beta, \gamma) = (0.005, 0.004, 0.002)$ (measured in radians per frame). The egomotion parameters and the depth values are such that the average translational component of the flow fields is comparable to the average rotational component. The angle between the direction of translation and the optical axis is about 46 degrees. The noise mean was increased to 12% in steps of 1% and the standard deviation was kept equal to 2%. As expected, the error increases with noise but remains acceptable even with very large amounts of noise. The error in the case of Gaussian depths is smaller since in this case the translational component of motion is larger than that in the case of uniformly distributed depths; this is further explained in the discussion of the experiments related to the magnitude of translation below.

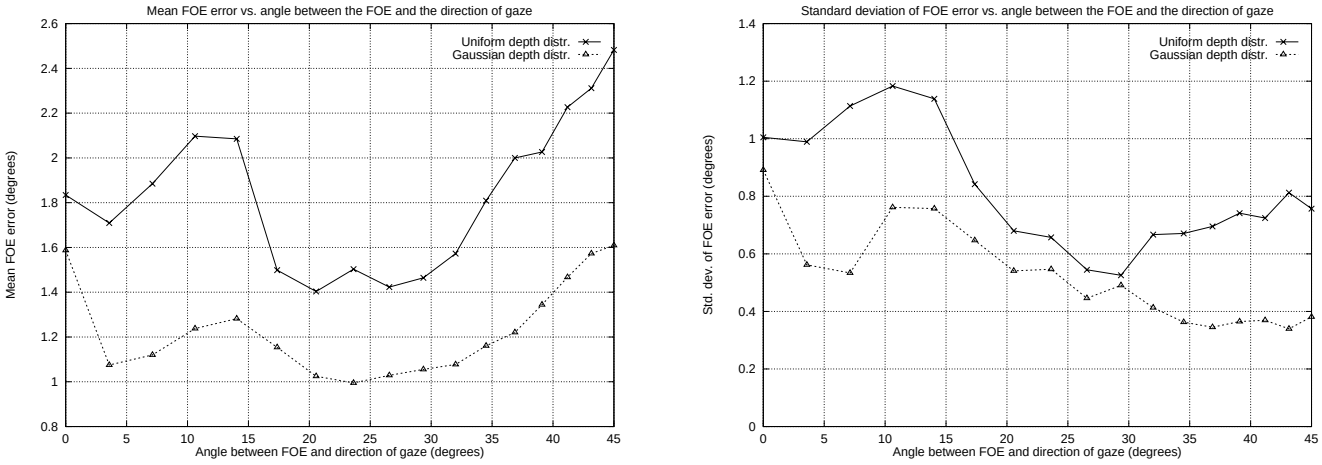


(a) (b)

Figure 4.1: (a) Mean FOE error versus noise and (b) Standard deviation of FOE error versus noise.

It has been observed in previous work on egomotion estimation that the error of the estimated FOE increases with the angle between the direction of translation and the direction of gaze (i.e. the direction defined by the optical axis) [65]. The second set of experiments studies the dependence of the FOE error on this angle for the proposed method. Figures (4.2)(a) and (b) show the mean and standard deviation of the FOE error with respect to the angle between the direction of translation and the direction of gaze. The direction of translation was varied from $(0, 0, f)$ to $(f, 0, f)$, where f is the focal length. In other words, the translations considered range from a straight ahead motion to a sideways motion forming an angle of 45 degrees with the direction of gaze. The rotation parameters were again equal to $(\alpha, \beta, \gamma) = (0.005, 0.004, 0.002)$ and the magnitude of translation has been kept constant, equal to 216.565 pixels per frame, which is the magnitude of translation used in the first set of experiments. Each point in the graphs has been computed from 100 trials, performed with Gaussian noise of mean 8% and standard deviation of 2%. As can be seen from Fig. (4.2)(a), the FOE error does not vary considerably when the angle between the direction of translation and the direction of gaze is increased. This is a desirable characteristic of the proposed method, since it implies that the observer does not need to fixate on the estimated FOE to ensure

small errors in the FOE estimates.

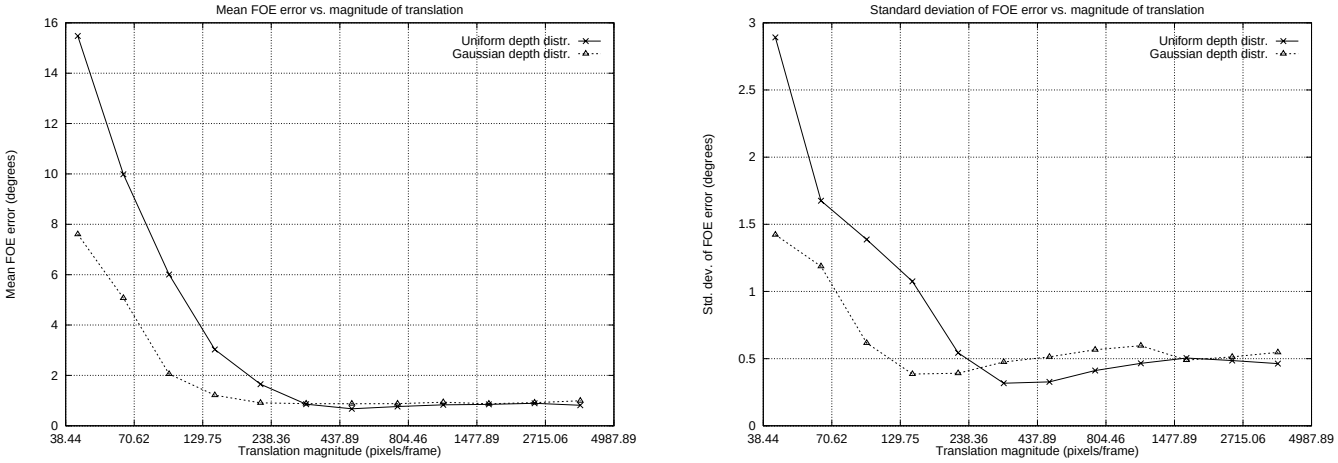


(a) (b)

Figure 4.2: (a) Mean FOE error versus the angle between the direction of translation and the direction of gaze and (b) Standard deviation of FOE error versus the angle between the direction of translation and the direction of gaze.

The last set of experiments evaluates the performance of the method when the ratio between the magnitude of translation and that of rotation is varied. More specifically, assuming that the rotation is constant, Figures (4.3)(a) and (b) depict the effect of varying translation magnitude on the mean and the standard deviation of the FOE error. In this series of experiments, the direction of translation is identical to that defined by $(U, V, W) = (-120, 100, 150)$, but its magnitude is increased by a multiplicative factor of 1.5 between successive experiments. The rotation has been kept constant at $(\alpha, \beta, \gamma) = (0.005, 0.004, 0.002)$ and 100 runs were made for each set of motion parameters. The noise was Gaussian with mean 8% and standard deviation 2%. As can be clearly seen from the plots, the FOE error is significant when the translation magnitude is small (less than 130 pixels per frame in Fig. (4.3)(a)). This is due to the fact that in this case, the translational components of the optical flow vectors are negligible compared to the rotational ones. Therefore, noise has a more pronounced effect on the translational components from which the FOE is recovered. However, as the magnitude of translation increases beyond 130 pixels per frame, the translational parts become comparable or

even larger than the rotational ones. Thus, the translational parts are more immune to noise, giving rise to small FOE errors which are almost constant with respect to the magnitude of translation.

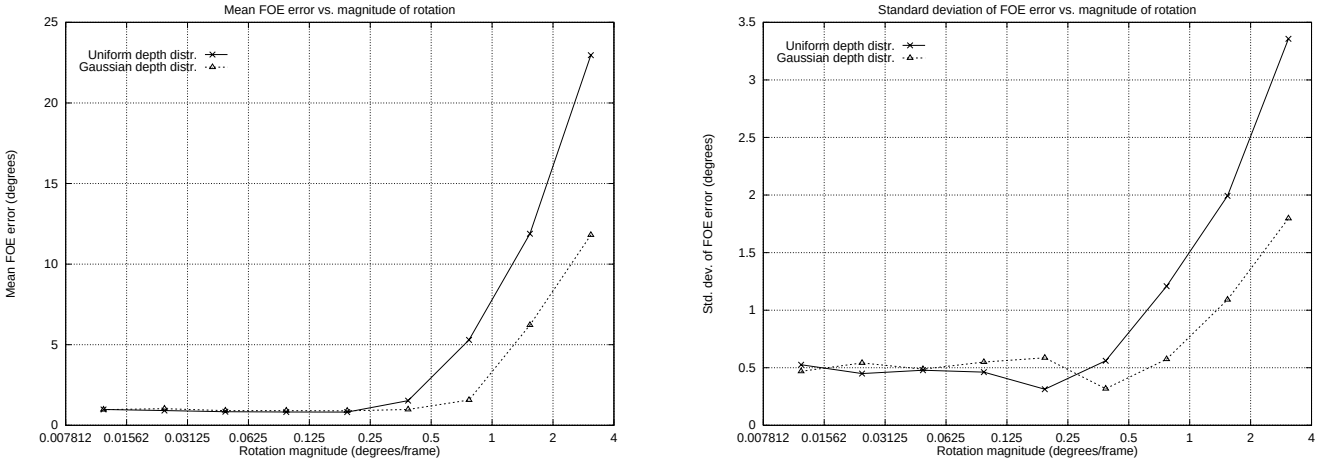


(a) (b)

Figure 4.3: (a) Mean FOE error versus magnitude of translation (b) Standard deviation of FOE error versus magnitude of translation. Note that the scale on the horizontal axes is logarithmic with base 1.5.

Assuming constant translation, Figures (4.4)(a) and (b) show the effect on the mean and the standard deviation of the FOE error caused by altering the rotation magnitude. Here, the behavior of the method is the converse of that observed in the case of constant rotation investigated in the previous paragraph. As can be seen from Fig. (4.4)(a), the error in the FOE estimates is almost constant for realistic amounts of rotation (less than 0.5 degrees per frame). When the rotation increases too much, the flow field becomes mainly rotational, with the rotational components accounting for a large percent of the full flow field. Thus, noise has an increased impact on the translational parts, resulting in large errors for the FOE estimates. During the experiments outlined in Fig. (4.4), translation was kept fixed at $(U, V, W) = (-120, 100, 150)$, the rotation magnitude was increased by a multiplicative factor of 2.0 between successive experiments and 100 runs were made for each experiment. The noise was Gaussian with mean 8% and standard deviation 2%. Note that a rotation of $(\alpha, \beta, \gamma) = (0.005, 0.004, 0.002)$ has a magnitude

of 0.3845 degrees. When assuming continuous image motion (i.e. fine time sampling), rotations having magnitudes larger than one degree per frame are very large and thus unrealistic.



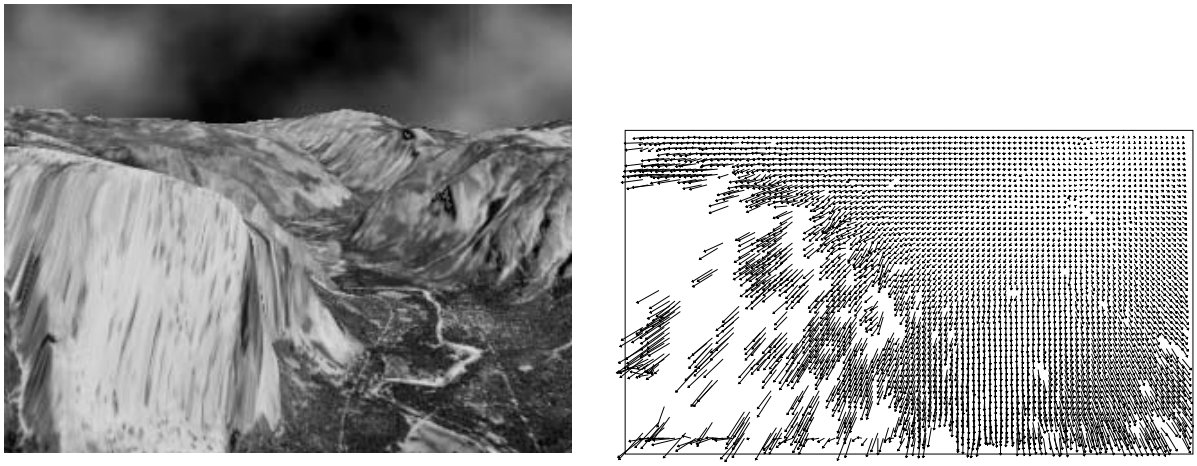
(a) (b)

Figure 4.4: (a) Mean FOE error versus magnitude of rotation and (b) Standard deviation of FOE error versus magnitude of rotation. Note that the scale on the horizontal axes is logarithmic with base 2.0.

4.3.2 Real Image Sequences

The method has also been tested using flow fields computed from real imagery for which the ground truth was known a priori. Throughout all experiments, optical flow was computed using an implementation of the Lucas & Kanade algorithm [151]. The first experiment employed the ‘‘yosemite’’ image sequence, one frame of which is shown in Fig. 4.5(a). This sequence contains both translation and rotation and depicts a flight through Yosemite valley. Since the clouds are moving independently, only the optical flow vectors computed at the lower portion of the images have been employed. This portion of the original images corresponds to a field of view equal to 49.6 degrees horizontally and 29 degrees vertically. The true FOE is rather close to the center of the

field of view, namely at (0, 58)[1] while the estimate computed by the proposed method was (-17.3, 72.3), a value that corresponds to an error of 22.4 pixels or 3.7 degrees. This amount of error compares favorably to errors in the ''yosemite'' FOE estimates appearing in the literature. More specifically, Heeger and Jepson [100] report an error of 3.5 degrees for the ''yosemite'' sequence and Daniilidis [62] reports an error of 4.0 degrees.



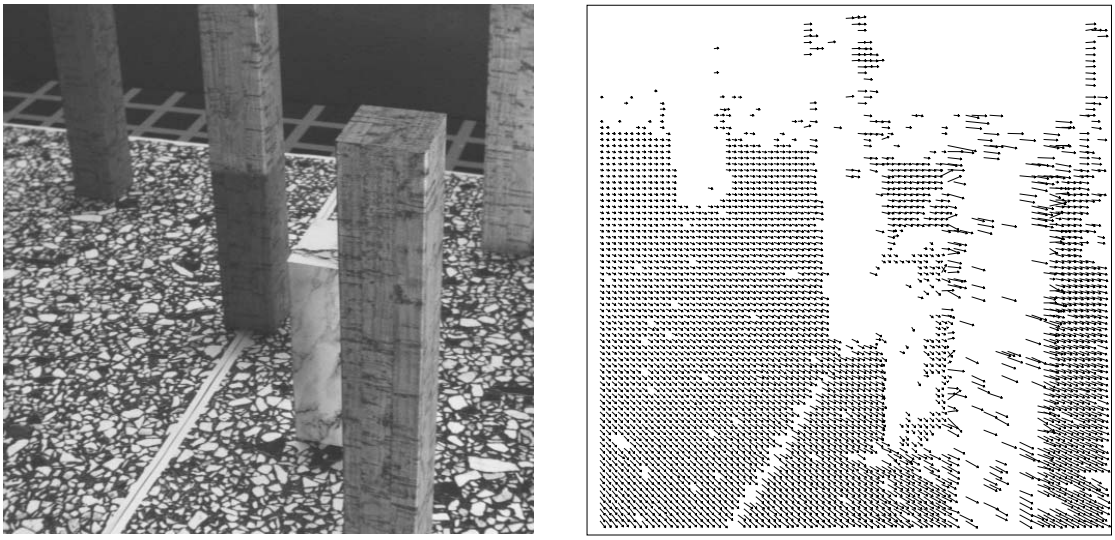
(a) (b)

Figure 4.5: (a) One frame from the ''yosemite'' image sequence (b) The optical flow field used for egomotion estimation.

The second experiment refers to the ''marbled block'' sequence, one frame of which is shown in Fig. 4.6(a). The sequence is described in [187] and contains many sharp discontinuities in depth and motion. The sequence was captured by a translating camera mounted on a robot arm that was moving above a textured floor in a right to left direction. The four dark blocks that lie on the floor are stationary, while the white block in the middle of the scene is moving independently with a right to left direction. The images of the ''marbled block'' sequence subtend 25.6 degrees of visual angle. The true FOE is outside the field of view, specifically at (777, 95.6). Thus, the angle between the direction of translation and the optical axis is about 35 degrees. The proposed method estimated

[1]These are ''calibrated'' image coordinates, defined with respect to the image principal point.

the FOE at (625.0, 111.4), in error by 152.7 pixels or 5.65 degrees. For comparison, the FOE estimate reported by Daniilidis in [62] amounts to an error of 7.17 degrees.

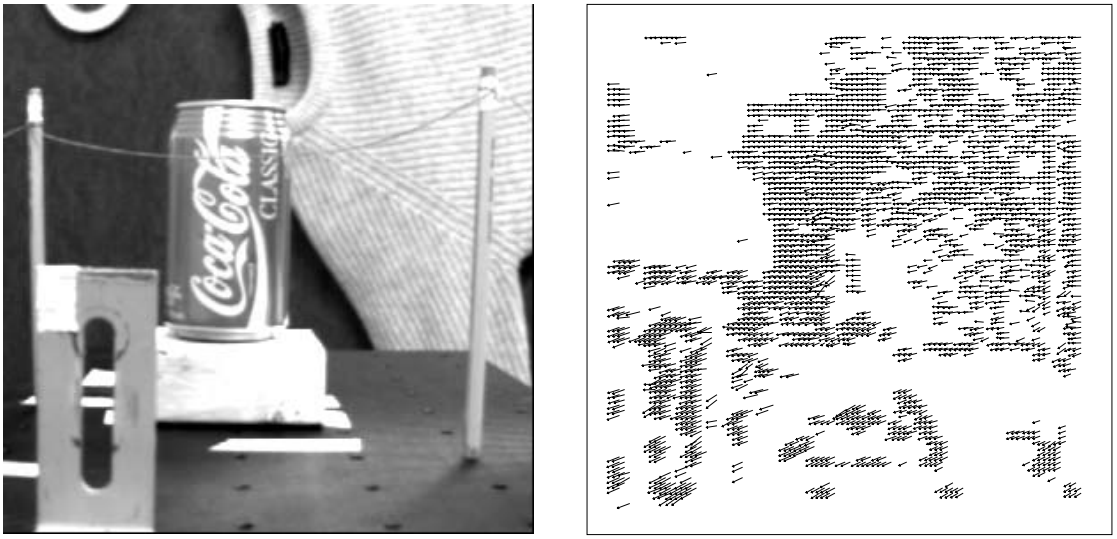


(a) (b)

Figure 4.6: (a) One frame from the ‘‘marbled block’’ image sequence (b) The optical flow field used for egomotion estimation.

The last experiment is based on the ‘‘nasa’’ image sequence, shown in Fig. 4.7(a). Since the camera undergoes a purely translational motion, a rotation of $(\alpha, \beta, \gamma) = (-0.00025, -0.0018, 0.00030)$ was added synthetically in order to make the experiment more challenging. The ground truth for the FOE is (-5, -8) while the recovered FOE was (2.21, 49.29), in error by 57.74 pixels or 5.5 degrees. The images of the ‘‘nasa’’ sequence subtend 24 degrees of visual angle.

At this point, it should be noted that the errors in the FOE estimate are larger for small field of view image sequences. This observation agrees with the findings of [81, 75], which conclude that due to the inhomogeneous flow characteristics of a large field of view, the latter is more helpful for determining the singularities of the flow field (i.e. FOE and axis of rotation) compared to a narrow field of view.



(a) (b)

Figure 4.7: (a) One frame from the ‘‘nasa’’ image sequence (b) The optical flow field used for egomotion estimation.

4.4 Summary

Accurate estimation of camera motion is important for many vision based tasks. In this work, a novel constraint regarding the parameters of 3D motion has been presented. This constraint was used to develop a method for egomotion estimation that has several advantages. First, the method does not impose any constraints on the egomotion that can be recovered or on the structure of the viewed scene. Second, egomotion is computed through closed form solutions of linear equations, avoiding searching the space of possible solutions. Third, instead of employing local information derived from small image regions, redundancy is exploited by combining information across the whole visual field. Fourth, the method does not assume the availability of a dense optical flow field. This is very important for practical applications, since image sequences often have uniform, textureless areas that give rise to sparse optical flow fields. Finally, the use of a robust estimator such as LMedS safeguards against errors in the input, which could otherwise have a significant effect on the accuracy of the computations. Experimental results collected from extensive simulations as well as real image sequences indicate the

effectiveness and robustness of the proposed method. Regarding future enhancements to the proposed method, temporal filtering of the estimated FOE with the aid of a Kalman filter would further increase the accuracy of the recovered egomotion.

Chapter 5

Obstacle Detection

5.1 Introduction

In order to avoid collisions, autonomous vehicles need a means for sensing obstacles obstructing their path. This can be achieved though the combination of two distinct tasks, namely the tasks of obstacle detection and obstacle avoidance. The first task addresses the questions of what is the visual information signaling the presence of obstacles and how this can be extracted from a set of images. The second task, deals with employing sensory input for generating an appropriate sequence of control commands that will drive a mobile vehicle away from the detected obstacles. In this chapter, we assume a vehicle capable of acquiring images of its surroundings and propose a vision-based approach to obstacle detection.

Most approaches to visual obstacle detection exploit motion cues for locating obstacles. Furthermore, an assumption that is often made is that vehicle motion is confined to a surface that is either planar or can be approximated locally by planes [68, 48, 123, 212, 85, 295]. The existence of a planar ground gives rise to a phenomenon termed as *motion parallax* in the psychophysics literature [89]: A moving observer, perceives objects extending vertically from the ground to move differently from their

immediate background (see also appendix A). Various techniques for obstacle detection based on motion parallax have been proposed. Enkelmann [68] for example, uses a calibrated camera to compute a reference flow related to the motion of the ground and then compares it with the flow estimated from images captured by a monocular observer. Inconsistencies between these two flows signal the presence of obstacles. Enkelmann assumes that the camera pursues a purely translational motion. However, such an assumption about egomotion is not always valid and should be avoided when possible. Carlsson and Eklundh [48] assume a camera with unrestricted motion and predict the egomotion and the equation of the ground plane from long image sequences. Obstacles are identified in regions whose motion differs from that predicted. Matthies [158] employs stereo maps and uses the range of points to determine whether they lie close to the ground plane or not. Jenkin and Jepson [123] apply the EM algorithm to obtain maximum likelihood estimates of the parameters of a mixture model describing the disparity field computed with phase-based techniques from a calibrated stereo pair. The probability that a point does not belong to the floor is then computed from the ownership probabilities of the mixture model. Santos-Victor and Sandini [212] employ the normal flow field estimated with an uncalibrated camera and detect obstacles lying on a planar floor by performing an inverse perspective transformation that maps the normal flow onto a horizontal (parallel to the floor) plane. Their method, however, uses an approximate parametric model of the flow generated by the ground plane, deals with outliers in an ad hoc manner and requires the camera to remain in a fixed position relative to the vehicle. Fornland [85] uses the normal flow field measured from a camera moving parallel to the ground plane to derive a linear equation relating motion parameters to the spatiotemporal derivatives of the image intensity function. Obstacles are then detected as the outliers of a robust fit estimated by RANSAC [82] over the image points. Zhang et al [295] present three algorithms for obstacle detection. The first algorithm employs optical flow and a calibrated camera to derive a linear system whose solvability implies the absence of obstacles. The second algorithm does not require camera calibration and exploits the homography of the ground plane to derive a linear system relating corresponding image coordinates in two views. Similarly to the first

algorithm, inconsistency of this linear system signals the presence of obstacles. This algorithm is discussed in more detail in section 5.3. The third algorithm uses sequences of partially calibrated stereo pairs to estimate the equation of the ground plane and the height of obstacles. Although not specifically intended for obstacle detection, the scheme described by Sinclair and Blake in [225] can be used for finding the floor as the dominant plane recovered from a scene. In doing so, they employ pairs of matched points extracted from a stereo pair and assign them to planes according to the conservation of the two five-point projective invariants [168] (see also section 3.2). Fornland and Schnörr [86] solve a similar problem, with their major contribution being that they do not assume that correspondence among points has been established.

Approaches that do not assume a planar ground have also been suggested. Nelson and Aloimonos [180] show that the directional divergence of the 2D motion field can be used as a qualitative cue indicating the presence of obstacles in the field of view of a monocular observer pursuing unrestricted motion. Young et al [286] make no assumptions regarding the structure of the viewed scene and examine geometric properties of the flow field to achieve obstacle detection. Ringach and Baram [206] define an immediacy measure, representing the imminence of collision between an object and a moving observer. A diffusion process, initialized by estimates of the normal flow, is shown to converge asymptotically to the immediacy measure, thus permitting the detection of objects moving towards the camera. Santos-Victor and Sandini [214] present a navigation system driven by a divergent stereo setup. The driving cue they use mimics the behavior of free flying honeybees and is based on qualitative optical flow information, computed on non-overlapping areas of the visual field of the two cameras. Coombs et al [56, 47] also use the divergence of a wide angle optical flow field for collision detection and employ two peripheral flow fields for steering. Kundur et al [137] introduce the a cue that provides a measure for changes in relative range as well as absolute clearances between a 3D surface and a moving observer. The cue is dependent on translation only and can be extracted from a sequence acquired by a fixating camera in motion.

In this work, we propose a method that uses two images to detect obstacles along the path of an autonomous vehicle. There are three main motivations behind our work. First, a basic observation is that the task of establishing dense correspondences between image points, either in the form of optical flow [165] or in the form of stereo disparities [67], is both algorithmically and computationally difficult. We thus avoid one of the major shortcomings of many of the previously published works by using very sparse point correspondences. Second, we refrain from imposing any restrictive assumptions regarding the acquisition of the image pair, since they are hard to guarantee in practice. Finally, we seek to segment the obstacles out of the viewed scene, instead of giving a binary answer concerning their presence or absence. Based on the above, our method assumes that the ground is planar and starts by computing its motion in the two images using the plane homography estimated from a small set of matched points. Subsequently, compensation of the motion of the ground is performed by warping the second image with respect to the first, according to the computed motion. This warping registers the image of the ground in the two views, so that the obstacles exhibit relative motion between the two images. Finally, a change detection operation between the first and the warped second image locates the obstacles present in the scene.

The proposed method does not rely on the reconstruction of 3D structure, does not require a solution to the correspondence problem for every image point, poses no restrictions on egomotion and does not need any calibration information. This last feature is particularly attractive in the context of active vision [6], where the camera position in 3D as well as the zoom and focus are actively controlled, resulting in frequent changes in the extrinsic and intrinsic parameters of the camera.

The rest of this chapter is organized as follows. Sections 5.2 and 5.3 present the proposed obstacle detection method in detail. Experimental results from the application of the method on real images are presented in section 5.4. The chapter is concluded with a brief discussion in section 5.5.

5.2 Estimation of the Ground Homography

The proposed method starts by extracting a set of corners from each of the two images, using the SUSAN corner detector [231]. A local similarity measure based on normalized cross-correlation assigns to each corner in the first image a set of candidate matches in the second image. Corner correspondence is then established by an iterative algorithm that disambiguates multiple candidate matches using a relaxation labeling scheme. Relaxation labeling is based on the assumption that neighboring features have similar disparities. More details regarding the matching algorithm can be found in [290, 294]. Matched corners are the reference points on which the procedure for estimating the fundamental matrix $\mathbf{F}$ is based as follows: Let $\mathbf{f}$ be the 9×1 vector defined by the 9 unknown elements of matrix $\mathbf{F}$, i.e. $\mathbf{f} = (F_{11}, F_{12}, F_{13}, F_{21}, F_{22}, F_{23}, F_{31}, F_{32}, F_{33})^T$. Then, Eq. (2.19) can be written as

$$(m_1 m_1',\ m_2 m_1',\ m_1',\ m_1 m_2',\ m_2 m_2',\ m_2',\ m_1,\ m_2,\ 1)\mathbf{f} = 0 \tag{5.1}$$

Considering N matched pairs, the N constraints given by Eq. (5.1) can be written more compactly as the following linear homogeneous equation:

$$\mathbf{A}\mathbf{f} \ = \ \mathbf{0},$$

where $\mathbf{A}$ is a $N \times 9$ matrix.

The fundamental matrix is then estimated from the solution of the following minimization problem:

$$min_{\mathbf{f}} ||\mathbf{A}\mathbf{f}||^2 \quad subject\ to \quad ||\mathbf{f}||^2 = 1, \tag{5.2}$$

where $||\ ||$ denotes the vector 2-norm [90]. The solution to this constrained minimization problem is known to be the eigenvector of the matrix $\mathbf{A}^T\mathbf{A}$ that corresponds to the smallest eigenvalue (see appendix B for more details).

As noted in [98, 291], $\mathbf{A}^T\mathbf{A}$ is inhomogeneous in image coordinates and, therefore, ill-conditioned. To improve its condition number and to derive a more stable linear

system, the coordinates of the matched corners are normalized by a pair of linear transformations $\mathbf{L}$ and $\mathbf{L}'$ as follows: $\mathbf{L}$ defines a translation of the corners in the first image, such that their centroid is brought to the origin of the coordinate system, followed by an isotropic scaling that maps the average corner coordinates to $(1, 1, 1)$. $\mathbf{L}'$ is defined similarly for corners in the second image. As shown in [98, 291], these transformations result in a more stable system, from which a fundamental matrix $\hat{\mathbf{F}}$ can be estimated. $\mathbf{F}$ is then computed from $\hat{\mathbf{F}}$ as $\mathbf{F} = \mathbf{L}'^T \hat{\mathbf{F}} \mathbf{L}$. At this point, it should be noted that there exist more sophisticated nonlinear methods for estimating $\mathbf{F}$ [153, 291, 253]. However, for the purposes of the present work, the simple linear technique outlined above gives results with satisfactory accuracy.

Since the normalized matching pairs that are given as input to the estimation process will contain errors due to false matches and errors in the localization of corners, care must be taken so that these errors do not corrupt the computed estimate. Thus, instead of using the whole set of matched corners to estimate $\mathbf{F}$, the LMedS estimator is employed to find an estimate that is consistent with the majority of the matched corners. Using a predetermined number of iterations, LMedS picks random samples of matching pairs and computes an estimate of $\mathbf{F}$ from each of them. The estimate that yields the smallest median error is returned as the fundamental matrix which best fits the set of matched corners. The singularity constraint $det\mathbf{F} = 0$ can be enforced a posteriori, by using Singular Value Decomposition (SVD) to compute the singular matrix that is closest to the estimated one in terms of the Frobenius norm [98, 236].

The procedure for estimating the ground homography $\mathbf{H}$ is similar to that for estimating $\mathbf{F}$ above. The 9 unknown elements of matrix $\mathbf{H}$ define a 9×1 vector $\mathbf{h}$ such that $\mathbf{h} = (H_{11}, H_{12}, H_{13}, H_{21}, H_{22}, H_{23}, H_{31}, H_{32}, H_{33})^T$. For each pair of corresponding points m and m', Eq. (2.20) yields the following pair of constraints:

$$\begin{aligned} H_{11}m_1 + H_{12}m_2 + H_{13} &= H_{31}m_1m'_1 + H_{32}m_2m'_1 + H_{33}m'_1 \\ H_{21}m_1 + H_{22}m_2 + H_{23} &= H_{31}m_1m'_2 + H_{32}m_2m'_2 + H_{33}m'_2 \end{aligned} \tag{5.3}$$

Using N matching pairs, the $2N$ constraints given by Eq. (5.3), combined with the 6 constraints arising from the skew-symmetry constraint defined by Eq. (2.21) [1], can be written as

$$\mathbf{Bh} \quad = \quad \mathbf{0},$$

where $\mathbf{B}$ is a $(2N + 6) \times 9$ matrix. $\mathbf{H}$ is then estimated by solving

$$min_{\mathbf{h}} ||\mathbf{Bh}||^2 \quad subject\ to \quad ||\mathbf{h}||^2 = 1 \tag{5.4}$$

The solution to the above problem is the eigenvector of the matrix $\mathbf{B}^T\mathbf{B}$ that corresponds to the smallest eigenvalue. As can be clearly seen from Eq. (2.21) and Eq. (5.3), $\mathbf{B}^T\mathbf{B}$ is inhomogeneous in image coordinates. Thus, the normalization procedure that was previously employed for estimating $\mathbf{F}$ is also used for determining $\mathbf{H}$.

Assuming that at least 50% of the matched corners belong to the ground, LMedS is employed to compute a robust estimate of the ground plane homography $\hat{\mathbf{H}}$ defined by the normalized matching pairs. $\mathbf{H}$ is then computed as $\mathbf{L}'^{-1}\hat{\mathbf{H}}\mathbf{L}$. It should be noted that the use of $\mathbf{F}$ in estimating $\mathbf{H}$ is not necessary. $\mathbf{H}$ has 8 degrees of freedom and, since each pair of corresponding ground corners provides 2 constraints, 4 pairs of corresponding ground corners in general position (no three corners are collinear) give rise to 8 constraints regarding the elements of $\mathbf{H}$ and, therefore, suffice to provide a solution. However, knowledge of $\mathbf{F}$ provides 5 constraints regarding $\mathbf{H}$, enabling us to estimate $\mathbf{H}$ using only 2 pairs of matching corners. Thus, the size of the random samples selected by LMedS during the estimation of $\mathbf{H}$ is equal to 2, implying that fewer iterations are required to guarantee that the correct solution is found with a given probability of error (see Eq. (2.30)). In the case of a robot that is continuously looking for obstacles using a stereo rig where each camera remains in a fixed position with respect to the other, $\mathbf{F}$ does not change in time. Hence, the extra cost of estimating it can be paid only once.

Apart from corners, line segments that have been matched [145] between two images

[1] Actually only 5 of these 6 constraints are linearly independent; see [95].

can be used for providing additional constraints regarding the ground homography. More specifically, an image line defined by $ax+by+c=0$, is represented by the vector (a,b,c) in projective coordinates. Every line in 3D, which belongs to the ground plane, projects to two corresponding line segments in two images. Denoting these line segments with the vectors $\mathbf{l}$ and $\mathbf{l}'$, they are related by

$$\mathbf{l}' \simeq \mathbf{H}^{-T}\mathbf{l}, \tag{5.5}$$

where $\mathbf{H}^{-T}$ denotes the inverse transpose of $\mathbf{H}$. The above equation provides 2 constraints regarding $\mathbf{H}$. However, corresponding line segments were not employed in our implementation for estimating the ground plane homography. At this point, it should be pointed out that the linear technique outlined above for estimating $\mathbf{H}$ minimizes the algebraic distance expressed by Eq. (2.20). Nonlinear methods that estimate the homography by minimizing the euclidian distance between the points in the second view and the corresponding transformed points of the first view, can be found in [152, 256, 298].

5.3 Ground Registration and Obstacle Detection

After the homography of the ground plane has been computed, we can compensate for the motion of the ground by warping the second image with respect to the first using bilinear interpolation and the motion defined at each image point by Eq. (2.20). This transformation results in the image of the ground being registered in the two views, leaving all obstacles extruding from the ground plane unregistered. Subtracting the first image from the warped one, we can declare points where the absolute value of the computed difference is above a threshold as belonging to obstacles. For more accurate results that will not be sensitive to changes in the illumination, the change detection method described in [230] is employed. This method is based on a test regarding the variance of the intensity ratios in small neighborhoods in the two images. In some cases, change detection can produce small noisy areas that do not correspond to obstacles. A

size filtering step can effectively eliminate these areas as follows. Pixels that do not belong to a connected component of a minimum size are assumed to be due to noise and can be masked out. Only regions having area greater than some predefined threshold are retained in the final obstacle map.

In the case that a fast binary decision regarding the presence or absence of obstacles is required, the steps of ground registration and change detection in the process described in the preceding paragraph can be omitted, with the process terminating immediately after the application of LMedS. The presence of obstacles is signaled by the existence of sufficient numbers of outlying corners that are closely located in the image plane. If necessary, a rough estimate of the location of obstacles can be computed by a clustering algorithm that will group nearby outlying corners together. Each cluster picked up by the clustering algorithm is then assumed to correspond to an obstacle.

As mentioned in section 5.1, the second algorithm for obstacle detection proposed by Zhang et al in [295] uses the homography of the ground plane, similarly to the method proposed here. There are, however, important differences between the two methods. Zhang et al use a test based on the ratio of singular values obtained from SVD to determine whether the linear system relating corresponding image points in two views is solvable or not. This test requires the specification of an ad-hoc threshold, and as shown in the synthetic experiments reported in [295], is sensitive to noise. A single erroneous correspondence can provide a constraint that makes the linear system unsolvable. The noise sensitivity measured by Zhang et al is expected to increase when their method has to cope with real noisy data instead of simulated ones. This is due to the fact that the noise model they employ during simulation accounts only for small scale deviations from the ground plane, ignoring many other possible sources of noise. In contrast, the method described in this work uses robust regression techniques to ensure that the existence of corresponding pairs of points that are contaminated by noise do not cause the obstacle detection algorithm to fail. Moreover, the algorithm by Zhang et al provides a simple yes/no answer regarding the presence of obstacles, while our method provides

a map indicating the exact location of obstacles in the field of view of the observer.

5.4 Experimental Results

A set of experiments has been conducted in order to test the performance of the proposed method. Representative results from three of these experiments are given in this section. The first two experiments were performed with the aid of stereo pairs acquired by a binocular head mounted on a mobile robot. The third experiment employs two frames from a publicly available monocular image sequence. In all three experiments, the estimates of the ground homography obtained with and without the use of the fundamental matrix were almost identical. MPEG videos showing the results reported here, are available at http://www.ics.forth.gr/proj/cvrl/demos/lourakis/ .

The first experiment refers to the stereo pair shown in Figure 5.1(a) and 5.1(b). The viewed scene consists of a planar floor on which lies a textured poster. A box in the middle and a flower-pot on the right side of the scene are the obstacles to be detected. White rectangles in Fig. 5.1(c) indicate the corners that do not conform to the estimated plane homography. Corners that agree with the estimated plane homography are marked with gray rectangles. As can be seen in Fig. 5.1(c), some of the corners belonging to the floor are marked as outliers after the estimation of the floor homography. These corners have been erroneously matched between the two views, forming pairs that do not satisfy Eq. (2.20).

Fig. 5.1(d) shows the right image warped according to the estimated homography of the ground plane. It is clear from Fig. 5.1(a) and Fig. 5.1(d) that image warping according to the estimated floor homography registers the image of the floor. The obstacles detected after change detection between Fig. 5.1(a) and Fig. 5.1(d) are shown in black in Fig. 5.1(e). No size filtering was necessary. Note that the detected obstacles correspond to the box and the flower-pot.

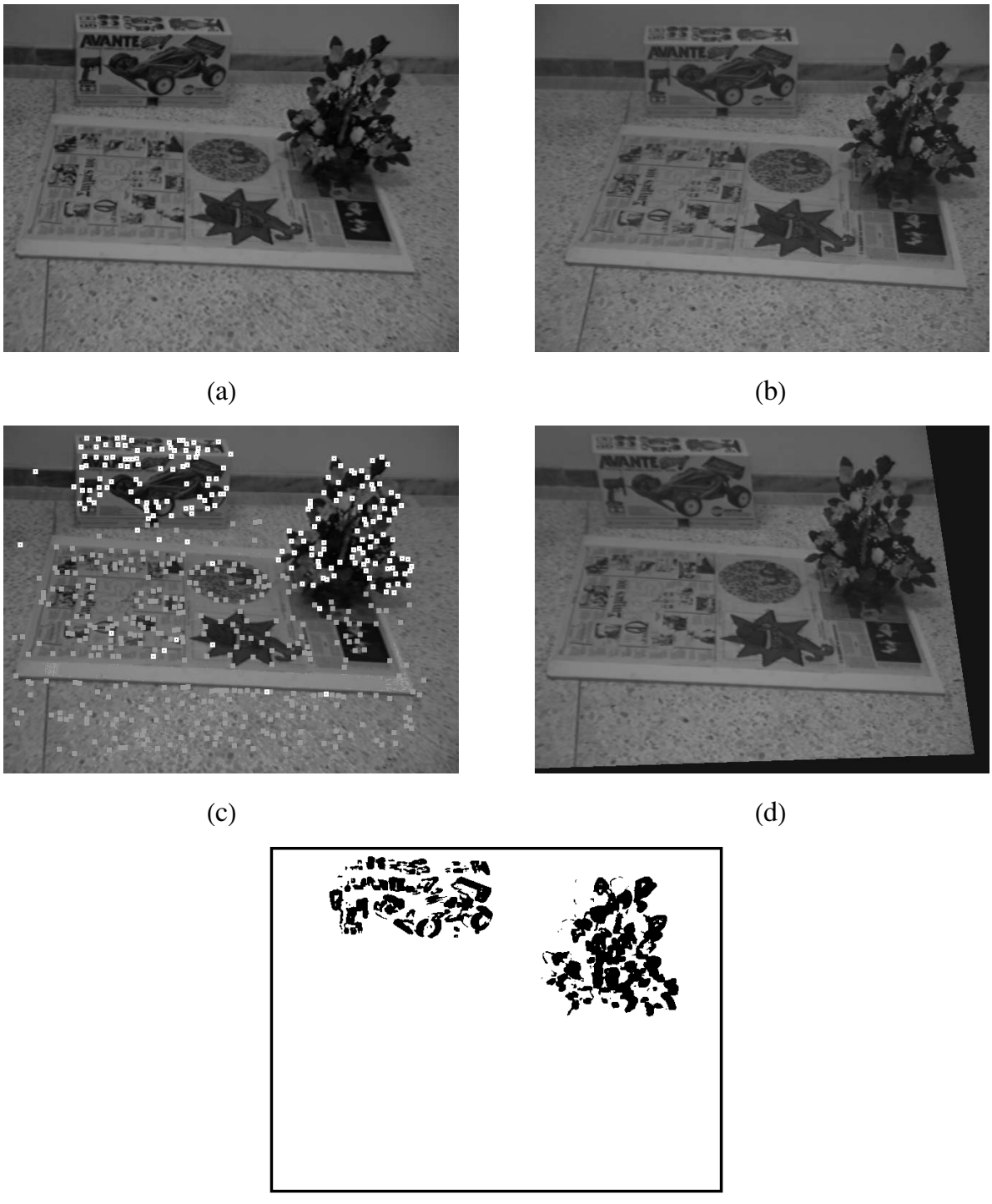


Figure 5.1: First stereo pair: (a), (b) left and right view, (c) outliers detected by LMedS during the estimation of the ground homography, (d) right image warped according to ground homography, (e) detected obstacles (see text for explanation).

The second experiment is based on the stereo pair shown in Figure 5.2(a) and 5.2(b). A textured poster has been placed on a planar floor and a chair on the left side of the scene, along with a box on the right side, are the obstacles to be detected. Corners that do not belong to the floor are characterized as outliers by LMedS during the estimation of the ground homography, and are shown as white rectangles in Fig. 5.2(c). Corners belonging to the floor are marked by gray rectangles. In this particular experiment, a large number of outliers was tolerated. More specifically, LMedS concluded that 91 matched corners from a total of 201 (a percentage of about 45%) are outliers. This clearly demonstrates the robustness of the proposed method.

Fig. 5.2(d) shows the right image warped according to the estimated homography of the ground plane. This warping registers the image of the floor between Fig. 5.2(a) and Fig. 5.2(d). The obstacles detected after change detection between Fig. 5.2(a) and Fig. 5.2(d) are shown in black in Fig. 5.2(e). Again, size filtering on the output of change detection was not required. Note that the chair and the box have been successfully identified as obstacles.

The third experiment tests the obstacle detection method using two frames from the ‘‘marbled block’’ sequence. Frames 20 and 30 of this sequence are shown in Figure 5.3(a) and 5.3(b). The sequence is described in [187] and contains many sharp discontinuities in depth and motion. The sequence was captured by a camera mounted on a robot arm that was moving above a textured floor in a right to left direction. The four dark blocks that are standing on the floor are stationary, while the white block in the middle of the scene is moving independently with a right to left direction. The ‘‘marbled block’’ sequence has a lot of texture, which yields a large number of corners. Corners lying above the floor violate Eq. (2.20) and are marked as outliers by LMedS. Outliers and inliers are shown respectively with white and gray rectangles in Fig. 5.3(c).

Fig. 5.3(d) shows frame 30 warped according to the estimated ground homography. The obstacles detected after change detection between Fig. 5.3(a) and Fig. 5.3(d) are shown in black in Fig. 5.3(e). No size filtering was performed on this result. Clearly, all

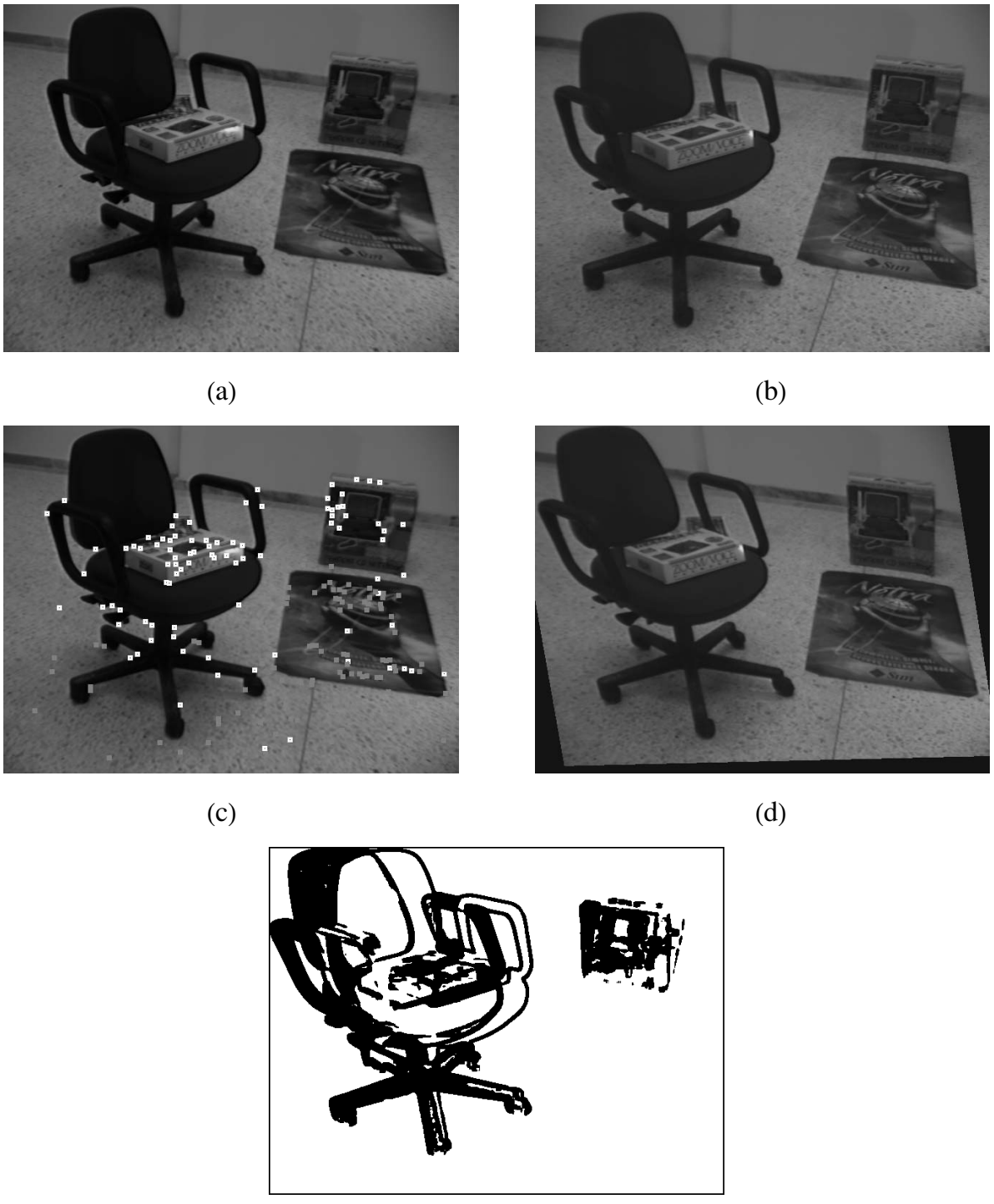


Figure 5.2: Second stereo pair: (a), (b) left and right view, (c) outliers detected by LMedS during the estimation of the ground homography, (d) right image warped according to ground homography, (e) detected obstacles (see text for explanation).

five blocks have been successfully identified as obstacles. This experiment demonstrates that the proposed method can detect obstacles even in the presence of independent motion.

5.5 Summary

The capability of obstacle avoidance is crucial for a robot moving in an unknown environment. In this chapter, a novel method for obstacle detection has been presented. The method is based on the registration of the ground between two views of the environment, which leaves objects not belonging to the ground unregistered. Registration exploits geometrical constraints imposed by the planarity of the ground and is achieved with the aid of a sparse set of corners that have been matched in the two images. The proposed method has several advantages. First, it does not require any calibration information to be known. This feature is particularly important for a mobile vehicle, since in this case the calibration parameters may be continuously changing. Second, the method does not require the computation of a dense set of disparities between the two views and, therefore, solving the correspondence problem for each image point is avoided. Third, there is no need for explicitly recovering the 3D structure of the viewed scene. Fourth, the method is usable either by a monocular vehicle moving in the environment or by a binocular one. Finally, the use of a robust estimator such as LMedS guards against errors in the input, which could otherwise have a significant effect on the accuracy of the computations.

The main disadvantage of the proposed method is that it requires at least 50% of the matched corners to be on the ground, a constraint imposed by the breakdown point of LMedS. One way to overcome this limitation is to use robust estimators having higher breakdown points, such as the one proposed in [235]. A related shortcoming is that the method assumes that the ground is textured, in order to be able to extract corners. It should be noted, however, that most vision algorithms are expected to run

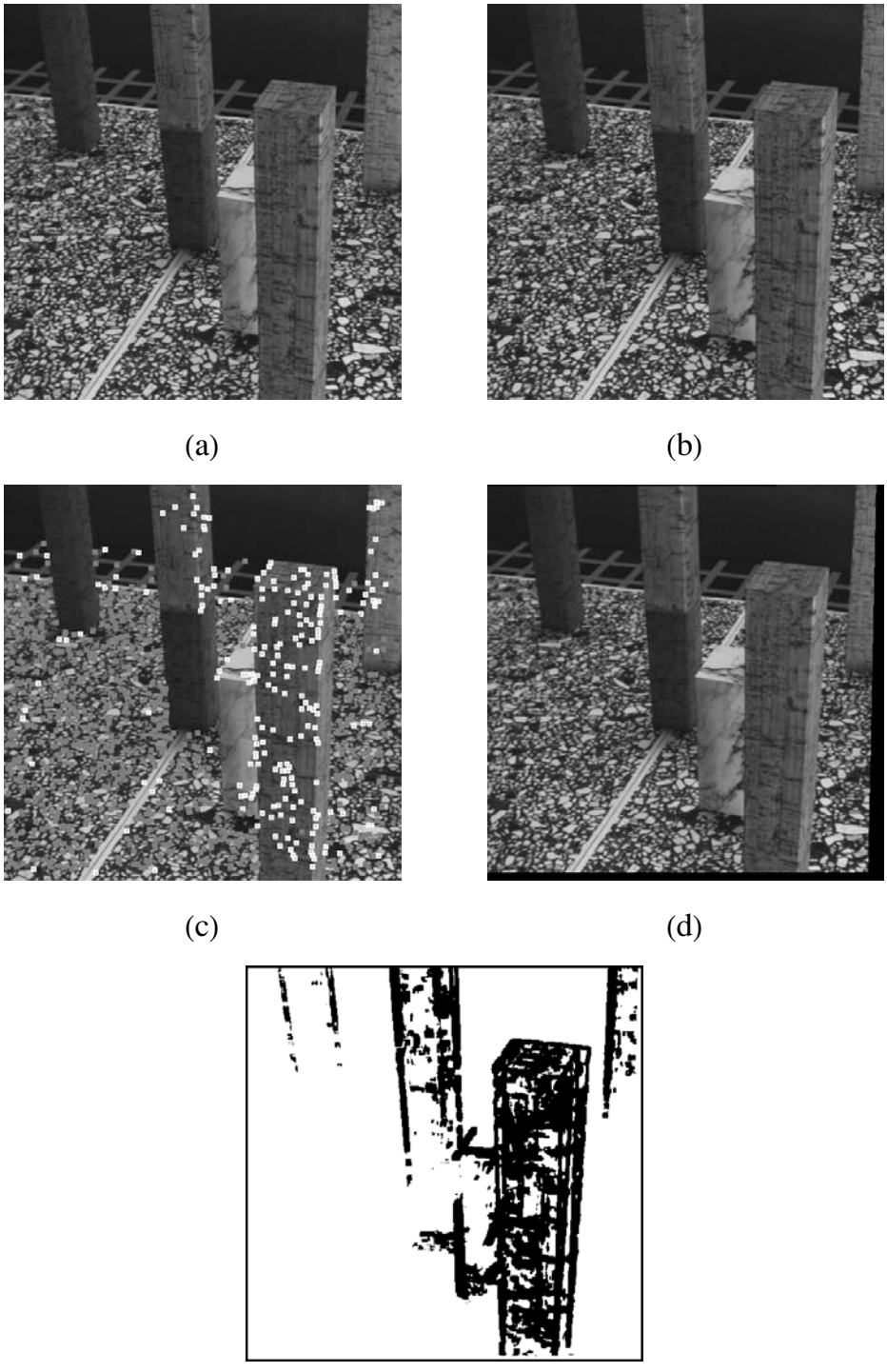


Figure 5.3: The ‘‘marbled block’’ monocular sequence: (a), (b) frames 20 and 30, (c) outliers detected by LMedS during the estimation of the ground homography, (d) frame 30 warped according to ground homography, (e) detected obstacles (see text for explanation).

into difficulties in the absence of texture.

Chapter 6

Time-to-Contact Estimation

6.1 Introduction

A vehicle which is intended to move autonomously in the environment should possess a means for avoiding obstacles obstructing its path. According to the purposive vision paradigm [6], the vision system of such a vehicle should attempt to recover only those aspects of the world that are relevant to the task the vehicle has to perform, instead of reconstructing a 3D representation of the world. Indeed, as it has been previously demonstrated in [180], obstacle avoidance can be achieved using very little information, namely the time-to-contact. The time-to-contact (also known as the time-to-collision or time-to-impact) with a point in the field of view of a moving observer, is defined as the time that remains before the point in question collides with the observer, provided that they continue to maintain the same relative translational velocity. Knowledge of the time-to-contact is more useful for obstacle avoidance compared to estimates of the distance between the obstacles and the observer, since the former is by definition taking into account the dynamics of the observer's motion. There is also strong biological evidence suggesting that insects rely heavily on estimates of the time-to-contact when landing or avoiding obstacles [110]. Owing to the inherent scale ambiguity that characterizes visual motion (see section 2.1.2), the information related to the scene structure that can

be recovered from Eqs. (2.7) is at most the time-to-contact.

The most common approach in the literature for estimating the time-to-contact is to employ the first or even second order derivatives of the optical flow [238, 250, 11, 20]. Such methods share the drawback of being sensitive to errors in the estimates of optical flow, since the latter are always corrupted by noise which is amplified by the process of differentiation. Subbarao [238], for example, presents a theoretical derivation of upper and lower bounds for the time-to-contact using retinal velocities and their first order derivatives. Tistarelli and Sandini [250], exploit the geometric properties of special space variant retinas to estimate the time-to-contact using the derivatives of the image flow. Ancona and Poggio [11], employ a conventional image sensor and compute the time-to-contact from the first order spatial derivatives of the optical flow. Arnspang et al [21] assume an observer with constant egomotion and employ the concept of optic acceleration [20] to develop a method for estimating the time-to-contact from the normal flow field and its first order derivatives, without any knowledge of the camera intrinsic calibration.

In an attempt to overcome the problems associated with the computation of retinal velocity derivatives, Meyer [163] has proposed a technique for estimating the time-to-contact from long monocular sequences. He assumes that the optical flow field can be segmented into regions whose motion can be described by affine models. The coefficients of these models along with their temporal derivatives are estimated through a multiresolution scheme combined with temporal filtering by a Kalman filter. The time-to-contact is then recovered using the estimated coefficients. It should be noted that Meyer's approach assumes that the viewed surfaces are smooth and far from the camera, so that their motion can be approximated by affine models. An additional shortcoming is that it is assumed that the vertical component of the translational velocity of the camera is zero. However, even if the observer is moving on planar ground, this hypothesis is valid only in the case that the camera optical axis is kept parallel to the ground. Cipolla and Blake [51] take a different approach and relate the temporal derivative of the area of

a tracked closed contour and its moments to the time-to-contact. Although they avoid the computation of dense image velocity fields and their derivatives, their method is likely to be sensitive to occlusions of the tracked contour.

At this point it should be pointed out that a few of the methods described above (e.g. [51, 11]) estimate the time-to-contact in special cases only. This is because they estimate the time-to-contact based on the image flow *divergence*. The divergence is one of the differential invariants of the optical flow field [238] and expresses the isotropic expansion or contraction of the flow around an image point. Although there is a connection between the divergence and the time-to-contact which has been exploited in [180] to achieve obstacle avoidance in a qualitative way, the time-to-contact can be recovered from the divergence alone only in the case that the viewed surface is frontoparallel with respect to the camera [163].

It is also worth noting that given the egomotion, i.e. the quantities $\frac{Uf}{W}, \frac{Vf}{W}, \alpha, \beta$ and γ, Eqs. (2.7) can be solved for the time-to-contact. In practice, however, the egomotion is computed from the optical flow and is therefore subject to errors. These errors are thus propagated to the time-to-contact estimates. Besides, the estimation of egomotion might require a considerable amount of time. Hence, our work is motivated by the need to develop a method for time-to-contact estimation that will circumvent the problem of egomotion estimation. Towards this direction, we propose a novel method which assumes that the observer is moving on a planar ground. After estimating the time-to-contact with points on the ground, the concept of planar parallax (see appendix A) is employed to recover the time-to-contact with obstacle points. The method avoids both the numerically unstable computation of high order derivatives of image flow and the estimation of the egomotion.

The rest of this chapter is organized as follows. Section 6.2 develops a technique for determining the time-to-contact with a planar surface and combines it further with the concept of planar parallax for recovering the time-to-contact with points that do not belong to the plane. Experimental results from a prototype implementation are presented

in section 6.3, followed by a concluding discussion in section 6.4.

6.2 The proposed method

We start by presenting an overview of the proposed method for estimating the time-to-contact. Following subsections present issues related to the method in more detail. The basic assumption that we make is that the surface on which the robot is moving can be locally approximated by a plane. Thus, the first step of the proposed method is to detect obstacles in order to identify the image points that belong to the ground. To achieve this, we have employed the technique developed in chapter 5 (see also [150]). Briefly, this technique consists in recovering the ground homography for estimating the motion of the ground between two successive images, then warping one of the two images so as to compensate for the estimated ground motion and finally detecting obstacles as those image regions that appear to be nonstationary after the motion compensation. The second step of the proposed method is to estimate the coefficients of the parametric model defining the motion of the floor. This is done by using Eq. (2.10) to fit an eight parameter linear model directly to the spatiotemporal derivatives of image intensity. To reduce the effects that noise in the estimates of the spatiotemporal derivatives might have on the accuracy of the recovered coefficients, fitting is achieved through the use of the Least Median of Squares robust estimator [209]. Using the estimates of the planar flow coefficients, the time-to-contact with points on the plane is computed next. Finally, based on the time-to-contact with plane points, planar parallax (see appendix A) is employed to compute the time-to-contact with points not on the plane.

6.2.1 Time-to-contact with a planar surface

In this subsection, the time-to-contact with each point of a planar surface viewed by a moving camera will be derived. Let (S_X, S_Y) be the slopes of the plane at $(X, Y) = (0, 0)$

and Z_0 the distance of the surface along the optical axis. The equation of the plane in the 3D camera coordinate frame that was defined in Fig. 2.2 is

$$Z = Z_0 + S_X X + S_Y Y$$

or

$$\frac{1}{Z} = \frac{1}{fZ_0}(f - S_X x - S_Y y) \tag{6.1}$$

in image plane coordinates. Substituting $\frac{1}{Z}$ from Eq.(6.1) into Eqs. (2.7), the well-known eight parameter linear model which describes the optical flow for a planar surface is derived [239]:

$$\begin{aligned} u^\pi &= a_1 x^2 + a_2 xy + a_3 x + a_4 y + a_5 \\ v^\pi &= a_2 y^2 + a_1 xy + a_6 y + a_7 x + a_8 \end{aligned} \tag{6.2}$$

where

$$\begin{aligned} a_1 = -\tfrac{\beta}{f} - \tfrac{W_0}{f} S_X \quad , \quad a_2 = \tfrac{\alpha}{f} - \tfrac{W_0}{f} S_Y \\ a_3 = W_0 + U_0 S_X \quad , \quad a_4 = U_0 S_Y + \gamma \\ a_5 = -U_0 f - \beta f \quad , \quad a_6 = W_0 + V_0 S_Y \\ a_7 = V_0 S_X - \gamma \quad , \quad a_8 = -V_0 f + \alpha f \end{aligned} \tag{6.3}$$

and

$$U_0 = \frac{U}{Z_0} \ , \ V_0 = \frac{V}{Z_0} \ , \ W_0 = \frac{W}{Z_0}$$

In the following, we will assume that the camera rotation around its optical axis, i.e. γ, is zero. This is a reasonable assumption for a camera mounted on a mobile robot, since in this case possible rotations are restricted to a combination of pan and tilt. The problem of using the coefficients $a_1 \cdots a_8$ of the optical flow field to recover the motion and orientation of a moving planar surface has been extensively studied in the past

[263, 143, 239, 194]. It has been proved that there exist two possible motions and plane orientations that give rise to the same optical flow field. In other words, the problem is ambiguous, having two possible dual solutions (see for example [239], p. 212). It is also known that these two solutions share the same translational component W_0 along the optical axis. This component is equal to the middle root of a cubic equation whose coefficients are defined in terms of the eight parameters a_i of the optical flow field in Eqs. (6.2) (see [239], p. 220-221). Using the additional constraint $\gamma = 0$, it will be shown that given W_0, the ambiguity can be raised and a unique solution can be found.

As can easily be seen from Eqs. (6.3),

$$\frac{V_0}{U_0} = \frac{1}{2}\left(\frac{a_7}{a_3 - W_0} + \frac{a_6 - W_0}{a_4}\right) \tag{6.4}$$

and

$$\frac{S_Y}{S_X} = \frac{1}{2}\left(\frac{a_4}{a_3 - W_0} + \frac{a_6 - W_0}{a_7}\right) \tag{6.5}$$

Denoting $\frac{S_Y}{S_X}$ and $\frac{V_0}{U_0}$ by λ and κ respectively, it can also be shown that

$$\begin{aligned} a_1 f^2 - a_5 &= U_0 f - W_0 S_X f \\ a_2 f^2 - a_8 &= \kappa U_0 f - \lambda W_0 S_X f \end{aligned} \tag{6.6}$$

Using the above system of two equations, S_X is found to be

$$S_X = \frac{a_2 f^2 - a_8 - \kappa(a_1 f^2 - a_5)}{W_0 f(\kappa - \lambda)} \tag{6.7}$$

and then S_Y can be computed from the known ratio $\frac{S_Y}{S_X}$. If required, solutions for U_0 and V_0 can be obtained in a similar manner. Knowledge of W_0, S_X and S_Y enables us to compute the inverse of the time-to-contact with the point of the planar surface that is projected on image point (x, y) as

$$\frac{1}{ttc^\pi} = \frac{W}{Z^\pi} = \frac{W_0}{f}(f - S_X x + S_Y y) \tag{6.8}$$

Knowledge of the time-to-contact with plane points is exploited in the next subsection for estimating the time-to-contact with points not on the plane.

6.2.2 Time-to-contact with points not on the plane

First we state and prove a lemma which will be combined later with planar parallax to provide estimates of the time-to-contact with points not lying on a plane.

Lemma 6.1 *Suppose that two image points lie on a line that goes through the origin of the image coordinate system (i.e. the principal point). The difference of the projections of their corresponding optical flow vectors along the direction that is normal to the line does not depend on rotation.*

Proof. Let $\mathbf{p_1} = (x_1, y_1)$ and $\mathbf{p_2} = (x_2, y_2)$ be two points in the image and $\vec{\mathbf{n}} = (n_x, n_y)$ be a unit vector that is normal to the line $\mathcal{L}$ defined by $\mathbf{p_1}$ and $\mathbf{p_2}$. The assumptions that there is no cyclotorsion in the egomotion and that $\mathcal{L}$ goes through the image principal point, result in the terms γ and ν in Theorem C.1 being equal to zero. Therefore, the rotational component vanishes and the difference of the projections depends on translation only:

$$un_1 - un_2 \;\; = \;\; DW(\frac{1}{Z_1} - \frac{1}{Z_2}) = D(\frac{1}{ttc_1} - \frac{1}{ttc_2}) \tag{6.9}$$

As shown in the proof of Theorem C.1, the term D in Eq. (6.9) expresses the distance of the FOE from the line $\mathcal{L}$ and is equal to $(x - x_0)n_x + (y - y_0)n_y \;\; \forall \; (x, y) \; \in \; \mathcal{L}$

□

Suppose now that $\mathbf{q_1}$ is a point not lying on the plane and let $\mathbf{q_2}$ be a point on the line defined by $\mathbf{q_1}$ and the principal point. Let un_1 denote the projection of the optical flow at $\mathbf{q_1}$ along the direction that is normal to the line. Eqs. (6.2) can be used to predict the planar optical flow vectors at points $\mathbf{q_1}$ and $\mathbf{q_2}$. Let un_1^π and un_2^π denote the projections along the normal direction of these optical flow vectors. Using Eq. (A.1), the projection of the residual optical flow field at point $\mathbf{q_1}$ is equal to

$$un_1 - un_1^\pi = D(\frac{1}{ttc_1} - \frac{1}{ttc_1^\pi}) \tag{6.10}$$

Also, according to Eq. (6.9), the difference of the planar flow projections in terms of the corresponding times-to-contact is given by

$$un_1^{\pi} - un_2^{\pi} = D(\frac{1}{ttc_1^{\pi}} - \frac{1}{ttc_2^{\pi}}) \tag{6.11}$$

Dividing the last two equations in terms yields

$$\frac{un_1 - un_1^{\pi}}{un_1^{\pi} - un_2^{\pi}} = \frac{\frac{1}{ttc_1} - \frac{1}{ttc_1^{\pi}}}{\frac{1}{ttc_1^{\pi}} - \frac{1}{ttc_2^{\pi}}} \tag{6.12}$$

Consequently, after computing the quantities $\frac{1}{ttc_1^{\pi}}$ and $\frac{1}{ttc_2^{\pi}}$ with the aid of Eq. (6.8), the time-to-contact for point $\mathbf{q_1}$ can be computed by solving Eq. (6.12) for ttc_1. Note that in the previous derivation it has been assumed that the FOE does not lie on the line defined by $\mathbf{q_1}$ and the principal point, so that D is nonzero. In other words, the time-to-contact with points on the line defined by the FOE and the principal point cannot be estimated using the above method.

6.3 Experimental Results

A set of experiments has been conducted in order to test the performance of the described method using both real and synthetic flow fields. Representative results from two of these experiments are given in this section.

The first experiment aims to quantitatively evaluate the proposed method. In order to achieve this, a simulator has been developed, which given appropriate values for the intrinsic parameters of the simulated camera (focal length and principal point), the translational and rotational motion parameters, the dimensions of the retina and the depth corresponding to each image point, employs Eqs. (2.7) to synthesize an optical flow field. Range images are employed to supply the depths of image points. To make the simulation more realistic, noise is added to the synthetic optical flows. The noise we employ is generated according to the model suggested in [141]:

$$u_{noisy} = u + sign_1 * N(a, b) * 0.01 * u$$

$$v_{noisy} = v + sign_2 * N(a, b) * 0.01 * v$$

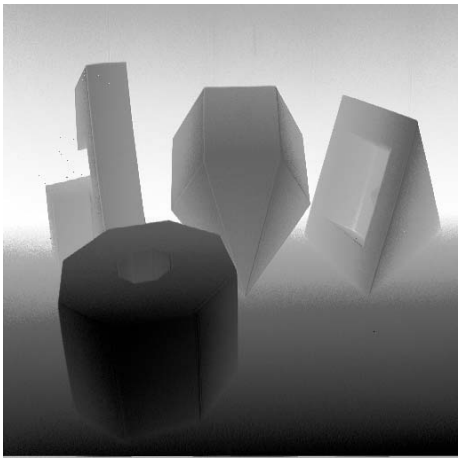
where $sign_1$ and $sign_2$ are binary values that are randomly chosen with equal probability and $N(a, b)$ is a Gaussian random variable with mean a and standard deviation b. This noise model is referred to as ''Gaussian noise with mean $a\%$ and $\sigma = b\%$''. As noted in [141], 8% and 2% are realistic values for the noise mean and the standard deviation respectively, accounting for most of the errors observed in actual flow fields.

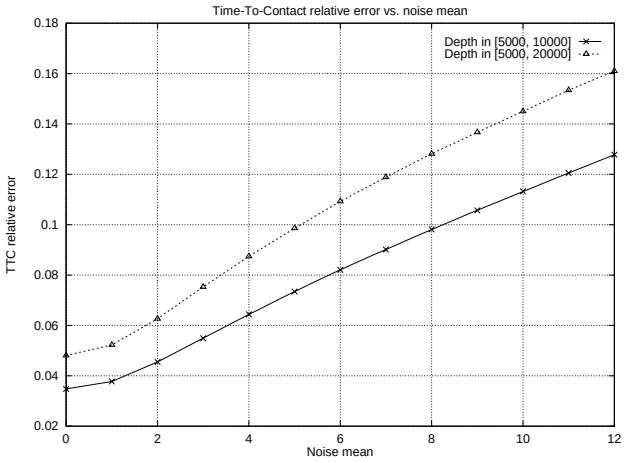
For the first experiment reported here, a range image from the collection [84] has been utilized and is shown in Fig. 6.1(a). To examine the effects of depth variations, two different scenarios for the scene depth were simulated. In the first, all depths have uniformly been scaled to the range $[5000,\ 10000]$ pixels while in the second they range within $[5000,\ 20000]$ pixels. Throughout all trials, the simulated image size was 512×512 pixels, the principal point was assumed to be in the center of the image and the focal length was 256 pixels, amounting to a field of view of 90 degrees. The 3D motion parameters used to synthesize flow were $(U, V, W) = (-120, 100, 150)$ (measured in pixels per frame) and $(\alpha, \beta, \gamma) = (0.005, 0.004, 0.0)$ (measured in radians per frame). The ''ground'' for this scene has been extracted manually. To assess the accuracy of the recovered time-to-contact, the computed estimates have been compared with the known ground truth values and the *relative error* has been computed as

$$\frac{|ttc - \hat{ttc}|}{ttc} \tag{6.13}$$

where $\hat{ttc}$ is the estimated time-to-contact and ttc the correct value. The noise mean was increased from 0% to 12% in steps of 1% and the standard deviation was kept equal to 2%. The average relative error computed from 10 runs versus the noise mean is shown in Fig. 6.1(b) for both scenarios. As expected, the relative error increases with the noise and assumes larger values for the scene having the largest depth variation. The latter is due to the fact that points having large depths give rise to optical flow vectors having small magnitudes, which are thus sensitive even to small amounts of noise.

The second experiment tests the method using two frames from the ''marbled block'' sequence. Frames 20 and 30 of this sequence are shown in Figures 6.2(a) and





(a) (b)

Figure 6.1: (a) The range image for the synthetic experiments; intensity is proportional to depth, with distant objects being brighter. (b) The average relative error in the estimated time-to-contact versus the noise mean.

(b). The sequence is described in [187] and contains many sharp discontinuities in depth and motion. The sequence was captured by a camera mounted on a robot arm that was moving above a textured floor in a right to left direction. The four dark blocks that lie on the floor are stationary, while the white block in the middle of the scene is moving independently with a right to left direction.

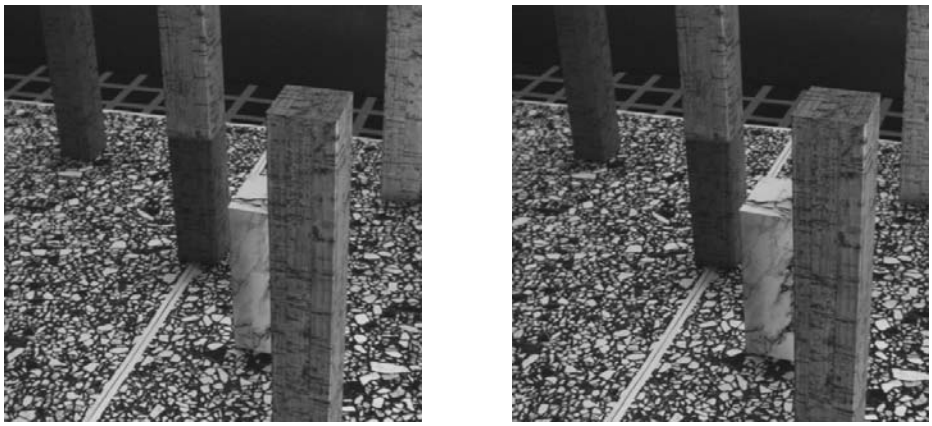
Optical flow was computed using an implementation of the Lucas & Kanade algorithm [151] and is illustrated in Fig. 6.2(c). The obstacles detected by applying the technique described in [150] are shown in Fig. 6.2(d), where white represents the union of ground and textureless points and black corresponds to the detected obstacles. The estimated time-to-contact for each point in the field view is shown in the form of an image in Fig. 6.2(e). In this image, the intensity is proportional to the time-to-contact, with dark points being closer compared to bright ones. Red pixels correspond to points where the time-to-contact could not be computed either due to the lack of reliable optical flow estimates or due to the denominator of the left hand side in Eq. (6.12) being very close to zero. As can be seen from Fig. 6.2(e), the latter case is more frequent in a locus of points along a line through the image center. This is due to the fact that this line goes

through the FOE and as explained in section 6.2, the time-to-contact cannot be estimated for points close to it.

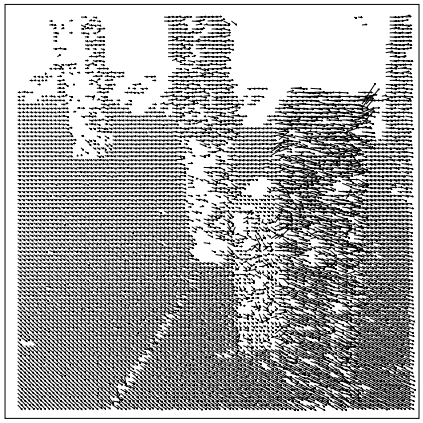
Since the ground truth for the time-to-contact is unavailable, the results can only be qualitatively evaluated. Clearly, the method has successfully recovered the spatial ordering of the four stationary blocks, since closer blocks in Fig. 6.2(e) appear to be darker than further ones. However, as indicated by the white color in Fig. 6.2(e), the method is confused by the independent motion of the white block in Fig. 6.2(a), deducing that is far away from the camera. This is because the block in question is moving in the same direction with the camera (i.e. from right to left), which results in its combined apparent motion being small. Therefore, the method mistakenly concludes that the time-to-contact with the block is large, as if it were far from the camera. By fusing the maps in Figs. 6.2(d) and (e), an autonomous system should be able to identify the free space and also rank obstacles according to their potential for causing a collision.

6.4 Summary

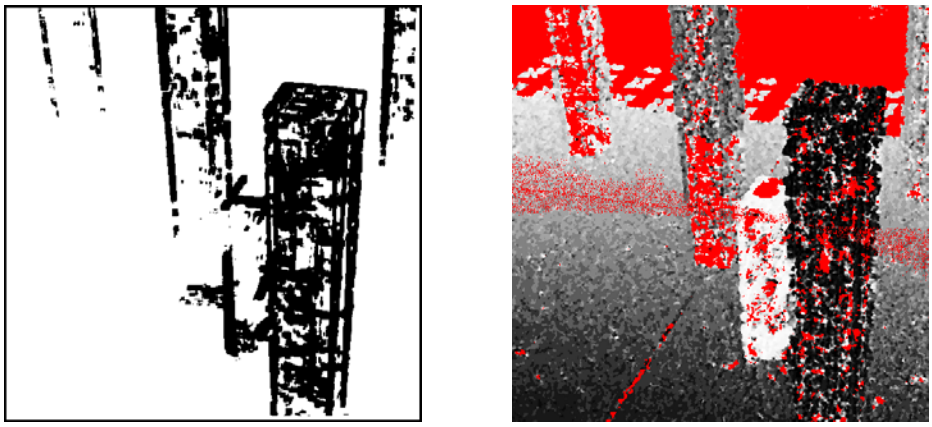
The capability of obstacle avoidance is crucial for a robot moving in an unknown environment. Knowledge of the time-to-contact with points in the robot's field of view is adequate for avoiding collisions with obstacles. In this chapter, a novel method for estimating the time-to-contact has been presented. Assuming that the robot is moving on a planar ground, the time-to-contact with ground points is first estimated using the optical flow field and then planar parallax is exploited to recover the time-to-contact with points not on the ground. The main advantages of the proposed method is that it avoids the computation of high order derivatives of image flow and also does not need to recover the 3D velocity of the camera. In addition, no strict restrictions on the egomotion are posed. The method has been experimentally validated using both real and simulated optical flow fields.



(a) (b)



(c)



(d) (e)

Figure 6.2: The ‘‘marbled block’’ monocular sequence: (a), (b) frames 20 and 30, (c) optical flow shown scaled and subsampled, (d) detected obstacles, (e) estimated time-to-contact (see text for explanation)

Chapter 7

Conclusions

This thesis was concerned with visual navigation, that is the employment of vision to achieve autonomous movement in the environment. The study of navigation is interesting both from the practical standpoint of producing useful mechanisms, and from the abstract standpoint of providing working systems that can be adequately understood and described theoretically. The choise of employing visual motion as the primary sensory input has been motivated by the fundamental role played by motion processing in biological organisms.

It has been argued that the reconstructionist vision paradigm, according to which the goal of vision is the recovery of an accurate representation which captures every detail of the environment, is insufficient to study vision and visual navigation in particular. Instead, the behavioral vision paradigm is better suited to study visual navigation. This paradigm calls for the development of multiple, simple visual processes each of which implements one of the behaviors the system is supposed to exhibit. Each behavior has a distinct, well-defined goal and is tailored to the environment the vision system is expected to operate in. In other words, each behavioral process is responsible for one of the capabilities possessed by the system. Vision is thus realized by a set of cooperating processes, which pursue the system's goals in a synergistic manner.

Based on the premise that the behavioral approach to vision has the potential to lead to successful vision systems, this thesis is concentrated on the study of four visual capabilities, namely independent motion detection, egomotion estimation, obstacle detection and time-to-contact estimation. Independent motion detection refers to the ability of a mobile system to detect objects that are moving independently of it in the environment. Egomotion estimation deals with the problem of deducing the velocity of a mobile system using the images of the surroundings that it captures as it is moving. Obstacle detection addresses the problem of detecting obstacles obstructing the system's path. Time-to-contact estimation describes the ability of estimating the time that remains before a mobile observer collides with objects in his field of view. These capabilities can function as generic navigational tools for building various practical applications. Collectively, they constitute a solid arsenal of primitive algorithms that is able to support complex behavioral repertoires.

This thesis has also demonstrated the viability of a bottom up methodology for studying vision. More specifically, it has made clear that the specification of a set of visual processes that permit the incremental development of a navigation system, in which there is no need for the whole system to be constructed before experiments can be conducted. The research framework that was adopted is discussed in section 1.4.1 and therefore it will not be repeated here. In the remainder of this section, a brief overview of the contributions of the thesis is provided.

The problem of independent motion detection has been formulated as a problem of robust parameter estimation applied to the visual input acquired by a rigidly moving observer. The proposed method automatically selects a planar surface in the scene and the residual planar parallax normal flow field with respect to the motion of this surface is computed at two successive time instants. The two resulting normal flow fields are then combined in a linear model. The parameters of this model are related to the parameters of self-motion (egomotion) and their robust estimation leads to a segmentation of the scene based on 3D motion. The method avoids a complete solution

to the correspondence problem by selectively matching subsets of image points and by employing normal flow fields. In addition, no constraints regarding the scene structure or the observer's motion are imposed.

The method proposed for estimating egomotion relies on a novel linear constraint that involves quantities that depend on the sought egomotion. The constraint is defined in terms of the optical flow vectors corresponding to quadruples of image points which lie on lines going through the image principal point. Combined with robust linear regression techniques, the constraint enables the recovery of the FOE, thereby decoupling the 3D motion parameters. There is no need either for searching the high dimensional space of possible egomotions or for imposing restrictions on the egomotion and/or the scene structure.

The method for detecting obstacles uses two images of the environment and provides a binary labeling of image points, classifying them either as obstacles or as free space. Based on the assumption that the observer is moving on a locally planar ground, the method is able to compute an estimate of the motion of the ground. Subtracting this motion from the two images permits the compensation of the ground motion. Following this, obstacles are detected as areas in the image that appear nonstationary after the motion compensation. The method is particularly attractive, since it does not require camera calibration, it is applicable either to stereo pairs or to motion sequence images and it does not rely on a dense disparity/flow field.

For the case of time-to-contact estimation, a method has been developed that is complementary to the obstacle detection capability and is capable of estimating the time-to-contact without any knowledge of the egomotion. This method is again based on the assumption that the observer is moving on a locally planar ground. First, the time-to-contact with points on the ground is estimated. Then, the phenomenon of planar parallax is exploited to yield the time-to-contact with obstacles. The method has the desirable characteristic of avoiding the estimation of high order derivatives of image flow, which are known to be very difficult to compute accurately.

7.1 Further Research

The present section identifies and discusses possible extensions of the work described in this thesis. To begin with, one direction of further research could be that of improving the proposed navigational capabilities by eliminating any shortcomings that they might have or by loosening the assumptions under which they operate. More details on such improvements can be deduced from the descriptions of these methods in chapters 3, 4, 5 and 6. An other axis of future work would be to exploit deliberate motions of the observer, according to the active vision paradigm [6]. It is probable that some of the problems to be solved can be rendered simpler when the observer is actively involved in the image acquisition process, so as to simplify related calculations.

The visual capabilities presented in chapters 3, 4 and 6 rely upon knowledge of the intrinsic calibration parameters of the employed camera. However, this is an undesirable assumption in the case of a mobile robot. This is because the intrinsic parameters might change frequently due to changes in the camera zoom and focus, therefore any initial calibration information is soon outdated. Obviously, requiring a calibration object for periodic re-calibration is not practical. To remedy this, the continuous motion equations in the uncalibrated case should be studied. Although calibrated cameras can deliver more informative descriptions of the environment compared to uncalibrated ones[1], the latter have proven to be sufficient for dealing with non-trivial vision tasks [289]. Hence, it is a remarkable fact that despite the significant amount of research devoted to the study of uncalibrated discrete motion, continuous uncalibrated motion has received little attention [269].

Clearly, the behavioral repertoire of a system can be enriched by developing more of the visual capabilities outlined in section 1.3.2. The simpler of these capabilities employ precategorical visual processing, i.e. visual information is not linked to object

[1]For example, it is known that using uncalibrated cameras, one can at most recover the scene structure up to an unknown projective transformation [73]. In the case of calibrated cameras, scene structure can be recovered up to an unknown scale factor [258].

descriptions. However, in order to implement more complex capabilities such as homing, visual recognition issues will have to be addressed. In addition, capabilities that rely on other cues besides motion can also be investigated. Color, for instance, is a rich source of information that has been successfully employed for object recognition [242]. Furthermore, non-visual sensing modalities can be employed in order to simplify and accelerate vision tasks. Obstacle avoidance, for example, can be achieved with minimal processing using sonar or laser range sensors, while egomotion estimation is easier and perhaps more accurate when the rotational velocity is supplied by inertial sensors, i.e. gyros.

Due to time and resource limitations, the integration of all the proposed navigational capabilities on a mobile robot has not been attempted. Therefore, in spite of the fact that the ties between the task and the information that must be extracted from visual data have been established, visuomotor control issues have not been addressed. This is also common in the literature, where the coupling between perception and action is rarely studied in detail. Such studies, however, are essential for resolving many practical issues concerning the coordination of behavioral processes. These issues include, but are not limited to, action selection, resource allocation, real-time constraints, interprocess communication, error recovery, motor control, etc. The related research effort should aim at the specification of a framework that would allow the painless incorporation of new visual capabilities in an already working system. Košecka et al [135], for example, propose a formalism based on Discrete-Event Systems for modeling the coupling between different sensory and motor subsystems. This formalism provides systems with composite visual behaviors and well-defined properties. In this respect, machine learning techniques assume the important role of inferring the association between visual patterns and patterns of motor control. Thus, a system can generalize and employ its experience to react to unforeseen situations that are similar to situations that have successfully been dealt with in the past.

All four visual capabilities developed in this thesis rely on a behavior-oriented

notion of visual perception. This notion, however, has been applied only at the level of algorithms and representations. The behavioral paradigm can be extended to cover the visual sensors employed by an autonomous system as well. In other words, an autonomous system could be equipped with visual sensors tailored to accomplish different tasks. Factors determining the properties of each sensor, such as optics, the shape of the retina, the pattern of light-sensitive retinal elements, etc, will be properly selected to match the requirements of specific visual tasks. Early research in this direction has produced fruitful results. For example, the space-variant sensor described in [66], is characterized by high pixel density in the center of the CCD array and coarse resolution towards the periphery. Thus, it achieves both a wide field of view and high resolution foveal vision, while employing a limited number of pixels. As mentioned in chapter 4, Nelson [179] has demonstrated that egomotion estimation is particularly simple when a spherical eye is employed. Yet another example of an alternative camera is supplied by the OmniCam [186], developed at Columbia University. This design makes use of special optics to yield panoramic, i.e. 360^o, views of the environment from conventional CCD arrays. Apart from its obvious advantages in tasks such as surveillance and monitoring, this camera can also be helpful for visual homing by capturing in a single image all the visual information available from a specific location. More details regarding existing designs of alternative visual sensors can be found in [271].

Part III

Appendices

Appendix A

Planar Parallax

Most motion analysis methods express rigid image motion as the sum of two displacement fields, namely a translational and a rotational one. Recently, however, it has been shown that if image motion is expressed in terms of the motion of a parametric surface and a *residual parallax field*, important problems in motion analysis become considerably simpler [136, 215, 120, 146, 60]. In the following, the equations describing the residual field are derived, assuming that the employed parametric surface is a plane.

Let Π be a 3D plane in front of a pinhole camera and let p be a point on the image plane, as shown in Figure A.1. Assume that P and $P^\pi \in \Pi$ are two 3D points located on the optical ray defined by p, i.e. both P and P^π project on the same retinal point p. This can happen, for example, when viewing a scene through a transparent surface. Denoting the optical flow induced by the motion of points P and P^π by (u, v) and (u^π, v^π) respectively, Eqs. (2.7) can be employed to yield the residual flow (u^r, v^r) for points P and P^π [120, 50]:

$$\begin{aligned} u^r &= u - u^\pi = W(x - x_0)(\frac{1}{Z} - \frac{1}{Z^\pi}) \\ v^r &= v - v^\pi = W(y - y_0)(\frac{1}{Z} - \frac{1}{Z^\pi}), \end{aligned} \tag{A.1}$$

where $\frac{1}{Z}$ and $\frac{1}{Z^\pi}$ are the depths of points P and P^π respectively. As can been seen from

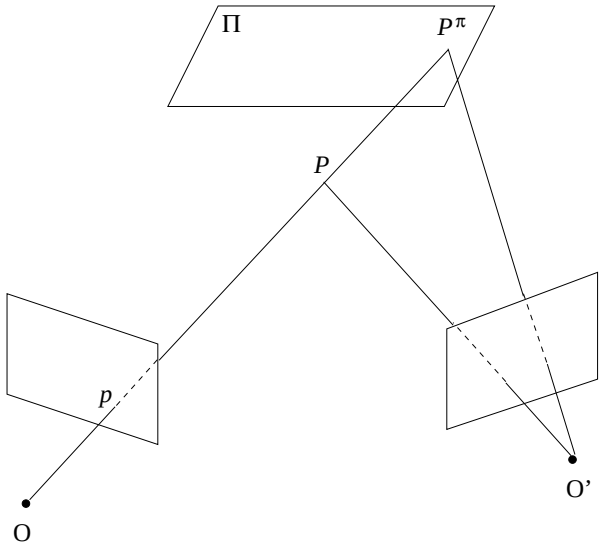


Figure A.1: Planar parallax. O and O' are the camera focal points, Π a 3D plane and P^π, P are two 3D points with $P^\pi \in \Pi$

Eq. (A.1), the residual flow field is purely translational. This is because the rotational components are identical at both points, since they do not depend on depth and are canceled out by subtracting the flow vectors. Consequently, all optical flow vectors of the residual flow point towards the FOE.

In practice, it is difficult to identify image points where two distinct 3D points project simultaneously. However, this difficulty can be alleviated by recalling that the motion of a planar surface can be expressed parametrically by a linear model with eight parameters [239]. Thus, assuming that the parameters defining the motion of the plane in an image have been estimated, the motion that would be induced if the moving plane covered the whole visual field can be predicted using the linear model. Then, subtracting this predicted flow field from the actual optical flow field estimated from the image sequence, yields a residual flow field which is zero at points belonging to the plane and nonzero elsewhere (see also [120]).

Appendix B

The Solution of the Vector Equation $\mathbf{Ax} = \mathbf{0}$ with $||\mathbf{x}|| = 1$

Let $\mathbf{A}$ be a $N \times M$ matrix with $N \geq M$ and $\mathbf{x}$ a $M \times 1$ column vector. The solution for $\mathbf{x}$ of the (possibly overdetermined) vector equation $\mathbf{Ax} = \mathbf{0}$ with $||\mathbf{x}|| = 1$, is equal to the solution of the following optimization problem:

$$Minimize \quad (\mathbf{Ax})^T(\mathbf{Ax}) \equiv \mathbf{x}^T\mathbf{Bx} \quad subject\ to \quad ||\mathbf{x}|| = 1, \tag{B.1}$$

where $\mathbf{B} = \mathbf{A}^T\mathbf{A}$. In the following, it will be shown that the solution to this problem is the eigenvector of the matrix $\mathbf{B}$ which corresponds to the smallest eigenvalue [292].

Recall that $\mathbf{B}$ is a $M \times M$ symmetric matrix, thus it can be decomposed as [90]

$$\mathbf{B} = \mathbf{U}\ \mathbf{E}\ \mathbf{U}^T,$$

with $\mathbf{E}$ a diagonal array formed by the the M eigenvalues v_i of $\mathbf{B}$, i.e. $\mathbf{E} = diag(v_1, v_1, \ldots, v_M)$ and $\mathbf{U}$ comprised by the eigenvectors $\mathbf{e}_i$ of $\mathbf{B}$, i.e. $\mathbf{U} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_M]$. Without loss of generality, let $v_1 \leq v_2 \leq \ldots \leq v_M$. The original problem described by (B.1) is now reduced to the following:

$$Find\ \ x_1, x_2, \ldots, x_M\ \ with\ \ \mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \ldots + x_M\mathbf{e}_M,$$
$$such\ that\ \mathbf{x}^T\mathbf{Bx}\ is\ minimized\ subject\ to\ {x_1}^2 + {x_2}^2 + \ldots + {x_M}^2 = 1$$

Some algebraic manipulation reveals that

$$\mathbf{x}^T\mathbf{B}\mathbf{x} = {x_1}^2 v_1 + {x_2}^2 v_2 + \ldots + {x_M}^2 v_M$$

Therefore, the problem becomes to minimize the following unconstrained function:

$${x_1}^2 v_1 + {x_2}^2 v_2 + \ldots + {x_M}^2 v_M + \lambda({x_1}^2 + {x_2}^2 + \ldots + {x_M}^2 - 1), \tag{B.2}$$

where λ is the Lagrange multiplier. Equating the partial derivatives of the above function with respect to $x_1, x_2 \ldots, x_M$ and λ to zero, yields

$$\begin{aligned} 2x_1 v_1 + 2x_1\lambda &= 0 \\ 2x_2 v_2 + 2x_2\lambda &= 0 \\ &\ldots\ldots \\ 2x_M v_M + 2x_M\lambda &= 0 \\ {x_1}^2 + {x_2}^2 + \ldots + {x_M}^2 - 1 &= 0 \end{aligned}$$

The above system of equations has M solutions, with the $i-$th solution given by

$$x_i = 1,\ x_j = 0\ \ \forall\ \ j \neq i\ \ and\ \ \lambda = -v_i$$

The value of the objective function (B.2) corresponding to the $i-$th solution is v_i. Since the eigenvalues have been assumed to be ordered in ascending order, the function assumes its minimum value for the first solution, i.e.

$$x_1 = 1,\ x_j = 0\ \ for\ \ j = 2, \ldots, M$$

Thus, the sought solution to problem (B.1) is the eigenvector of $\mathbf{B}$ corresponding to the minimum eigenvalue. In practice, the minimum eigenvalue is computed using Jacobi transformations of the symmetric matrix $\mathbf{B}$ [200].

Appendix C

Proofs of Theorems

This Appendix is devoted to the proofs of Theorems C.1 and C.2 which have been employed in the developments of Chapters 4 and 6. Both Theorems are repeated below for completeness.

Theorem C.1 *Let $\mathcal{L}$ be the line defined by a pair of image points $\mathbf{p_1} = (x_1, y_1)$ and $\mathbf{p_2} = (x_2, y_2)$. Let also un_1 and un_2 denote the projections of the optical flow vectors at points $\mathbf{p_1}$ and $\mathbf{p_2}$ along the direction (n_x, n_y) that is normal to $\mathcal{L}$. Then, the difference of the two projections is independent of the α and β components of rotation and, assuming that the equation of $\mathcal{L}$ is $y = -\frac{n_x}{ny}x + \nu$, equal to*

$$un_1 - un_2 = [(x_1 - x_0)n_x + (y_1 - y_0)n_y]W(\frac{1}{Z_1} - \frac{1}{Z_2}) + (\frac{\nu\ n_x}{f}\alpha + \frac{\nu\ n_y}{f}\beta + \frac{\gamma}{n_y})(x_2 - x_1)$$

Proof. The projections of the optical flow vectors at points $\mathbf{p_1}$ and $\mathbf{p_2}$ on the vector (n_x, n_y) are equal to $un_i = u_i n_x + v_i n_y, \quad i = 1,\ 2$, which by substitution from Eqs. (2.7) yields

$$un_i \quad = \quad D_i\frac{W}{Z_i} + R_i^\alpha\alpha + R_i^\beta\beta + R_i^\gamma\gamma,$$

where $D_i = (x_i - x_0)n_x + (y_i - y_0)n_y$, $i = 1,\ 2$ and

$$R_i^\alpha \quad = \quad \frac{x_i y_i}{f}n_x + (\frac{y_i^2}{f} + f)n_y$$

$$
\begin{aligned}
R_i^{\beta} &= -(\frac{x_i^2}{f} + f)n_x - \frac{x_i y_i}{f} n_y \\
R_i^{\gamma} &= (y_i n_x - x_i n_y)
\end{aligned}
$$

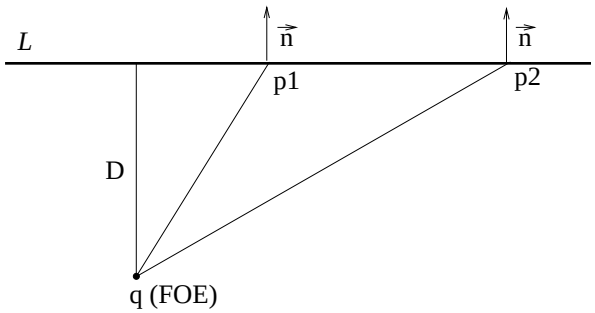


Figure C.1: The projections along $\vec{\mathbf{n}}$ of all vectors defined by the FOE $\mathbf{q}$ and some point on $\mathcal{L}$ are all equal to D.

As can be seen from Fig. C.1 where point $\mathbf{q}$ is assumed to be the FOE, $\mathbf{q}\overrightarrow{\mathbf{p_1}} \cdot \vec{\mathbf{n}} = \mathbf{q}\overrightarrow{\mathbf{p_2}} \cdot \vec{\mathbf{n}}$, i.e. $D_1 = D_2 = D$, which implies that the projection on $\vec{\mathbf{n}}$ of all vectors defined by the FOE and some point on $\mathcal{L}$ are equal to the distance D of $\mathbf{q}$ from $\mathcal{L}$. Therefore, the difference of the projections un_1 and un_2 after some algebraic manipulation can be shown to be equal to

$$un_1 - un_2 = DW(\frac{1}{Z_1} - \frac{1}{Z_2}) + (\frac{\nu\ n_x}{f}\alpha + \frac{\nu\ n_y}{f}\beta + \frac{\gamma}{n_y})(x_2 - x_1),$$

as ought to be shown.

□

A search scheme that exploits Theorem C.1 for recovering egomotion can be found in [148].

Theorem C.2 *Let* $\mathbf{p_1} = (x_1, y_1)$, $\mathbf{p_2} = (x_2, y_2)$ *and* $\mathbf{p_3} = (x_3, y_3)$ *be three collinear image points lying on a line whose equation is* $y = \kappa x + \nu$. *Let also* (x_0, y_0) *be the FOE and assume that* $\mathbf{p_2}$ *divides the vector* $\overrightarrow{\mathbf{p_1}\,\mathbf{p_3}}$ *in ratio* λ. *For the projections* $un_i, i = 1 \ldots 3$ *of the optical flow vectors at points* $\mathbf{p_1}$, $\mathbf{p_2}$ *and* $\mathbf{p_3}$ *along an arbitrary direction* (n_x, n_y), *the following equation holds*

$$un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3 = D_2W(\frac{1}{Z_2} - \frac{1}{1+\lambda}\frac{1}{Z_1} - \frac{\lambda}{1+\lambda}\frac{1}{Z_3}) + \frac{d_{21}}{1+\lambda}W(\frac{1}{Z_1} - \frac{1}{Z_3}) +$$

$$\frac{\kappa d_{21}(x_2 - x_3)}{f}\alpha - \frac{d_{21}(x_2 - x_3)}{f}\beta$$

In the above equation, $D_2 = (x_2 - x_0)n_x + (y_2 - y_0)n_y$ *and* $d_{21} = (x_2 - x_1)n_x + (y_2 - y_1)n_y$.

Proof. Due to the separability of the translational and rotational flow components, the linear sum of the flow projections can be written as

$$un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3 = (T_2 - \frac{T_1}{1+\lambda} - \frac{\lambda T_3}{1+\lambda}) + (R_2 - \frac{R_1}{1+\lambda} - \frac{\lambda R_3}{1+\lambda}),$$

where T_i and R_i are the projections of the translational and rotational part for point i respectively. More specifically,

$$T_i = D_i\frac{W}{Z_i} \quad with \quad D_i = (x_1 - x_0)n_x + (y_1 - y_0)n_y$$

and, employing the notation of Theorem C.1,

$$R_i = R_i^\alpha\alpha + R_i^\beta\beta + R_i^\gamma\gamma \ , \quad i = 1 \ldots 3$$

Noting that $D_1 = D_2 - d_{21}$ and $D_3 = D_2 + \frac{d_{21}}{\lambda}$, the sum of the translational parts can be expressed as

$$T_2 - \frac{T_1}{1+\lambda} - \frac{\lambda T_3}{1+\lambda} = \frac{D_2W}{Z_2} - \frac{1}{1+\lambda}\frac{D_1W}{Z_1} - \frac{\lambda}{1+\lambda}\frac{D_3W}{Z_3}$$

$$= D_2W(un_2 - \frac{1}{1+\lambda}un_1 - \frac{\lambda}{1+\lambda}un_3) + \frac{d_{21}}{1+\lambda}W(\frac{1}{Z_1} - \frac{1}{Z_3})$$

Regarding the sum of rotational parts, some algebraic manipulation reveals that it is equal to

$$R_2 - \frac{R_1}{1+\lambda} - \frac{\lambda R_3}{1+\lambda} \quad = \quad \frac{\kappa d_{21}(x_2 - x_3)}{f}\alpha - \frac{d_{21}(x_2 - x_3)}{f}\beta$$

Adding the right hand sides of the two equations for the translational and rotational parts above, yields the desired result.

□

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 Trans. on PAMI*, 7(4):384--401, Jul. 1985.

[3] P.E. Agre. *The Dynamic Structure of Everyday Life*. Cambridge University Press, 1991.

[4] J. Aloimonos and A. Badyopadhyay. Active Vision. In *Proceedings of ICCV'87*, pages 35--54, Jun. 1987.

[5] J. Aloimonos and C.M. Brown. Direct Processing of Curvilinear Motion from a Sequence of Perspective Images. In *Proceedings of Workshop on Computer Vision: Representation and Control*, pages 72--77, 1984.

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

[7] Y. Aloimonos and C. M. Brown. On the Kinetic Depth Effect. *Biological Cybernetics*, 60(6):445--455, 1989.

[8] Y. Aloimonos and Z. Duric. Estimating the Heading Direction Using Normal Flow. *IJCV*, 13(1):33--56, 1994.

[9] P. Anandan. A Computational Framework and an Algorithm for the Measurement of Visual Motion. *IJCV*, 2:283--310, 1989.

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

[11] N. Ancona and T. Poggio. Optical Flow from 1D Correlation: Application to a Simple Time-to-Crash Detector. *IJCV*, 14(2):131--146, Mar. 1995.

[12] H. Araujo, J. Batista, P. Peixoto, and J. Dias. Pursuit Control in a Binocular Active Vision System Using Optical Flow. In *Proceedings of ICPR'96*, 1996.

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

[14] A. Argyros. *Visual Detection of Independent 3D Motion by a Moving Observer.* PhD Dissertation, Department of Computer Science, University of Crete, Heraklion, Crete, Creece, 1996.

[15] A.A. Argyros and F. Bergholm. Reactive Robot Navigation Based on a Combination of Central and Peripheral Vision. In *VIRGO/SMART/MobiNet EU TMR Joint Workshop*, Santorini, Greece, Sep. 1998.

[16] 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 *Proceedings of BMVC'96, Edinburgh, UK*, Sep. 9-12 1996.

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

[18] A.A. Argyros and S.C. Orphanoudakis. Independent 3D Motion Detection Based on Depth Elimination in Normal Flow Fields. In *Proceedings of CVPR '97, San Juan, Puerto Rico*, pages 672--677, Jun. 17-19 1997.

[19] A.A. Argyros, P.E. Trahanias, and S.C. Orphanoudakis. Robust Regression for the Detection of Independent 3D Motion by a Binocular Observer. *Journal of Real Time Imaging*, 4(2):125--141, Apr. 1998.

[20] J. Arnspang. Optic Acceleration. In *Proceedings of ICCV'88*, pages 364--373, 1988.

[21] J. Arnspang, K. Henriksen, and R. Stahr. Estimating Time to Contact with Curves, Avoiding Calibration and Aperture Problem. In *Proceedings of CAIP'95*, pages 856--861, 1995.

[22] S. Ayer, P. Schroeter, and J. Bigun. Segmentation of Moving Objects by Robust Motion Parameter Estimation over Multiple Frames. In *Proceedings of ECCV'96*, pages 316--327, 1994.

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

[24] 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.

[25] D.H. Ballard and O.A. Kimball. Rigid Body Motion from Depth and Optical Flow. *CVGIP*, 22:95--115, 1983.

[26] C. Bard, C. Laugier, C. Milesibellier, J. Troccaz, B. Triggs, and G. Vercelli. Achieving Dextrous Grasping by Integrating Planning and Vision-Based Sensing. *International Journal of Robotics Research*, 14(5):445--464, Oct. 1995.

[27] S.T. Barnard and W.B. Thompson. Disparity Analysis of Images. *IEEE Trans. on PAMI*, PAMI-2(4):333--340, Jul. 1980.

[28] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of Optical Flow Techniques. *IJCV*, 12(1):43--77, 1994.

[29] V. Beranger and J. Herve. Recognition of Intersections in Corridors Environment. In *Proceedings of ICPR'96*, 1996.

[30] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. Hierarchical Model-Based Motion Estimation. In *Proceedings of ECCV'92*, pages 237--252, 1992.

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

[32] P.J. Besl, R.C. Jain, and L.T. Watson. Robust Window Operators. In *Proceedings of ICCV'88*, pages 591--600, 1988.

[33] A. Del Bimbo, P. Nesi, and J.L.C. Sanz. Analysis of Optical Flow Constraints. *IEEE Trans. on Image Processing*, 4(4):460--469, Apr. 1995.

[34] 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, 1996.

[35] 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, Dec. 1994.

[36] P. Bouthemy and E. Francois. Motion Segmentation and Qualitative Dynamic Scene Analysis from an Image Sequence. *IJCV*, 10(2):157--182, 1993.

[37] 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 Trans. on PAMI*, PAMI-16:987--1001, 1994.

[38] K.J. Bradshaw, P.F. McLauchlan, I.D. Reid, and D.W. Murray. Saccade and Pursuit on an Active Head Eye Platform. *Image and Vision Computing*, 12(3):155--163, Apr. 1994.

[39] V. Braitenberg. *Vehicles - Experiments in Synthetic Psychology*. MIT Press, Cambridge, MA, 1984.

[40] R.A. Brooks. A Robust Layered Control System For a Mobile Robot. *IEEE J. Robotics Automation*, RA-2(7):14--23, Apr. 1986.

[41] R.A. Brooks. Intelligence Without Reason. Technical Report AILAB Memo 1293, Massachusetts Institute of Technology Artificial Intelligence Laboratory, Apr. 1991.

[42] R.A. Brooks. Intelligence Without Representation. *Artificial Intelligence*, 47:139--160, 1991.

[43] J. Bruske, M. Hansen, L. Riehn, and G. Sommer. Biologically Inspired Calibration Free Adaptive Saccade Control of a Binocular Camera Head. *Biological Cybernetics*, 77(6):433--446, Dec. 1997.

[44] A. R. Bruss and B.K.P. Horn. Passive Navigation. *CVGIP*, 21:3--20, 1983.

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

[46] P.J. Burt, C. Yen, and X. Xu. Multiresolution Flow Through Motion Analysis. In *Proceedings of CVPR'83*, pages 246--252, 1983.

[47] T. Camus, D. Coombs, M. Herman, and T. Hong. Real-Time Single-Workstation Obstacle Avoidance Using Only Wide-Field Flow Divergence. In *Proceedings of ICPR'96*, pages 323--330, 1996.

[48] S. Carlsson and J.-O. Eklundh. Object Detection Using Model Based Prediction and Motion Parallax. In *Proceedings of ECCV'90, LNCS*, pages 134--138, 1990.

[49] B.A. Cartwright and T.S. Collett. Landmark Maps for Honeybees. *Biological Cybernetics*, 57:85--93, 1987.

[50] R. Cipola, Y. Okamoto, and Y. Kuno. Robust Structure from Motion Using Motion Parallax. In *Proceedings of ICCV'93*, pages 374--382, 1993.

[51] R. Cipolla and A. Blake. Surface Orientation and Time to Contact from Image Divergence and Deformation. In *Proceedings of ECCV'92*, pages 187--202, 1992.

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

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

[54] T.S. Collett, S.N. Fry, and R. Wehner. Sequence Learning by Honeybees. *Journal of Comparative Physiology A*, 172:693--706, 1993.

[55] C. Colombo, B. Allotta, and P. Dario. Affine Visual Servoing for Robot Relative Positioning and Landmark-Based Docking. *Advanced Robotics*, 9(4):463--480, 1995.

[56] D. Coombs, M. Herman, T. Hong, and M. Nashman. Real-Time Obstacle Avoidance Using Central Flow Divergence and Peripheral Flow. In *Proceedings of ICCV'95*, pages 276--283, 1995.

[57] J.R. Cooperstock and E.E. Milios. Self-Supervised Learning for Docking and Target Reaching. In *Proceedings of IAS'93*, 1993.

[58] T.H. Cormen, C.H. Leiserson, and R.L. Rivest. *Introduction to Algorithms.* MIT Press, Cambridge, MA, 1990.

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

[60] A. Criminisi, I. Reid, and A. Zisserman. Duality, Rigidity and Planar Parallax. In *Proceedings of ECCV'98*, 1998.

[61] S.M. Culhane and J.K. Tsotsos. An Attentional Prototype for Early Vision. In *Proceedings of ECCV'92*, pages 551--560, 1992.

[62] K. Daniilidis. Fixation Simplifies 3D Motion Estimation. *CVIU*, 68(2):158--169, Nov. 1997.

[63] K. Daniilidis and E. Bayro-Corrochano. The Dual Quaternion Approach to Hand-Eye Calibration. In *Proceedings of ICPR'96*, 1996.

[64] K. Daniilidis and H.-H. Nagel. Analytical Results on Error Sensitivity of Motion Estimation from two Views. *Image and Vision Computing*, 8:297--303, 1990.

[65] K. Daniilidis and M.E. Spetsakis. Understanding Noise Sensitivity in Structure From Motion. In Y. Aloimonos, editor, *Visual Navigation: From Biological Systems to Unmanned Ground Vehicles*, chapter 4. Lawrence Erlbaum Associates, Hillsdale, NJ, 1997.

[66] 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*. DeKluwer Pubs, Boston, 1989.

[67] U.R. Dhond and J.K. Aggarwal. Structure From Stereo - A Review. *IEEE Trans. on SMC*, SMC-19:1489--1510, 1989.

[68] W. Enkelmann. Obstacle Detection by Evaluation of Optical Flow Fields From Image Sequences. *Image and Vision Computing*, 9(3):160--168, 1991.

[69] A. Argyros et al. Analysis of Current Approaches in Automated Vision-based Navigation. VIRGO TMR task 1 report.

[70] H.R. Everett. *Sensors for Mobile Robots: Theory and Application*. A K Peters, Wellesley, MA, 1995.

[71] O. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, Cambridge, MA, 1993.

[72] O. Faugeras. Stratification of 3-D Vision: Projective, Affine, and Metric Representations. *Journal of the Optical Society of America A*, 12(3):465--484, Mar. 1995.

[73] O.D. Faugeras. What Can Be Seen in Three Dimensions with an Uncalibrated Stereo Rig? In *Proceedings of ECCV'92*, pages 563--578, 1992.

[74] O.D. Faugeras and S.J. Maybank. Motion from Point Matches: Multiplicity of Solutions. *IJCV*, 4:225--246, 1990.

[75] S. Fejes and L.S. Davis. Direction-Selective Filters for Egomotion Estimation. Technical Report CS-TR-3814, Center for Automation Research, University of Maryland, Jul. 1997.

[76] S. Fejes and L.S. Davis. What Can Projections of Flow Fields Tell Us About the Visual Motion. In *Proceedings of ICCV'98*, pages 979--986, 1998.

[77] J. A. Feldman. Four Frames Suffice: A Provisional Model of Vision and Space. *The Behavioral and Brain Sciences*, 8:265--289, 1985.

[78] C.L. Fennema and W. Thomson. Velocity Determination in Scenes Containing Several Moving Objects. *CGIP*, 9:301--315, 1979.

[79] C. Fermüller. Passive Navigation as a Pattern-Recognition Problem. *IJCV*, 14(2):147--158, Mar. 1995.

[80] C. Fermüller and Y. Aloimonos. Vision and Action. *Image and Vision Computing*, 13(10):725--744, Dec. 1995.

[81] C. Fermüller and Y. Aloimonos. The Confounding of Translation and Rotation in Reconstruction from Multiple Views. In *Proceedings of CVPR'97*, pages 250--256, 1997.

[82] 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.

[83] D.J. Fleet and A.D. Jepson. Computation of Component Image Velocity from Local Phase Information. *IJCV*, pages 77--104, 1990.

[84] P. Flynn. The MSU/WSU Range Image Database, 1989. Available online at http://www.eecs.wsu.edu/IRL/RID/.

[85] P. Fornland. Direct Obstacle Detection and Motion from Spatio-Temporal Derivatives. In *Proceedings of CAIP'95, LNCS*, pages 874--879, Prague, Sep. 1995.

[86] P. Fornland and C. Schnörr. A Robust and Convergent Iterative Approach for Determining the Dominant Plane from Two Views Without Correspodence and Calibration. In *Proceedings of CVPR'97*, pages 508--513, Puerto Rico, Jun. 1997.

[87] C.S. Fuh and P. Maragos. Motion Displacement Estimation Using an Affine Model for Image Matching. *Optical Engineering*, 30(7):881--887, Jul. 1991.

[88] B. Galvin, B. McCane, K. Novins, D. Mason, and S. Mills. Recovering Motion Fields: An Evaluation of Eight Optical Flow Algorithms. In *Proceedings of BMVC'98, Southampton, UK*, volume 1, pages 195--204, Sep. 14-17 1998.

[89] J.J. Gibson. *The Perception of the Visual World.* Houghton-Mifflin, Boston, 1950.

[90] G.H. Golub and C.F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, 3rd edition, 1996.

[91] W.E.L. Grimson. Computational Experiments with a Feature Based Stereo Algorithm. *IEEE Trans. on PAMI*, 7(1):17--34, Jan. 1985.

[92] E. Grossmann and J. Santos-Victor. The Precision of 3D Reconstruction from Uncalibrated Views. In *Proceedings of BMVC'98, Southampton, UK*, volume 1, pages 115--124, Sep. 14-17 1998.

[93] E. Grosso, G. Metta, A. Oddera, and G. Sandini. Robust Visual Servoing in 3-D Reaching Tasks. *IEEE Trans. on Robotics and Automation*, 12(5):732--742, Oct. 1996.

[94] K.J. Hanna. Direct Multi-Resolution Estimation of Ego-Motion and Structure from Motion. In *IEEE Workshop on Visual Motion*, pages 156--162, 1991.

[95] R. Hartley and R. Gupta. Computing Matched-Epipolar Projections. In *Proceedings of CVPR'93*, pages 549--555, 1993.

[96] R. Hartley, R. Gupta, and T. Chang. Stereo from Uncalibrated Cameras. In *Proceedings of CVPR'92*, pages 761--764, 1992.

[97] R.I. Hartley. In Defense of the 8-Point Algorithm. In *Proceedings of ICCV'95*, pages 1064--1070, 1995.

[98] R.I. Hartley. In Defense of the 8-Point Algorithm. *IEEE Trans. on PAMI*, PAMI-19(6):580--593, Jun. 1997.

[99] D. Heeger. Optical Flow Using Spatiotemporal Filters. *IJCV*, 1:279--302, 1988.

[100] D.J. Heeger and A.D. Jepson. Subspace Methods for Recovering Rigid Motion I: Algorithm and Implementation. *IJCV*, 7(2):95--117, 1992.

[101] E.C. Hildreth. Computations Underlying the Measurement of Visual Motion. *Artificial Intelligence*, 23(3):309--355, 1984.

[102] E.C. Hildreth and C. Koch. The Analysis of Visual Motion: From Computational Theory to Neuronal Mechanisms. *Ann. Rev. Neurosci.*, 10:477--533, 1987.

[103] J. Hochberg. Machines Should Not See as People Do, but Must Know How People See. *CVGIP*, 37:221--237, 1987.

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

[105] M. Holder, M. Trivedi, and S. Marapane. Mobile Robot Navigation by Wall Following Using a Rotating Ultrasonic Scanner. In *Proceedings of ICPR'96*, 1996.

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

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

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

[109] B.K.P. Horn and E.J. Weldon. Direct Methods for Recovering Motion. *IJCV*, 2:51--76, 1988.

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

[111] A. Horridge. Bee Vision of Pattern and 3D. The Bidder Lecture, delivered at the Annual Meeting of the Society for Experimental Biology at Swansea, Apr. 1994.

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

[113] T.S. Huang and A.N. Netravali. Motion and Structure from Feature Correspondences: A Review. *Proceedings of the IEEE*, 82(2):252--268, Feb. 1994.

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

[115] R. Hummel and V. Sundareswaran. Motion Parameter Estimation from Global Flow Field Data. *IEEE Trans. on PAMI*, 15(5):459--476, May 1993.

[116] S. Hutchinson, G.D. Hager, and P.I. Corke. A Tutorial on Visual Servo Control. *IEEE Trans. on Robotics and Automation*, 12(5):651--670, Oct. 1996.

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

[118] M. Irani and P. Anandan. A Unified Approach to Moving Object Detection in 2D and 3D Scenes. In *Proceedings of ICPR'96*, pages 712--717, Vienna, Austria, 1996.

[119] M. Irani, B. Rousso, and S. Peleg. Computing Occluding and Transparent Motions. *IJCV*, 12(1):5--16, 1994.

[120] M. Irani, B. Rousso, and S. Peleg. Recovery of Ego-Motion Using Region Alignment. *IEEE Trans. on PAMI*, 19(3):268--272, Mar. 1997.

[121] R. Jain, R. Kasturi, and B.G. Schunck. *Machine Vision*. McGraw-Hill, NY, 1995.

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

[123] M.R.M. Jenkin and A. Jepson. Detecting Floor Anomalies. In *Proceedings of BMVC'94*, pages 731--740, 1994.

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

[125] D.G. Jones and J. Malik. Computational Framework for Determining Stereo Correspondence from a Set of Linear Spatial Filters. *IVC*, 10(10):699--708, Dec. 1992.

[126] J. Jones and A. Flynn. *Mobile Robots: Inspirations to Implementation*. A K Peters, Wellesley, MA, 1993.

[127] T. Kanade, M.L. Reed, and L.E. Weiss. New Technologies and Applications in Robotics. *CACM*, 37(3):58--67, Mar. 1994.

[128] K. Kanatani. Computational Cross Ratio for Computer Vision. *CVGIP*, 60(3):371--381, 1994.

[129] K.I. Kanatani. Structure and Motion from Optical Flow Under Orthographic Projection. *CVGIP*, 35(2):181--199, Aug. 1986.

[130] K.I. Kanatani. Structure and Motion from Optical Flow Under Perspective Projection. *CVGIP*, 38(2):122--146, May 1987.

[131] K.I. Kanatani. Computational Projective Geometry. *CVGIP*, 54(3):333--348, Nov. 1991.

[132] K.I. Kanatani. Hypothesizing and Testing Geometric Properties of Image Data. *CVGIP*, 54(3):349--357, Nov. 1991.

[133] I. Kant. *Critique of Pure Reason*. Prometheus Books, Buffalo, NY, 1990.

[134] J.J. Koenderink and A.J. van Doorn. Affine stucture from motion. *Journal of the Optical Society of America*, 8:377--385, 1991.

[135] J. Košecka, H.I. Christensen, and R. Bajcsy. Discrete-Event Modeling of Visually Guided Behaviors. *IJCV*, 14(2):179--191, Mar. 1995.

[136] R. Kumar, P. Anandan, and K. Hanna. Direct Recovery of Shape from Multiple Views: A Parallax Based Approach. In *Proceedings of ICPR'94*, pages 685--688, Jerusalem, Israel, 1994.

[137] S. R. Kundur, D. Raviv, and E. Kent. An Image Based Visual Motion Cue For Autonomous Navigation. In *Proceedings of CVPR'97*, pages 7--14, 1997.

[138] Y. Lamdan, J.T. Schwartz, and H.J. Wolfson. Object Recognition by Affine Invariant Matching. In *Proceedings of CVPR'88*, pages 235--344, 1988.

[139] T. S. Levitt and D. T. Lawton. Qualitative Navigation for Mobile Robots. *AI Journal*, 44(3):305--360, Aug. 1990.

[140] L. Li and J. Duncan. 3-D Translational Motion and Structure from Binocular Image Flows. *IEEE Trans. on PAMI*, PAMI-15(7):657--667, Jul. 1993.

[141] N. V. Lobo and J. K. Tsotsos. Computing Egomotion and Detecting Independent Motion from Image Motion Using Collinear Points. *CVIU*, 64(1):21--52, Jul. 1996.

[142] H.C. Longuet-Higgins. A Computer Algorithm for Reconstructing a Scene from two Projections. *Nature*, 293:133--135, 1981.

[143] H.C. Longuet-Higgins. The Visual Ambiguity of a Moving Plane. In *Proceedings of the Royal Society*, volume 223, pages 165--175. London B, 1984.

[144] 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.

[145] M.I.A. Lourakis. Establishing Straight Line Correspondence. Technical Report 208, ICS/FORTH, Aug. 1997.

[146] M.I.A. Lourakis, A.A. Argyros, and S.C. Orphanoudakis. Independent 3D Motion Detection Using Residual Parallax Normal Flow Fields. In *Proceeedings of ICCV'98*, pages 1012--1017, Bombay, India, Jan. 1998.

[147] M.I.A. Lourakis, S.T. Halkidis, and S.C. Orphanoudakis. Matching Disparate Views of Planar Surfaces Using Projective Invariants. In *Proceedings of BMVC'98, Southampton, UK*, volume 1, pages 94--104, Sep. 14-17 1998.

[148] M.I.A. Lourakis and S.C. Orphanoudakis. Egomotion Estimation Using FOE Constraint Lines. In *VIRGO/SMART/MobiNet EU TMR Joint Workshop*, pages 107--114, Santorini, Greece, Sep. 1998.

[149] M.I.A. Lourakis and S.C. Orphanoudakis. Using Planar Parallax to Estimate the Time-to-Contact. In *VIRGO/SMART/MobiNet EU TMR Joint Workshop*, pages 123--130, Santorini, Greece, Sep. 1998.

[150] M.I.A. Lourakis and S.C. Orphanoudakis. Visual Detection of Obstacles Assuming a Locally Planar Ground. In *Proceedings of ACCV'98, LNCS No. 1352*, volume 2, pages 527--534, Hong Kong, China, Jan. 1998.

[151] B.D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings DARPA IU Workshop*, pages 121--130, 1981.

[152] Q.T. Luong and O.D. Faugeras. Determining the Fundamental Matrix with Planes: Instability and New Algorithms. In *Proceedings of CVPR'93*, pages 489--494, 1993.

[153] Q.T. Luong and O.D. Faugeras. The Fundamental Matrix: Theory, Algorithms and Stability Analysis. *IJCV*, 17(1):43--76, Jan. 1996.

[154] W.J. MacLean, A.D. Jepson, and R.C. Frecker. Recovery of Egomotion and Segmentation of Independent Motion Using the EM Algorithm. In *Proceedings of BMVC'94*, 1994.

[155] P. Maes. Modeling Adaptive Autonomous Agents. *Artificial Life Journal*, 1(1&2):135--162, 1994.

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

[157] L. Matthies, R. Szeliski, and T. Kanade. Kalman Filter-Based Algorithms for Estimating Depth from Image Sequences. *IJCV*, 3(3):209--238, Sep. 1989.

[158] L.H. Matthies. Toward Stochastic Modeling of Obstacle Detectability in Passive Stereo Range Imagery. In *Proceedings of CVPR'92*, 1992.

[159] S.J. Maybank and O.D. Faugeras. A Theory of Self-Calibration of a Moving Camera. *IJCV*, 8(2):123--151, 1992.

[160] C. Medioni and R. Nevatia. Matching Images Using Linear Features. *IEEE Trans. on PAMI*, PAMI-6(6):675--686, 1984.

[161] P. Meer, R. Lenz, and S. Ramakrishna. Efficient Invariant Representations. *IJCV*, 26(2):137--152, Feb. 1998.

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

[163] F.G. Meyer. Time-to-Collision from First-Order Models of the Motion Field. *IEEE Trans. on RA*, 10(6):792--798, 1994.

[164] E. De Micheli, V. Torre, and S. Uras. The Accuracy of the Computation of Optical Flow and of the Recovery of Motion Parameters. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-15(5):434--447, May 1993.

[165] A. Mitiche and P. Bouthemy. Computation and Analysis of Image Motion: A Synopsis of Current Problems and Methods. *IJCV*, 19(1):29--55, Jul. 1996.

[166] R. Mohr. Projective Geometry and Computer Vision. In C.H. Chen, L.F. Pau, and P.S.P. Wang, editors, *Handbook of Pattern Recognition and Computer Vision*, chapter 2.4, pages 369--393. World Scientific, River Edge, NJ, 1993.

[167] R. Mohr and B. Triggs. Projective Geometry for Image Analysis. A tutorial given at ISPRS, Vienna, July 1996. Available online at http://www-kogs.iitb.fhg.de/~cveducat/ECV_Tut_Proj_Geom/ProjGeometry.html.

[168] J.L. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision*. MIT Press, Cambridge, MA, 1992.

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

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

[171] H.-H. Nagel. On the Estimation of Optical Flow: Relations between Different Approaches and Some New Results. *AI Journal*, 33(3):299--324, Nov. 1987.

[172] H.-H. Nagel and W. Enkelmann. An Investigation of Smoothness Constraints for the Estimation of Displacement Vector Fields from Image Sequences. *IEEE Trans. on PAMI*, PAMI-8(9):565--593, Sept. 1986.

[173] K. Nakayama. Biological Image Motion Processing: A Review. *Vision Research*, 25(5):625--660, 1985.

[174] V.S. Nalwa. *A Guided Tour of Computer Vision*. Addison-Wesley, NY, 1993.

[175] S. Negahdaripour. Direct Computation of the FOE with Confidence Measures. *CVIU*, 64(3):323--350, 1996.

[176] S. Negahdaripour and B.K.P. Horn. A Direct Method for Locating the Focus of Expansion. *CVGIP*, 46:303--326, 1989.

[177] R.C. Nelson. Qualitative Detection of Motion by a Moving Observer. *IJCV*, 7(1):33--46, 1991.

[178] R.C. Nelson. Visual Homing Using an Associative Memory. *Biological Cybernetics*, 65:281--291, 1991.

[179] R.C. Nelson and J. Aloimonos. Finding Motion Parameters from Spherical Flow Fields (or the Advantages of Having Eyes in the Back of Your Head). *Biological Cybernetics*, 58:261--273, 1988.

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

[181] B. Nilsson, J. Nygards, and A. Wernersson. On-Range Sensor Feedback for Mobile Robot Docking within Prescribed Posture Tolerances. *Journal of Robotic Systems*, 14(4):297--312, Apr. 1997.

[182] 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.

[183] J.M. Odobez and P. Bouthemy. Detection of Multiple Moving Ojects Using Multiscale MRF with Camera Motion Compensation. In *Proceedings of ICIP'94*, Austin, TX, 1994.

[184] Y. Ohta and T. Kanade. Stereo by Intra-and Inter-Scanline Search Using Dynamic Programming. *IEEE Trans. on PAMI*, 7(2):139--154, Mar. 1985.

[185] T.J. Olson and D. Coombs. Real-Time Vergence Control for Binocular Robots. *IJCV*, 7(1):67--89, Nov. 1991.

[186] The Columbia OmniCamera web site: http://www.cs.columbia.edu/cave/omnicam/.

[187] M. Otte and H.-H. Nagel. Optical Flow Estimation: Advances and Comparisons. In *Proceedings of ECCV'94, LNCS*, pages 51--60, 1994.

[188] K. Pahlavan and J.O. Eklundh. A Head-Eye System: Analysis and Design. *CVGIP*, 56(1):41--56, Jul. 1992.

[189] 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.

[190] K. Pahlavan, T. Uhlin, and J.O. Eklundh. Dynamic Fixation and Active Perception. *IJCV*, 17(2):113--135, Feb. 1996.

[191] N.P. Papanikolopoulos. Selection of Features and Evaluation of Visual Measurements During Robotic Visual Servoing Tasks. *Journal of Intelligent and Robotic Systems*, 13(3):279--304, Jul. 1995.

[192] N.P. Papanikolopoulos, B.J. Nelson, and P.K. Khosla. 6-Degree-of-Freedom Hand Eye Visual Tracking with Uncertain Parameters. *IEEE Trans. on Robotics and Automation*, 11(5):725--732, Oct. 1995.

[193] N. Paragios and G. Tziritas. Detection and Location of Moving Objects Using Deterministic Relaxation Algorithms. In *Proceedings of ICPR'96*, volume 1, pages 201--205, Vienna, Austria, 1996.

[194] S.C. Pei and L.G. Liou. What Can Be Seen in a Noisy Optical Flow Field Projected by a Moving Planar Patch in 3D Space. *Pattern Recognition*, 30(9):1401--1413, 1997.

[195] T. Poggio, V. Torre, and C. Koch. Computational Vision and Regularization Theory. *Nature*, 317:314--319, 1985.

[196] S.B. Pollard, J.E.W. Mayhew, and J.P. Frisby. PMF: a Stereo Correspondence Algorithm Using a Disparity Gradient Constraint. *Perception*, 14:449--470, 1985.

[197] K. Prazdny. Egomotion and Relative Depth from Optical Flow. *Biological Cybernetics*, 36:87--102, 1980.

[198] K. Prazdny. Determining the Instantaneous Direction of Motion From Optical Flow Generated by a Curvilinearly Moving Observer. *CVGIP*, 17:238--248, 1981.

[199] K. Prazdny. On the Information in Optical Flows. *CVGIP*, 22:239--259, 1983.

[200] W.H. Press, S.A. Teukolsky, A.W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, 1992.

[201] P. Pritchett and A. Zisserman. Wide Baseline Stereo Matching. In *Proceedings of ICCV'98*, pages 754--760, Bombay, India, Jan. 1998.

[202] K. Rangarajan and M. Shah. Interpretation of Motion Trajectories Using Focus of Expansion. *IEEE Trans. on PAMI*, 14(12):1205--1210, Dec. 1992.

[203] R.P.N. Rao and D.H. Ballard. An Active Vision Architecture Based on Iconic Representations. *AI Journal*, 78(1-2):461--505, Oct. 1995.

[204] The REALiZ web site: http://www.realiz.com.

[205] J.H. Reiger and D.T. Lawton. Processing Differential Image Motion. *Journal of the Optical Society of America A*, 2:354--359, 1985.

[206] D.L. Ringach and Y. Baram. A Diffusion Mechanism for Obstacle Detection from Size-Change Information. *IEEE Trans. on PAMI*, PAMI-16(1):76--80, Jan. 1994.

[207] L. Robert and O.D. Faugeras. Relative 3D Positioning and 3D Convex Hull Computation from a Weakly Calibrated Stereo Pair. *Image and Vision Computing*, 13(3):189--196, Apr. 1995.

[208] A. Rosenfeld. *Multiresolution Image Processing and Analysis*. Springer-Verlag, 1984.

[209] P.J. Rousseeuw. Least Median of Squares Regression. *Journal of American Statistics Association*, 79:871--880, 1984.

[210] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. John Wiley and Sons Inc., New York, 1987.

[211] B. Rousso, S. Avidan, A. Shashua, and S. Peleg. Robust Recovery of Camera Rotation from Three Frames. In *Proceedings of CVPR'96*, 1996.

[212] J. Santos-Victor and G. Sandini. Uncalibrated Obstacle Detection Using Normal Flow. *Machine Vision and Applications*, 9(3):130--137, 1996.

[213] J. Santos-Victor and G. Sandini. Visual Behaviors for Docking. *CVIU*, 67(3):223--238, Sep. 1997.

[214] J. Santos-Victor, G. Sandini, F. Curotto, and S. Garibaldi. Divergent Stereo in Autonomous Navigation: From Bees to Robots. *IJCV*, 14(2):159--177, Mar. 1995.

[215] H. Sawhney. Simplifying Motion and Structure Analysis Using Planar Parallax and Image Warping. In *Proceedings of ICPR'94*, pages 403--408, Jerusalem, Israel, 1994.

[216] C. Schmid and A. Zisserman. Automatic Line Matching Across Views. In *Proceedings of CVPR'97*, pages 666--671, 1997.

[217] I. Schmidt, T.S. Collett, F.X. Dillier, and R. Wehner. How Desert Ants Cope with Enforced Detours on Their Way Home. *Journal of Comparative Physiology A*, 171:285--288, 1992.

[218] R. Sedgewick. *Algorithms.* Addison-Wesley, Reading, MA, 1988.

[219] T. Shakunaga. 3-D Corridor Scene Modeling from a Single View under Natural Lighting Conditions. *IEEE Trans. on PAMI*, 14(2):293--298, Feb. 1992.

[220] R. Sharma and Y. Aloimonos. Early Detection of Independent Motion from Active Control of Normal Image Flow Patterns. *IEEE Trans. on SMC*, SMC-26(1):42--53, Feb. 1996.

[221] A. Shashua. Algebraic Functions for Recognition. *IEEE Trans. on PAMI*, PAMI-17(8):779--789, Aug. 1995.

[222] C. Silva and J. Santos-Victor. Robust Egomotion Estimation from the Normal Flow Using Search Subspaces. *IEEE Trans. on PAMI*, 19(9):1026--1034, Sep. 1997.

[223] E. Simoncelli. *Distributed Representation and Analysis of Visual Motion.* PhD Dissertation, Electrical Engineering and Computer Science Department, MIT, 1993.

[224] D. Sinclair. Motion Segmentation and Local Structure. In *Proceedings of ICCV'93*, pages 366--373, 1993.

[225] D. Sinclair and A. Blake. Quantitative Planar Region Detection. *IJCV*, 18(1):77--91, Apr. 1996.

[226] D. Sinclair, A. Blake, and D. Murray. Robust Estimation of Egomotion from Normal Flow. *IJCV*, 13:57--69, 1994.

[227] D. Sinclair, H. Christensen, and C. Rothwell. Using the Relation Between a Plane Projectivity and the Fundamental Matrix. In *Proceedings of SCIA'95*, 1995.

[228] A. Singh. *Optical Flow Computation: A Unified Perspective*. PhD Dissertation, Department of Computer Science, Columbia University, 1990.

[229] S.S. Sinha and B.G. Schunck. A Two Stage Algorithm for Discontinuity-Preserving Surface Reconstruction. *IEEE Trans. on PAMI*, PAMI-14:36--55, 1992.

[230] K. Skifstad and R. Jain. Illumination Independent Change Detection for Real World Image Sequences. *Computer Vision, Graphics and Image Processing*, 46:387--399, 1989.

[231] S. M. Smith and J. M. Brady. SUSAN - A New Approach to Low Level Image Processing. *IJCV*, 23(1):45--78, May 1997.

[232] M.E. Spetsakis and Y. Aloimonos. Optimal Motion Estimation. In *IEEE Workshop on Visual Motion*, pages 229--237, 1989.

[233] M.E. Spetsakis and Y. Aloimonos. Structure from Motion Using Line Correspondences. *International Journal of Computer Vision*, 4:171--183, 1990.

[234] A. Stein and M. Werman. Robust Statistics in Shape Fitting. *Computer Vision, Graphics and Image Processing*, pages 540--546, 1992.

[235] C.V. Stewart. MINPRAN: A New Robust Estimator for Computer Vision. *PAMI*, 17(10):925--938, Oct. 1995.

[236] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer-Verlag, 1992.

[237] M. Subbarao. Interpretation of Image Flow: A Spatio-temporal Approach. *IEEE Trans. on PAMI*, 11(3):266--278, 1989.

[238] M. Subbarao. Bounds on Time-to-Collision and Rotational Component from First-Order Derivatives of Image Flow. *CVGIP*, 50(3):329--341, Jun. 1990.

[239] M. Subbarao and A.M. Waxman. Closed Form Solutions to Image Flow Equations for Planar Surfaces in Motion. *CVGIP*, 36(2/3):208--228, Nov./Dec. 1986.

[240] B.J. Super and W.N. Klarquist. Patch-based Stereo in a General Binocular Viewing Geometry. *IEEE Trans. on PAMI*, 19(3):247--253, Mar. 1997.

[241] M. J. Swain and M. A. Stricker. Promising Directions in Active Vision. *International Journal of Computer Vision*, 11(2):109--126, 1993.

[242] M.J. Swain. and D.H. Ballard. Color Indexing. *IJCV*, 7(1):11--32, Nov. 1991.

[243] M.A. Taalebizhaad. Direct Recovery of Motion and Shape in the General Case by Fixation. *IEEE Trans. on PAMI*, 14(8):847--853, Aug. 1992.

[244] D. Terzopoulos. Regularization of Inverse Visual Problems Involving Discontinuities. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-8(4):413--425, Apr. 1986.

[245] D. Terzopoulos and T.F. Rabie. Animat Vision: Active Vision in Artificial Animals. *VIDERE*, 1(1), 1997. Available at http://mitpress.mit.edu/e-journals/Videre/.

[246] W.B. Thompson and T.C. Pong. Detecting Moving Objects. *IJCV*, 4:39--57, 1990.

[247] W.B. Thomson and J.K. Kearney. Inexact Vision. In *IEEE Workshop on Motion: Representation and Analysis*, pages 791--794, 1986.

[248] T.Y. Tian, C. Tomasi, and D.J. Heeger. Comparison of Approaches to Egomotion Computation. In *Proceedings of CVPR'96*, pages 315--320, 1996.

[249] M. Tistarelli and G. Sandini. Direct Estimation of Time-to-Impact from Optical Flow. In *Proceedings of IUW*, pages 226--233, 1991.

[250] M. Tistarelli and G. Sandini. On the Advantages of Polar and Log-Polar Mapping for Direct Estimation of Time-to-Impact from Optical Flow. *IEEE Trans. on PAMI*, 15(4):401--410, Apr. 1993.

[251] C. Tomasi and T. Kanade. Shape and Motion from Image Streams under Orthography: a Factorization Method. *IJCV*, 9(2):137--154, 1992.

[252] P.H.S. Torr, P.A. Beardsley, and D.W. Murray. Robust Vision. In *Proceedings of BMVC'94*, pages 145--155, 1994.

[253] P.H.S. Torr and D. W. Murray. The Development and Comparison of Robust Methods to Estimate the Fundamental Matrix. *IJCV*, 24(3):271--300, 1997.

[254] 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.

[255] P.H.S. Torr and D.W. Murray. Stochastic Motion Clustering. In J.-O. Eklundh, editor, *Proceedings of ECCV'94*, pages 328--337, 1994.

[256] P.H.S. Torr and A. Zisserman. Robust Computation and Parametrization of Multiple View Relations. In *Proceedings of ICCV'98*, pages 727--732, Bombay, India, 1998.

[257] O. Tretiak and L. Pastor. Velocity Estimation from Image Sequences with Second Order Differential Operators. In *Proceeedings of ICPR'84*, pages 16--19, 1984.

[258] E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, 1998.

[259] R.Y. Tsai. A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-The-Shelf TV Cameras and Lenses. *IEEE Trans. of Robotics and Automation*, 3:323--344, 1987.

[260] R.Y. Tsai and T.S. Huang. Estimating 3-D Motion Parameters of a Rigid Planar Patch I. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-29(12):1147--1152, Dec. 1981.

[261] R.Y. Tsai and T.S. Huang. Estimating Three-Dimensional Motion Parameters of a Rigid Planar Patch, III: Finite Point Correspondences and the Three-View Problem. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-32:213--220, 1984.

[262] R.Y. Tsai and T.S. Huang. Uniqueness and Estimation of Three-Dimensional Motion Parameters of Rigid Objects with Curved Surfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-6(1):13--26, Jan. 1984.

[263] R.Y. Tsai, T.S. Huang, and W.-L. Zhu. Estimating Three-Dimensional Motion Parameters of a Rigid Planar Patch II: Singular Value Decomposition. *IEEE Trans. on ASSP*, 30(4):525--534, Aug. 1982.

[264] G. Tziritas. Recursive and/or Iterative Estimation of Two-Dimensional Velocity Field and Reconstruction of Three-Dimensional Motion. *Signal Processing*, 16:53--72, Jan. 1989.

[265] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, Cambridge, MA, 1979.

[266] S. Uras, F. Girosi, A. Verri, and V. Torre. A Computational Approach to Motion Perception. *Biological Cybernetics*, 60:79--87, 1989.

[267] A. Verri and T. Poggio. Motion Field and Optical Flow: Qualitative Properties. *IEEE Trans. on PAMI*, PAMI-11(5):490--498, May 1989.

[268] T. Viéville, O. Faugeras, and Q.T. Luong. Motion of Points and Lines in the Uncalibrated Case. *IJCV*, 17(1):7--41, Jan. 1996.

[269] T. Viéville and O.D. Faugeras. The First-Order Expansion of Motion Equations in the Uncalibrated Case. *CVIU*, 64(1):128--146, July 1996.

[270] T. Viéville, C. Zeller, and L. Robert. Using Collineations to Compute Motion and Structure in an Uncalibrated Image Sequence. *IJCV*, 20(3):213--242, 1996.

[271] Vision Chips, or Seeing Silicon web site: http://www.eleceng.adelaide.edu.au:80/groups/gaas/bugeye/visionchips/.

[272] J.Y.A. Wang and E.H. Adelson. Representing Moving Images with Layers. *IEEE Trans. on Image Processing*, 3(5):625--638, Sep. 1994.

[273] W. Wang and J. H. Duncan. Recovering the Three-Dimensional Motion and Structure of Multiple Moving Objects from Binocular Image Flows. *CVIU*, 63(3):430--440, May 1996.

[274] 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.

[275] A.M. Waxman and J.H. Duncan. Binocular Image Flows: Steps Toward Stereo-Motion Fusion. *IEEE Trans. on PAMI*, PAMI-8(6):715--729, Nov. 1986.

[276] 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.

[277] A.M. Waxman and K. Wohn. Contour Evolution, Neighborhood Deformation and Global Image Flow: Planar Surfaces in Motion. *International Journal of Robotics*, 4:95--108, 1985.

[278] J. Weber and J. Malik. Robust Computation of Optical-Flow in a Multiscale Differential Framework. *IJCV*, 14(1):67--81, Jan. 1995.

[279] J. Weber and J. Malik. Rigid-Body Segmentation and Shape-Description from Dense Optical-Flow Under Weak Perspective. *IEEE Trans. on PAMI*, 19(2):139--143, Feb. 1997.

[280] J. Weng, N. Ahuja, and T.S. Huang. Optimal Motion and Structure Estimation. *IEEE Trans. on PAMI*, 15(9):864--884, Sep. 1993.

[281] J. Weng, T.S. Huang, and N. Ahuja. Motion and Structure from Two Perspective Views:Algorithms, Error Analysis and Error Estimation. *IEEE Trans. on PAMI*, 11(5):451--476, May 1989.

[282] S.W. Wilson. The Animat Path to AI. In J.-A. Meyer and S. Wilson, editors, *From Animals to Animats*, pages 15--21. MIT Press, Cambridge, MA, 1991.

[283] S. Wolfram. *Mathematica: A System for Doing Mathematics by Computer.* Addison-Wesley Publishing Company, 1988.

[284] R.Y. Wong and E.L. Hall. Sequential Hierarchical Scene Matching. *IEEE Trans. on Computers*, 27:359--366, 1978.

[285] M. Xie. Robotic Hand-Eye Coordination: New Solutions with Uncalibrated Stereo Cameras. *MVA*, 10(3):136--143, 1997.

[286] G.-S. Young, T.-H. Hong, M. Herman, and J.C.S. Yang. Safe Navigation for Autonomous Vehicles: A Purposive and Direct Solution. In *Proceedings of SPIE Int. Conf. on Intelligent Robots and Computer Vision XII*, pages 31--42, 1993.

[287] G.S. Young and R. Chellappa. 3-D Motion Estimation Using a Sequence of Noisy Stereo Images: Models, Estimation, and Uniqueness Results. *IEEE Trans. on PAMI*, 12(8):735--759, Aug. 1990.

[288] S. Zeki. The Visual Image in Mind and Brain. *Scientific American*, 267(3):43--50, Sep. 1992.

[289] C. Zeller and O. Faugeras. Application of Non-Metric Vision to Some Visual Guided Tasks. Technical Report RR-2308, INRIA Sophia Antipolis, Jul. 1994.

[290] Z. Zhang. A New and Efficient Iterative Approach to Image Matching. In *Proceedings of ICPR'94*, pages 563--565, Jerusalem, Israel, 1994.

[291] Z. Zhang. Determining the Epipolar Geometry and its Uncertainty: A Review. Technical Report RR-2927, INRIA Sophia Antipolis, Jul. 1996.

[292] Z. Zhang. Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting. *Image and Vision Computing*, 15(1):59--76, Jan. 1997.

[293] Z. Zhang. Image-Based Geometrically-Correct Photorealistic Scene/Object Modeling (IBPhM): A Review. In *Proceedings of ACCV'98, LNCS No. 1352*, volume 2, pages 340--349, Hong Kong, China, Jan. 1998.

[294] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry. *AI Journal*, 78:87--119, Oct. 1995.

[295] Z. Zhang, R. Weiss, and A.R. Hanson. Obstacle Detection Based on Qualitative and Quantitative 3D Reconstruction. *IEEE Trans. on PAMI*, PAMI-19(1):15--26, Jan. 1997.

[296] Z.Y. Zhang, Q.T. Luong, and O.D. Faugeras. Motion of an Uncalibrated Stereo Rig: Self-Calibration and Metric Reconstruction. *IEEE Trans. on RA*, 12(1):103--113, Feb. 1996.

[297] X. Zhuang, T. Wang, and P. Zhang. A Highly Robust Estimator through Partially Likelihood Function Modelling and its Application in Computer Vision. *IEEE Trans. on PAMI*, 14:19--35, 1992.

[298] I. Zoghlami, O. Faugeras, and R. Deriche. Using Geometric Corners to Build a 2D Mosaic from a Set of Images. In *Proceedings of CVPR'97*, pages 420--425, Puerto Rico, Jun. 1997.