FEATURE-BASED OBJECT TRACKING IN MARITIME SCENES

PETR VOLEŠ

A thesis submitted in partial fulfilment of the requirements of Bournemouth University for the degree of Doctor of Philosophy

February 2005

Bournemouth University

Feature-Based Object Tracking in Maritime Scenes

Abstract

Feature-Based Object Tracking in Maritime Scenes

Petr Voleš

Bournemouth University

A monitoring of presence, location and activity of various objects on the sea is essential for maritime navigation and collision avoidance. Mariners normally rely on two complementary methods of the monitoring: radar and satellite-based aids and human observation.

Though radar aids are relatively accurate at long distances, their capability of detecting small, unmanned or non-metallic craft that generally do not reflect radar waves sufficiently enough, is limited. The mariners, therefore, rely in such cases on visual observations.

The visual observation is often facilitated by using cameras overlooking the sea that can also provide intensified or infra-red images. These systems nevertheless merely enhance the image and the burden of the tedious and error-prone monitoring task still rests with the operator.

This thesis addresses the drawbacks of both methods by presenting a framework consisting of a set of machine vision algorithms that facilitate the monitoring tasks in maritime environment.

The framework detects and tracks objects in a sequence of images captured by a camera mounted either on a board of a vessel or on a static platform overlooking the sea. The detection of objects is independent of their appearance and conditions such as weather and time of the day. The output of the framework consists of locations and motions of all detected objects with respect to a fixed point in the scene. All values are estimated in real-world units, i.e. location is expressed in metres and velocity in knots. The consistency of the estimates is maintained by compensating for spurious effects such as vibration of the camera.

In addition, the framework continuously checks for predefined events such

as collision threats or area intrusions, raising an alarm when any such event occurs.

The development and evaluation of the framework is based on sequences captured under conditions corresponding to a designated application. The independence of the detection and tracking on the appearance of the scene and objects is confirmed by a final cross-validation of the framework on previously unused sequences.

Potential applications of the framework in various areas of maritime environment including navigation, security, surveillance and others are outlined. Limitations to the presented framework are identified and possible solutions suggested. The thesis concludes with suggestions to further directions of the research presented.

To Magda, Julie

and my parents,

for their limitless patience and support.

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

This thesis would have not existed without the efforts of many kind individuals. It gives me a great pleasure to acknowledge their contribution here. I would like to thank my first supervisor, Martin Teal, for his advice and encouragement. My special gratitude goes to Milan Vasilko who's enormous help, constructive criticism and support were invaluable during the whole project. Special thanks to Jim Roach and the rest of the management team of the School of Design, engineering and Computing who have provided generous support throughout the studies. I am grateful to my past and present colleagues, Andy Smith, Periklis Chatzimisios, Piotr Stepien, Pi Huang, David Mayer, Sylvaine Laing, Jakub Šena, Darrell Gibson, Jenny Longster, Carolyn Mair and many others for their friendship, help and excellent working environment they have provided over the years. Financial support provided during my PhD studies by the UK CVCP Overseas Research Award Scheme is greatly acknowledged. My work on this research would have never been possible without the support from my family. I am grateful to my parents for their limitless support and encouragement in my pursuit of the university studies. Above all, I am indebted to my wife Magdalena who has firmly supported me throughout my PhD studies and has patiently suffered all the consequences.

List of Abbreviations

AIS	Automatic Identification System
ARPA	Automatic Radar Plotting Aids
ATR	Automatic Target Recognition
CCTV	Closed Circuit Television
COLREGS	The International Regulations for Preventing Collisions at Sea
CRF	Corner Response Function
EMD	Earth Mover's Distance
GLD	Generalised Logistic Distribution
GPC	Ground Plane Constraint
GPS	Global Positioning System
HSC	High Speed Craft
ICA	Independent Components Analysis
IMO	International Maritime Organization
MAIB	Maritime Accident Investigation Branch
MTC	Maritime Transport Committee
NAVREGS	The International Regulations for Navigation
OECD	Organization for Economic Co-operation
PCA	Principal Components Analysis
RCS	Radar Cross Section

SSD	Sum of Squared Differences
SWH	Significant Wave Height
TTC	Time To Contact
VHF	Very High Frequency
VTS	Vessel Tracking System

Chapter 1

Introduction

As the maritime transport sector provides the transportation of goods and persons around the world on a massive scale, navigation safety and security must be paramount for all involved. Any disrupting events, either deliberate criminal acts such as piracy or accidental such as collisions, need to be reduced or avoided completely.

Visual and radar-based navigation together with sea navigation and collision avoidance regulations (NAVREGS, COLREGS), (International Maritime Organization, 2004) are established methods in maritime transport sector. Additional navigation aids such as Global Positioning System (GPS), very high frequency radio link communication (VHF) and recently Automatic Identification System (AIS) complement them. These additional technologies are, however, applicable only to adequately equipped craft which are willing or able to participate in the process of navigation.

Both visual and radar-based navigations suffer from various limitations with potentially devastating consequences. Nielsen and Petersen (2001) illustrate on real situations that even the combination of various navigation aids does not always guarantee safe navigation.

Following sections provide a detailed discussion of the limitations to established navigation methods. Three main areas of maritime transport where safety and security are essential are analysed in detail: collision avoidance, maritime piracy counter measures and Vessel Tracking Systems. Benefits of machine vision technology in each of these areas are suggested. The proposed machine vision framework is briefly introduced. The chapter concludes with an outline of the structure of this thesis.

1.1 Limitations to Radar-based Systems

Marine radars can detect and locate objects on the sea up to tens or hundreds of nautical miles with relatively high precision. Majority of marine craft ranging from small leisure yachts to massive cargo ships are equipped with Automatic Radar Plotting Aids (ARPA) that combine information obtained by radar with electronic charts, GPS data and AIS.

Despite continuous advances in the radar technology, a number of limitations can still be identified:

- Radar-based systems do not operate beyond a certain minimum range. For example, Furuno (2004) limit the minimum range of their products to 1/8 of nautical mile. This makes use of radar in confined areas such as harbour entrances unreliable.

- The strength of radar response for a particular object expressed as Radar Cross-Ratio (RCS) depends on multiple factors such as material from which the object is built, geometry and pose of the object, speed, etc. Weather phenomena such as rain, waves, etc. generate false responses that can be mistaken for genuine objects, (Kingsley and Quegan, 1992).

- Materials such as fiberglass are transparent to radar signal. Small and low objects can remain undetected as their RCS is insufficient. Smooth shaped objects such as hulls of fishing boats give a poor radar response as compared to rough shaped objects, (Australian Transport Safety Bureau, 2004).

- Use of radar requires experienced and knowledgeable operators who are able to associate information on the radar display with the underlying situation in the scene.

- Radar is a sophisticated piece of equipment. There are numerous options and settings that have to be correctly adjusted for radar to function properly. Radar antenna contains moving parts that are prone to mechanical failures especially in severe weather conditions that are not uncommon in maritime environment.

- There is a health concern associated with a long-term exposure to high frequency electromagnetic fields typically generated by radars, (World Health Organization, 1999).

The International Maritime Organization (IMO) and local authorities (Australian Transport Safety Bureau, 2004) appeal on mariners not to rely solely on ARPAs but to jointly use all available navigation aids with primary emphasis on vigilant watch-keeping. There are, however, serious limitations to human vigilance.

1.2 Limitations to Human Vigilance

Historically, the very first vigilance test in maritime domain was done by Mackworth (1950) at the request of the Royal Navy. This was concerned with the degradation of sonar operators detecting enemy submarines. The results showed that the vigilance could not be maintained at an optimum level for more than 30 minutes. After 30 minutes as many as 15% of omissions occurred.

A more recent study into factors influencing the vigilance of humans was presented by The Institute of Applied Anthropology (2001).

Although the main focus of this study was to draw up guidelines for lifeguards at swimming pools, the results can be extended to other similar activities requiring significant attention over an extended period of time, including maritime watch-keeping.

The study by The Institute of Applied Anthropology (2001) identifies three main factors influencing the vigilance:

- characteristics of the task
- physical surroundings
- temporal progress of the task

Characteristics of the task - the level of vigilance is proportional to the ratio of relevant over irrelevant information provided to the operator as proved by Hitchcock et al. (2003).

The physical surroundings - environmental factors such as noise and high temperature have an adverse effect on the level of vigilance.

Temporal progress - the study by The Institute of Applied Anthropology (2001) shows results from previous experiments that confirm that short term breaks have a positive influence on the level of vigilance. The influence of the time of the day was also confirmed; during low points in physiological activity (early morning, early afternoon) breaks should be longer than at other times.

The study supports the general perception that any vigilance task can only be maintained at the highest level for about a half an hour. The findings are confirmed in a study commissioned by Marine Accident Investigation Branch (2004).

The study shows that crew's fatigue is the major cause of naval accidents. Broken sleeping patterns, extended working hours and limited crew numbers all reduce the vigilance substantially. Many collisions occurred when watchmen on duty fell asleep or due to fatigue omitted various indications of developing emergency. An automated early warning system would certainly prevent many of these incidents.

1.3 Collision Avoidance

Similar to the Highway Code the traffic on the sea is governed by The International Regulations for Preventing Collisions at Sea simply known as COLREGS, (International Maritime Organization, 2004). A detailed knowledge of COLREGS is compulsory for any professional mariner involved in navigation. Despite that, the collisions still occur.

A report by Maritime Accident Investigation Branch (2002) shows that 6% of all 5138 marine accidents involving fishing vessels reported to Maritime Accident Investigation Branch between 1992 and 2000 account for collisions. Despite their relatively low occurrence, collisions on the sea usually have more devastating consequences, especially at high seas where a rescue is not always to hand. Another report by Marine Accident Investigation Branch (2004) shows that most collisions are due to crew's fatigue, breach and misinterpretation of COLREGS and misinterpretation of radar and charting data.

The importance of accurate correspondences between objects surveyed visually and by radar is illustrated by Nielsen and Petersen (2001) on a practical example of navigation of a large cargo vessel. The study shows that because the VHF radio link between vessels often cannot be established due to the absence of identification signatures intentions of the craft have to be assessed by visual and radar monitoring. Incorrect correspondence can lead to incorrect assessment of the intentions of the objects. This can consequently lead to an inadequate collision avoidance manoeuvre with disastrous consequences.

Kjerstad (2003) presents results of an extensive survey involving navigators of High Speed Craft (HSC) operating along the coast of Norway. The survey

shows that radar and electronic charting systems are considered the most important navigation aids. Nevertheless, more than half of the navigators participating in the study admit that the radar is unreliable in bad weather and high seas. More importantly, 90% of surveyed navigators consider night vision (Turn Ltd., 2001; Vector Developments Ltd., 2004; Vistar Night Vision Limited, 2004a; Vistar Night Vision Limited, 2004b; Vistar Night Vision Limited, 2004c; The Current Sales Corp., 2004) as a significant contribution to the safety in spite the fact that only 4% of craft are equipped with such a technology. In addition, 72% of navigators would not object to more navigational technology on the bridge.

However, mere provision of the enhanced image on the bridge still requires full-time attention of a designated operator. An automated highlighting of objects in a night vision image would relieve the operator of the constant pressure and help him/her to concentrate on other navigational duties.

An important issue recently raised by United States National Oceanic and Atmospheric Administration are collisions of HSC with large marine mammals such as whales and dolphins, (Jensen and Silber, 2003). The photographic evidence gathered at Mediterranean during the last decade published by Tethys Research Institute (2004) illustrates the scale of injuries these creatures sustain during encounters with HSC. Fatalities are not exceptional. Despite the fact that marine mammals are difficult to spot when near the sea surface there is a potential for improvement and automated whale detection based on a machine vision would potentially help to avoid harming and killing of these creatures.

1.4 Maritime Piracy Counter Measures

Threats to the maritime traffic infrastructure can be numerous due to many security weak points along a cargo route. One of the most damaging threats is the act of piracy and armed robbery during transportation by sea, (Hawkes, 2001; White and Wydajewski, 2002; International Maritime Organization, 2002). The statistics presented by Maritime Transport Comitee (2003) show that there were a total of 335 officially registered acts of piracy in 2001. Hawkes (2001) and White and Wydajewski (2002) indicate that the real number could actually be much higher as many incidents are not reported to the authorities. The data maintained by the International Maritime Organization

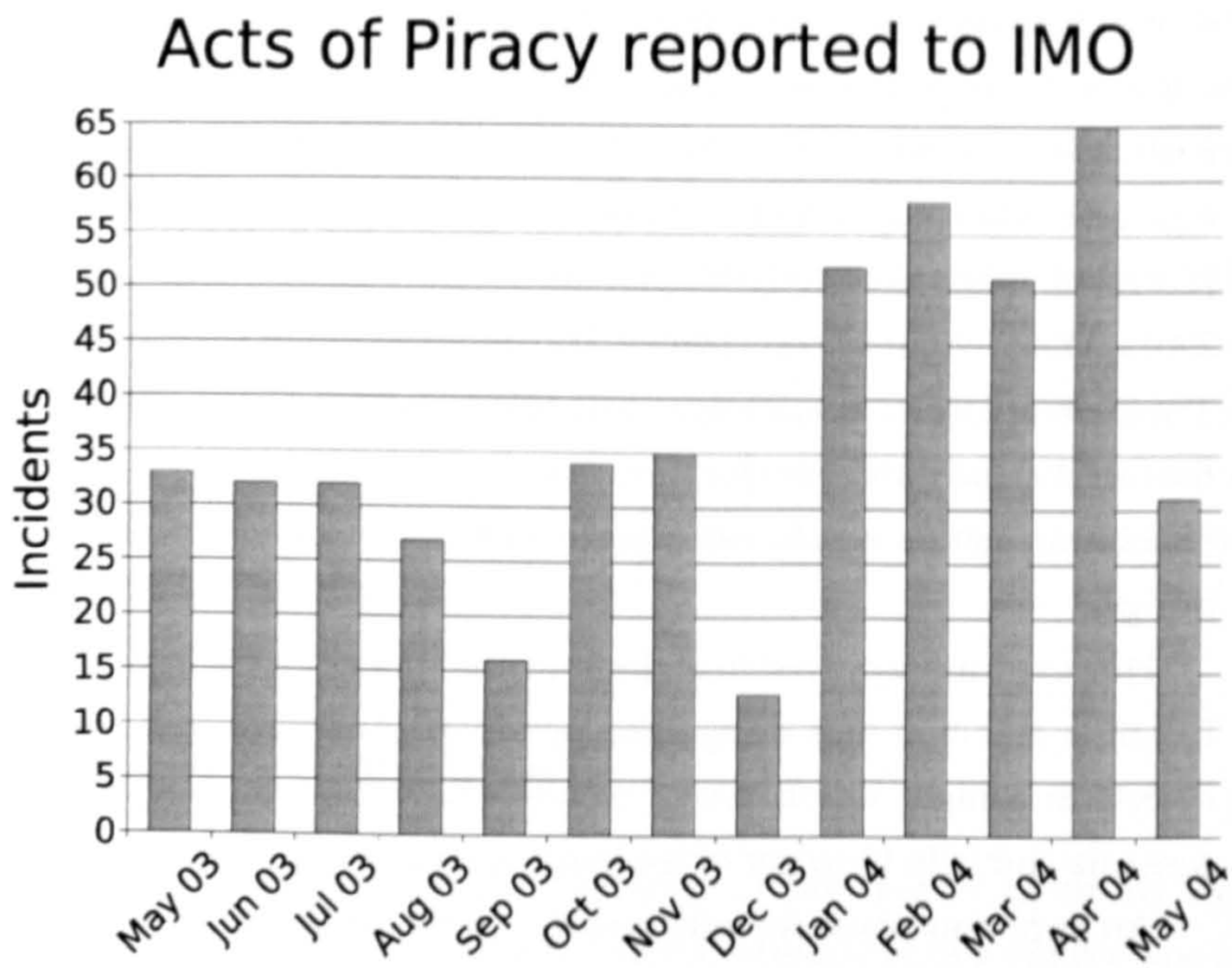


Figure 1.1: Acts of piracy reported to International Maritime Organisation between May 2003 and May 2004, (International Maritime Organization, n.d.)

(n.d.) indicate that the number of incidents is increasing (see Figure 1.1) as piracy is a very lucrative form of crime.

In some cases the capture of a vessel can lead to longer term profits by the operation of a 'ghost' ship, (Maritime Transport Comitee, 2003).

The main hot-spots of high-seas crime are the East Asian regions, Red Sea, African corner, South America and Central Africa where economic piracy goes hand in hand with the political situation (see Figure 1.2).

According to Maritime Transport Comitee (2003) 85% of attacks are committed either underway or at anchor which means that the typical scenario of such an attack involves a fleet of small craft approaching the vessel in an attempt to board it and overpower the crew. Due to the inability of radar to detect small non-metallic craft approaching a ship the crew must rely on careful watch-keeping.

A comprehensive guidance for minimisation and avoidance of piracy attacks on vessels published by International Maritime Organization (2002) considers watch-keeping as one of the main piracy counter measures. Careful watch-keeping helps to detect suspicious activities on the sea well in advance

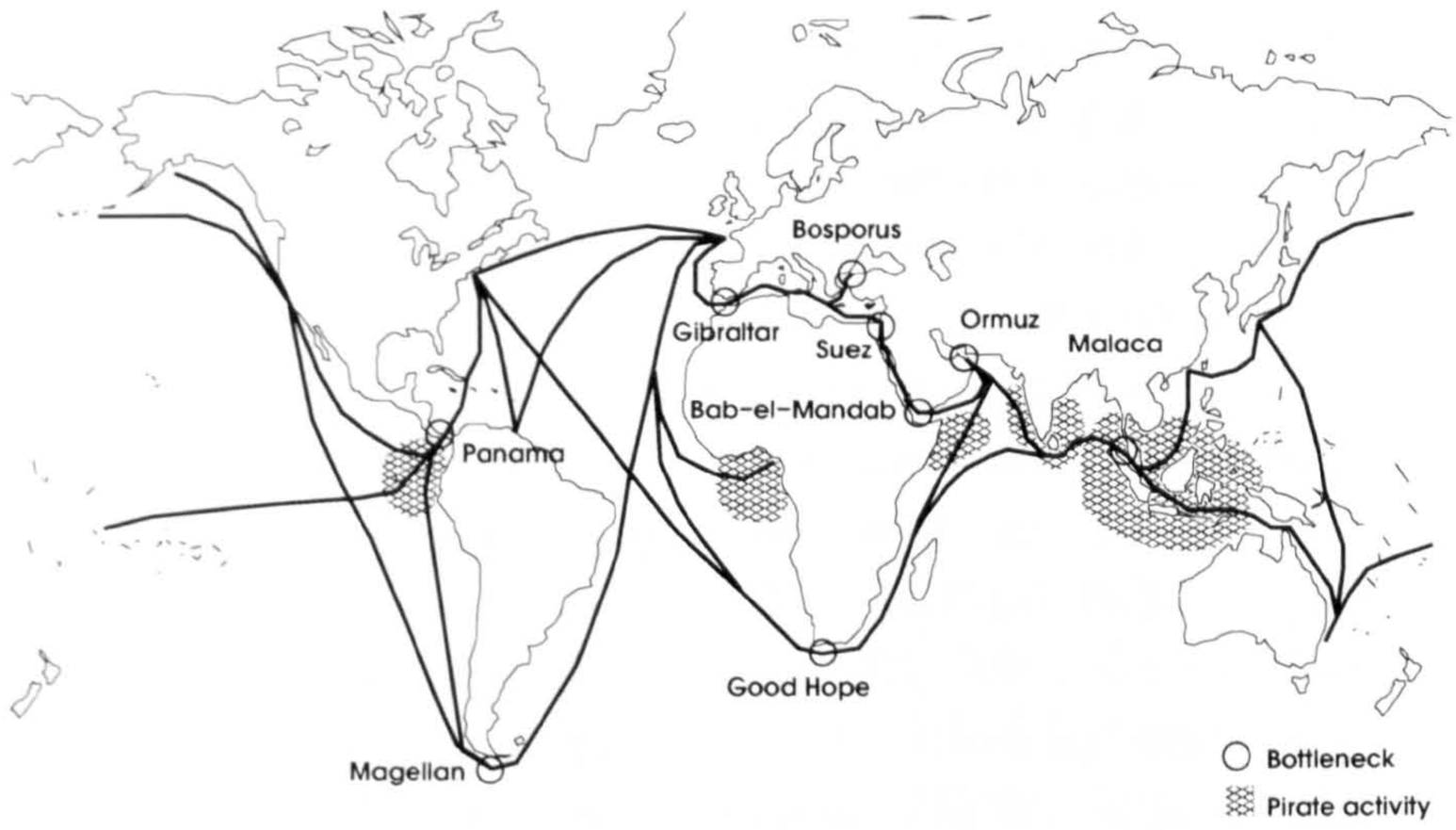


Figure 1.2: Map of global sea routes and the piracy hot-spots, (Maritime Transport Comitee, 2003).

allowing the crew to prepare for the attack. Nevertheless, proper watch-keeping is often difficult to maintain due to a limited number of personnel on modern cargo ships.

The conclusion is that maritime piracy is a serious issue which cannot be easily resolved by existing technologies such as radars. The vessel crews rely mainly on careful watch-keeping and increased vigilance when it comes to piracy threat assessment. Automation of watch-keeping task provided by a machine vision based system would relieve the human operator from necessary continuous attention. The operator would be alerted only when a potential risk is detected by the system.

1.5 Vessel Tracking Systems

Most harbours, port facilities and busy waterways are monitored and controlled by Vessel Traffic Systems (VTS), (Amiel, 2000; Slater, 1989). The VTS seamlessly integrate navigation, vessel tracking, surveillance, cargo and ship registration, administration and operational logistics in maritime environment. Systems typically consist of modules that are interconnected on a cooperative basis. This enables all the necessary information to be broadcast to various harbour authorities. For example, customs, traffic control, health and safety departments and other authorised bodies all have access to any information

required, such as traffic maps, ship identifications, cargo registrations, weather reports and forecasts, all through a single integrated system. The VTS also support communication links between control centres and vessels. The Automatic Identification System (AIS) is recently becoming substantial data source in the VTS.

The navigation and tracking modules are essential parts of any VTS. For open seas and coastal areas the service is provided by radars in combination with GPS, (Phinney, 1998). For traffic hot-spots such as harbour entrances, busy transport lines and crossings additional navigation aids such as buoys and VHF radio link communications are used. Locations where radar is inapplicable are monitored by CCTV cameras linked to the central control room in order to assist in navigation, (Amiel, 2000). The cameras provide only an overview of such locations and no further automated analysis based on the visual information is performed. The operator must survey the scene, assess activity of craft, make appropriate decisions and take actions.

The process would benefit from an automation of at least the first two stages alerting the operator only when decision and action are required. For example, the operator could be automatically warned by the system that "ship A is entering a forbidden zone" or "vessels C and D are on collision course". The system would be seamlessly integrated into the VTS structure, filling gaps in radar-based vessel tracking.

1.6 Proposed Framework

This thesis addresses the limitations of visual and radar-based navigations by proposing a machine vision-based framework that overcomes these limitations. The framework automates the visual surveillance task and provides human operator with relevant information about the activity of objects in the maritime scene. It identifies various activities requiring operator's attention such as collision threat, piracy threat, intrusion of forbidden zones and others. The framework facilitates the visual maritime navigation and complements other navigation technologies such as radar and GPS.

Pertinence of the proposed framework is illustrated in three most prominent areas of maritime transportation sector: collision avoidance, maritime piracy threat assessment and Vessel Tracking System. The use of the framework is not limited to these areas and there are certainly other optional

applications of the framework, either stand-alone such as perimeter intrusion detection in small private moorings, or embedded in more complex systems such as Automatic Radar Plotting Aid (ARPA) or Vessel Tracking System (VTS).

1.7 Thesis Outline

This chapter identifies limitations to established technologies of maritime navigation and proposes machine vision-based framework addressing these limitations together with various applications of the framework.

Chapter 2 characterises and analyses the problem domain, outlines the research objectives and describes research methodology used in the thesis.

Chapter 3 reviews the previous work in relevant areas of machine vision applications in maritime sector and object tracking.

Chapter 4 presents an adaptive texture-based segmentation algorithm for separation of objects from the background in the maritime scene.

Chapter 5 introduces an object representation consisting of a weak perspective projection of salient geometric features and submersion line.

Chapter 6 describes a temporal correspondence matching of geometric features necessary for motion estimation.

Chapter 7 presents the motion estimation by Kalman tracking as well as image stabilisation technique based on a registration of the horizon image projection.

Chapter 8 details the inverse geometrical mapping of objects' location and motion estimates, threat assessments and analysis of the mapping precision.

Chapter 9 presents the results of the cross-validation of the complete framework.

Chapter 10 concludes the thesis by discussion of the results and suggestions on future directions.

Chapter 2

Problem Characterisation

2.1 Introduction

This chapter provides several contexts of maritime scenes necessary to obtain characterisation of the problem domain. As the proposed framework is based on processing of visual information, contexts representing the appearance, geometry and dynamics of the scene and objects are analysed, (Strat, 1993):

- *optical context* - appearance attributes such as shape, colour, scale and structure of various components of a maritime scene including background and objects,
- *geometric context* - the geometric model of the scene, position of the camera with respect to scene and location of objects within the scene,
- *temporal context* - time-dependent properties of the scenes, such as motion dynamics of the sea and objects.

The contextual analysis carried out in this chapter leads to the research objectives together with constraints and assumptions imposed on the problem domain being outlined. Research methodology is introduced including the architecture of the proposed framework and details of development sequences.

The contextual analysis is based on several principal assumptions:

- The vision based framework is assumed to be composed of an input represented by one or more independent or related cameras overlooking

the scene; a processing unit that performs a dedicated task on the data provided by input; an output that is an outcome of the processing.

- An image is generated by projection of points in the scene through a camera lens onto a planar surface called the image plane.
- The video sequence is a sequence of images acquired at equal periods of time.

2.2 Optical Context

2.2.1 Background

The illumination of an outdoor scene depends mainly on environmental conditions, time of the day and the structure of the scene itself, (Narasimhan et al., 2002). Outdoor illumination obeys multiple models - light can be diffuse (cloudy day) or directional (sunny day). The direction of light depends on time of day and season of the year. Although these two models are distinctive, it is their combination that provides an approximation to a real situation. The parameters of a daylight model depend on factors such as humidity and air temperature, wind conditions, dispersion of minute particles, etc. that directly affect the way light passes through the air. The outdoor illumination model is a non-linear function of multiple variables quantifying these factors, (Preetham et al., 1999).

The background of maritime scenes is composed of the water, land and sky. The assumption is that land and sky are located above the water and that they are clearly separated by a horizon. If land is not present in the scene the horizon separates water directly from the sky. The focus of the proposed framework is only on the region of the water below the horizon.

Optical properties of water surfaces in outdoor scenes are influenced by weather factors such as wind, temperature, atmospheric pressure, dispersion of particles (turbidity), etc. Specular reflectance (Jain et al., 1995) of the water depends on an incident angle and it varies from between 5% to 100%, (Premoze and Ashikhmin, 2000). For reflectance close to 100% the sky is reflected with almost no loss. For reflectance close to 5% the light mostly comes from below the surface. This causes the familiar pattern of localised bright and dark spots.

While the water is predominantly specular or transparent, maritime objects

are typically composed of parts with mostly Lambertian reflectance. This is due to the physical properties of materials from which the objects are composed. Exceptions are shiny elements such as window panes, chrome railings, etc. which are mainly specular and reflect the incident light with little scatter.

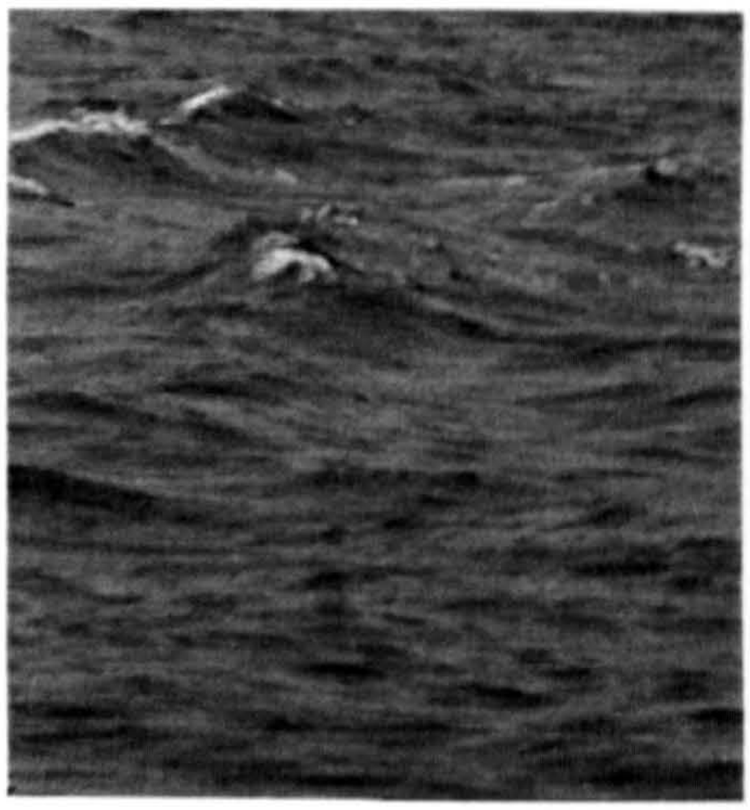
Water surface that undergoes perpetual motion which appears as waves, projects onto the image as a pattern of fragments of similar size and different intensity. The pattern of the water surface does not remain static, it changes continuously, (Doretto et al., 2003) (see Figures 2.1a,b). The fragments change intensity with changing illumination and incidence angle of the light. The position of the fragments changes as the wave oscillations propagate through the water surface. Wakes and crests that appear much brighter than the rest of the water are common during high wind speeds.

Regular patterns often form on the water surface. These patterns are caused by factors such as disturbances caused by some event such as boat passing by or underwater streams near the surface. They usually appear as straight lines or curves of varying width and intensity that differs from intensity of the surrounding water (see Figure 2.1c). The patterns usually change position over time (e.g. wakes travel away from the source of disturbance). They typically blend gradually into the background and disappear over time.

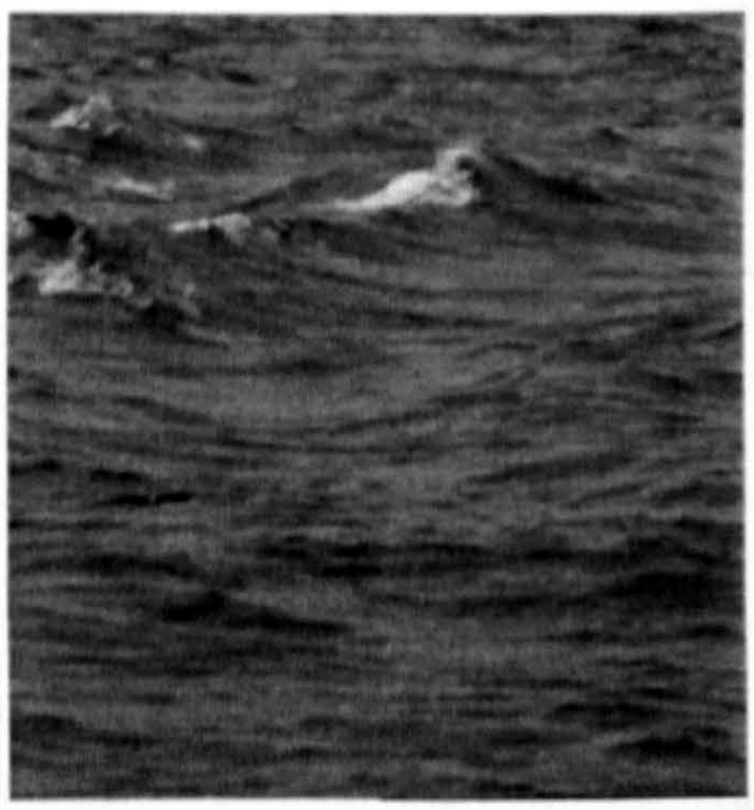
Some of these patterns can be permanent or change very slowly. Permanent patterns are usually caused by an interaction of water with sea vegetation, corals, rocks or other either natural or artificial objects that are close to the surface. Slowly changing patterns also appear on boundary of two underwater streams with different directions or speeds. Regions of different colours appear when the water is shadowed by clouds or at river deltas where fresh water enters the sea or where the depth of the sea suddenly changes.

2.2.2 Objects

Categorisation of objects in maritime scenes with respect to appearance is complicated due to variations in types, shapes, scales and colours. Figure 2.1d shows a minimal example of objects that can be encountered in a typical maritime scene. The scale of objects ranges from less than a metre for a buoy to hundreds of metres for a cargo ship. The appearance of the same object can change significantly within a short period of time. For example, yacht takes down its' sail, cargo ship unloads, etc.

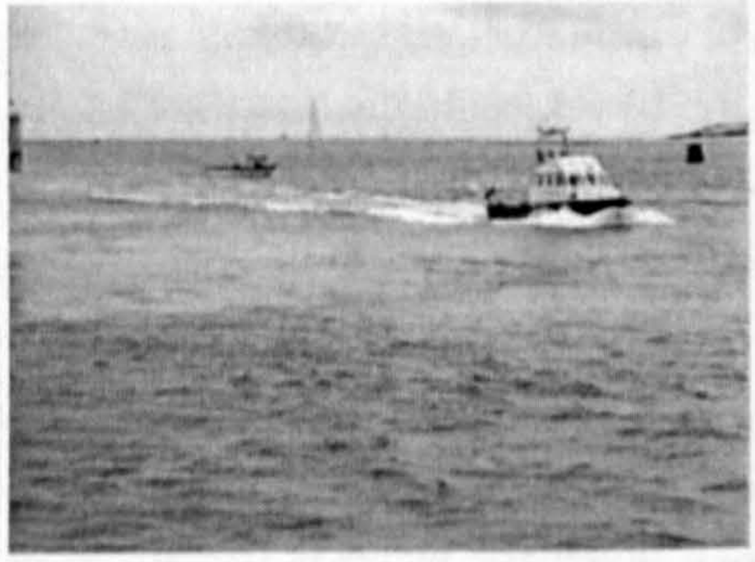


(a) initial state of the sea surface patch



(b) same patch after one second

(c) The wake generated by passing high speed craft



(d) various maritime objects

Figure 2.1: Maritime environment. (a) and (b) illustrate the spatio-temporal variability of the sea surface. Both frames are only approximately a second apart. (c) shows a typical 'wake' pattern occurring after high speed craft passes by. (d) illustrates maritime objects with various appearances and motion dynamics encountered in a maritime environment. The object on the far left is a part of a fixed structure embedded into a sea floor, vessels in the middle are both moving at different speeds and buoy on the right is either static or rolling due to the sea motion depending on the weather conditions.

A maritime scene might also contain objects that serve the purpose of navigation such as buoys, waterway signs, beacons, lighthouses, etc. Other objects such as piers, moorings, docks, etc. are used in maintenance of the craft and the cargo. All these objects come in various shapes, colours and scales which makes their categorisation based on visual attributes difficult.

In populated areas the sea surface can be littered with floating debris. The debris is mostly composed of waste coming either from passing vessels or it is washed off the land. The debris ranges from small and light household waste to logs of wood or cargo containers.

Natural objects such as rocks, cliffs and other geological phenomena protruding above the water are integral components of a maritime scene as well. Large sea mammals such as whales and dolphins surface regularly and in such a case they can pose a navigation challenge. They usually appear as dark and shiny bows above the water surface with fins occasionally protruding.

Smaller animals such as sea birds and fish are part of the maritime scene as well. They, however, hardly present a challenge to maritime traffic due to their natural tendency to avoid unwanted encounters with maritime craft.

2.3 Geometric Context

2.3.1 Scene Projection Model

The model of projection of the maritime scene onto an image is derived from a general projection from 3D scene to 2D image plane through a pinhole camera (Shapiro, 1995; Jain et al., 1995) as illustrated in Figure 2.2. The point in the scene at location $[X, Y, Z]$ is projected at position $[x, y]$ in the image through the centre of projection O. The focal length of the camera is f and the principal point is located at position $[0, 0]$ in the image. The image plane is a discrete array of pixels of a finite size s_{pix}. The pixels are assumed square which is acceptable for the majority of real cameras used in machine vision such as those by Hitachi Denshi (n.d.). It is often convenient to express the focal length in pixels, $f_{pix} = \frac{f}{s_{pix}}$.

The objects in maritime scenes are assumed to float on the planar sea surface. The assumption is known as Ground Plane Constraint (GPC) (Worral et al., 1995; Worrall et al., 1994). It stipulates that all objects are located on a plane with a constant Y-coordinate. The GPC is essential to many inland

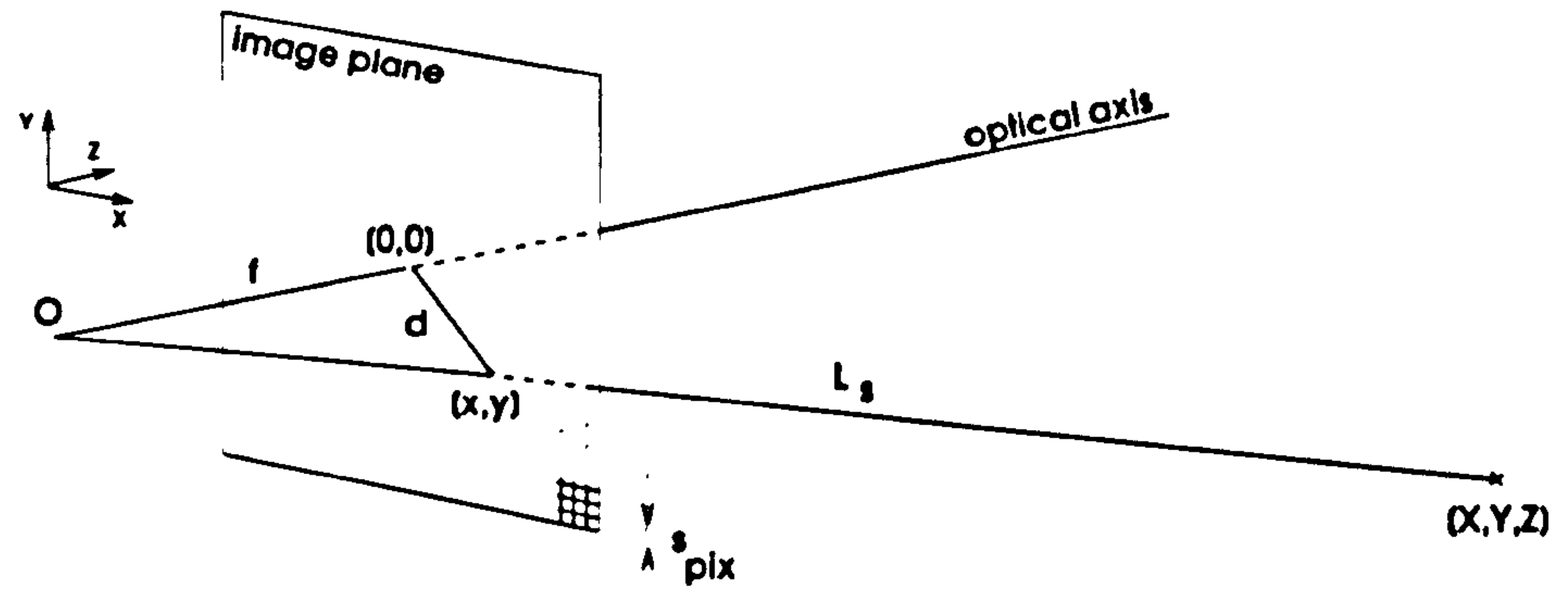


Figure 2.2: General projection from 3D scene to 2D image plane through a pinhole camera with projection centre O and focal length f. L_s is the line of sight connecting the scene point at $[X, Y, Z]$ with projection centre. The scene point projects at position $[x, y]$ on the image plane. d is the line connecting the projection with the principal point $[0, 0]$. The pixels are assumed square with side size s_{pix}.

traffic surveillance applications where cars and people are located on a road or a ground, (Dellaert and Thorpe, 1997; Magee, 2004; Williamson, 1998; Tai et al., 2004).

The model corresponds to a setup where the camera is mounted either on an articulated static point or on a mast of a moving vessel and it overlooks the monitored area. The model geometry is outlined in Figure 2.3.

The imaging device is approximated by the pinhole camera. The projection centre of the camera is located at height H above the plane Π corresponding to the sea surface. The camera is tilted by an angle ω so that a section of plane Π is projected onto the image plane. The projection can be expressed in terms of a plane-to-plane projection (Mohr and Triggs, 1996). No rotation about the optical axis is assumed so that horizontal edges of the image plane are aligned with the plane Π. The rotation about the Y-axis does not influence the projection as Π hypothetically stretches to infinity in all directions. Rotation around the Y-axis would imply that a different section of the plane Π is projected onto the image.

Optimally, the horizon in the scene projects at the top edge of the image, so that the largest possible section of the plane Π is projected onto the image. This allows for full utilisation of the image plane area.

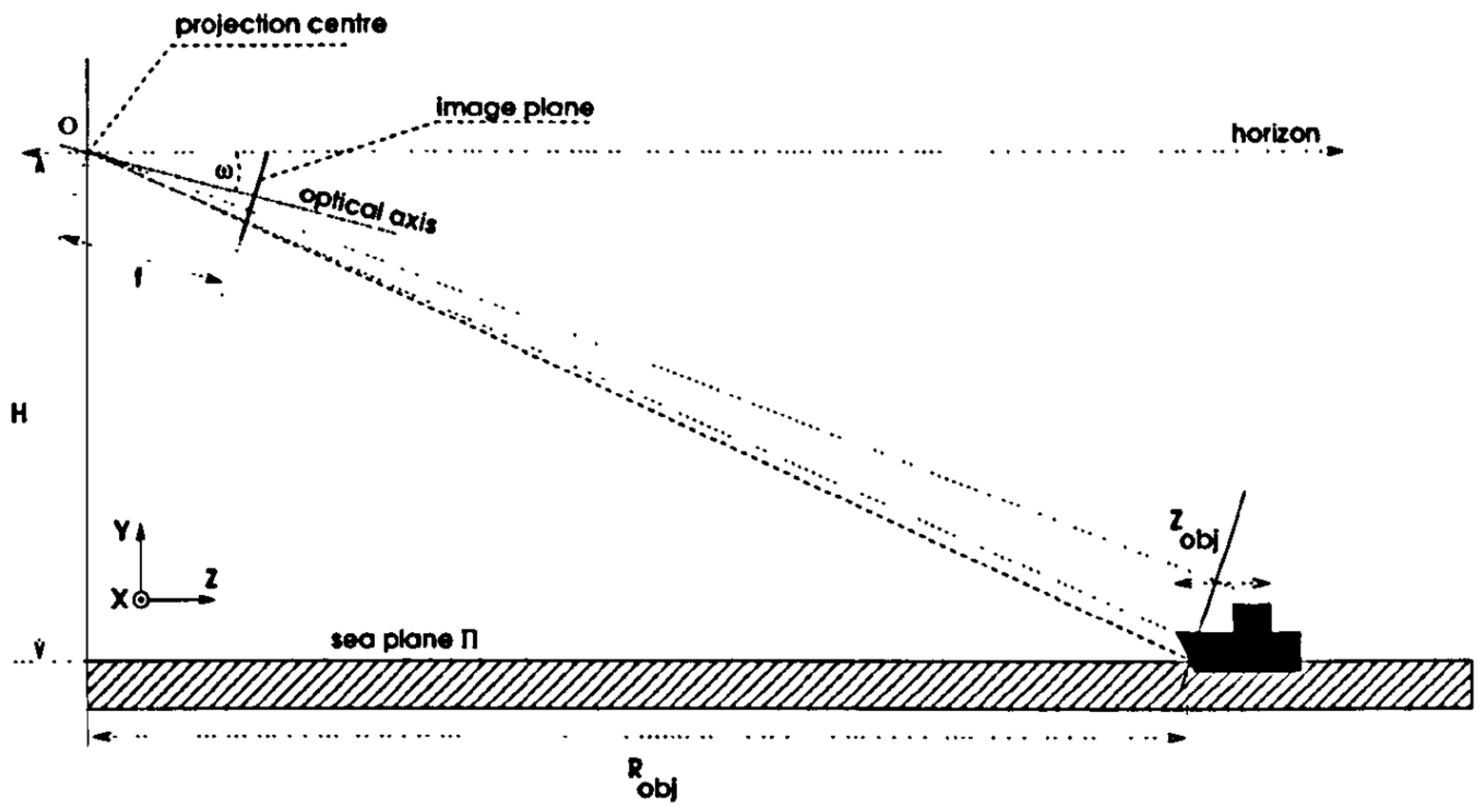


Figure 2.3: The projection model of the maritime scene. Objects are located on the plane Π representing the sea surface. The plane is projected through a pinhole camera onto the image plane. The camera is positioned at height H above the sea surface. It is tilted at an angle ω. The horizon is assumed at infinity and it projects onto the image as a horizontal line. The object is located at range R_{obj} in front of the camera and it's depth is Z_{obj}.

2.3.2 Deviations from Scene Projection Model

2.3.2.1 Curvature of the Earth

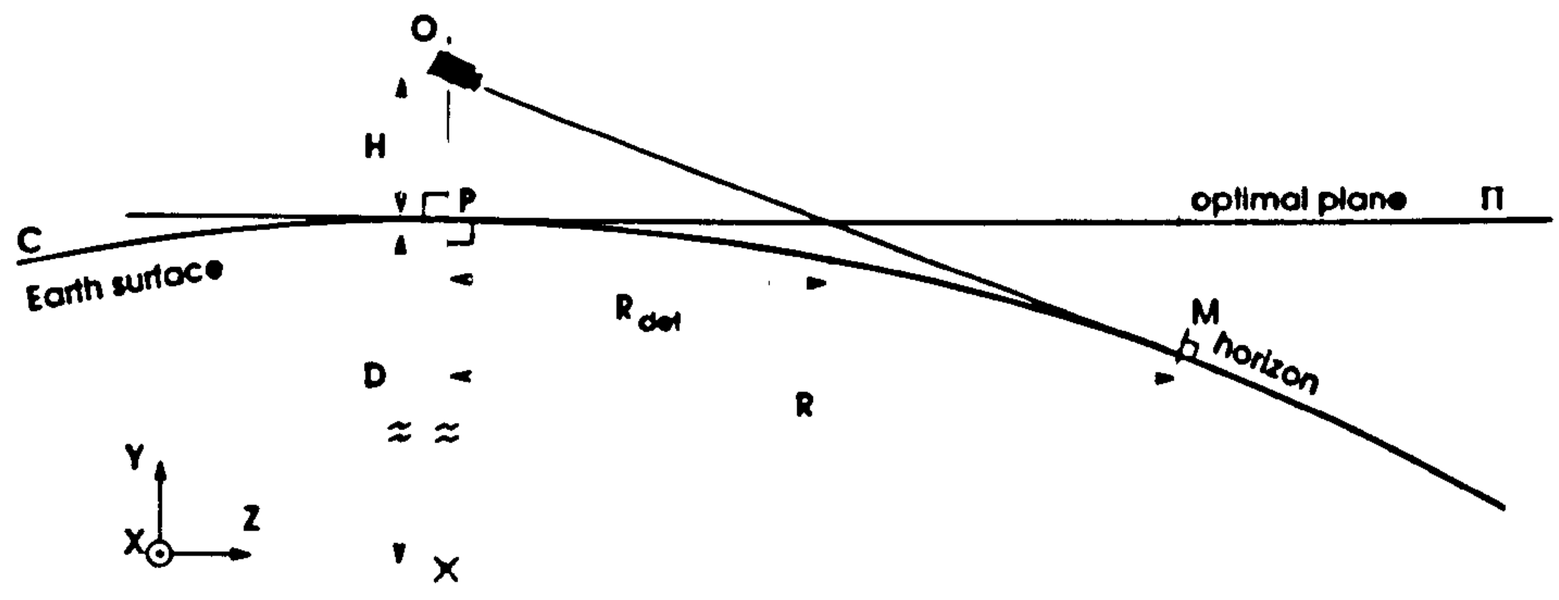
In reality, the sea surface is not perfectly flat. On a large scale it is a segment of a sphere representing the globe as shown in Figure 2.4a,b.

For camera placed at height H above the Earth surface the distance from the camera to the horizon is given as length of line $\widetilde{OM}$ connecting the projection centre O with the point M at the horizon (Figure 2.4a)

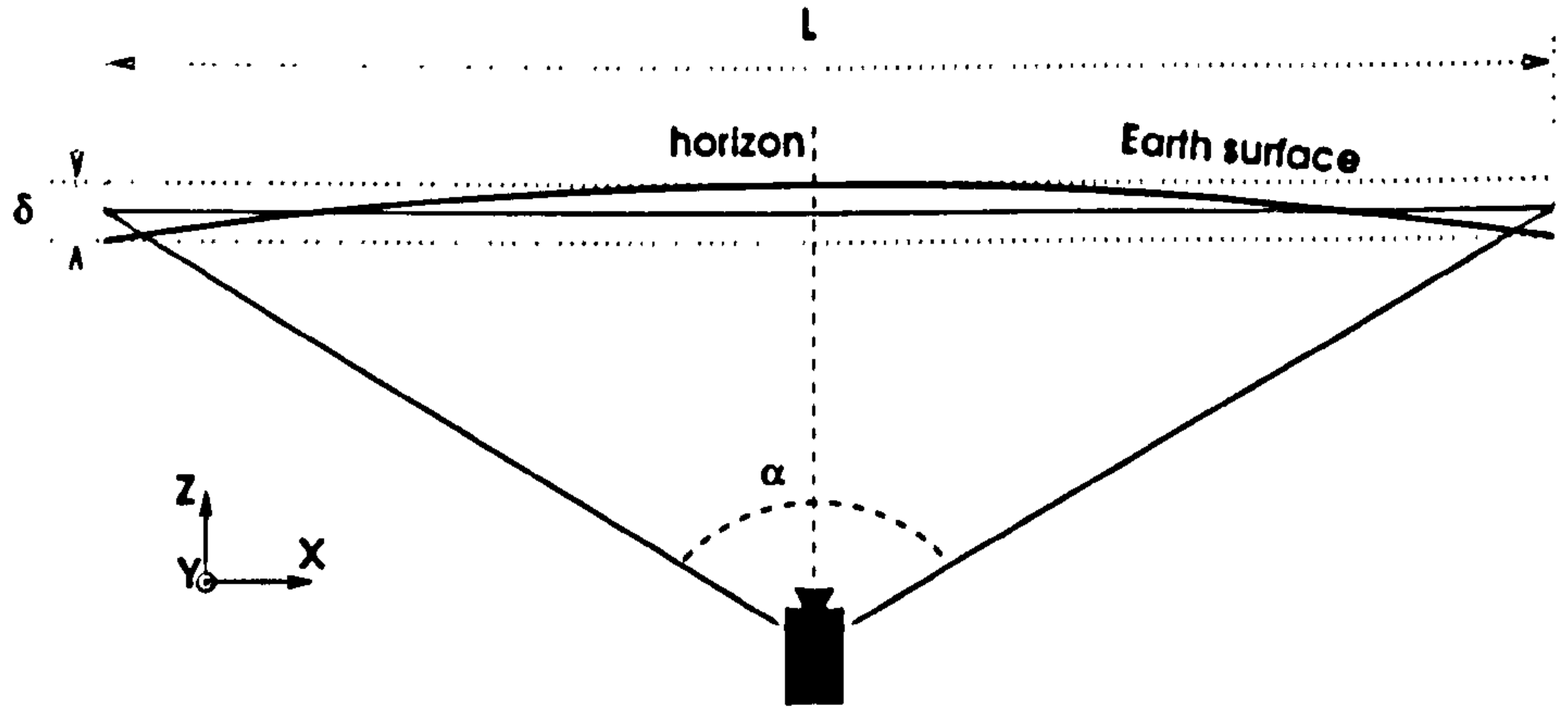
$$\left|\widetilde{OM}\right| = \sqrt{H(2D+H)} \tag{2.1}$$

where $D = 6378000\,m$ is the diameter of Earth. For example, if camera is at $H = 7\,m$ then $\left|\widetilde{OM}\right| \doteq 9449\,m$.

The perceived distance to the horizon in somewhat longer due to the refraction of the light towards the Earth's surface caused by the atmosphere. A commonly used approximation (Young, 2003,2004) of the refraction is to extend the diameter of the Earth by a factor if $\frac{7}{6}$. The modified distance to the horizon is then



(a) Side view



(b) Top view

Figure 2.4: Deviations from the optimal planar scene due to curvature of the Earth in (a) Z direction and (b) X direction of the scene coordinates.

$$\left|\widetilde{OM}\right| = \sqrt{H(\frac{14}{6}D + H)} = 10207\,m \tag{2.2}$$

The distance to the horizon is the theoretical limit of the framework range.

Figure 2.4a illustrates how the curvature of Earth violates the planar scene constraint. The point M in the scene detected at range R_{det} on the plane Π is actually located further on the globe at range R. The relation between the range R on the globe and range R_{det} detected on the optimal plane is obtained by solving for one of the two possible cross-points between line $\widetilde{OM}$ and circle sector $\mathbf{C}$ representing the Earth surface. The coordinate system is centered at point P, with Z-axis pointing to the right and aligned with the plane Π and with Y-axis perpendicular and pointing upwards. The crosspoint corresponding to the physical setting outlined in Figure 2.4a is the one with a smaller positive Z-coordinate. The other crosspoint lies behind the horizon. The coordinates of the crosspoint are obtained by solving a set of equations for line and circle

$$\widetilde{OM} : Y = H - \frac{H}{R_{det}}Z,\ Z \geq 0 \tag{2.3}$$

$$\mathbf{C} : Z^2 + (Y + D)^2 = D^2 \tag{2.4}$$

The Z-coordinate of the crosspoint corresponds to R

$$Z \equiv R = R_{det}\left(\frac{DH + H^2 \pm \sqrt{D^2H^2 - 2DHR_{det}^2 - H^2R_{det}^2}}{(H^2 + R_{det}^2)}\right) \tag{2.5}$$

Real values of R are obtained for points that lie no further than the horizon. For example, when substituting for $H = 7\,m$, $D = 6378000\,m$ and detected range $R_{det} = 500\,m$ the range on the globe $R = 501.41\,m$. Figure 2.5 illustrates the difference between the range detected on the optimal plane Π and on the globe. The deviation is non-linear and it significantly increases towards the horizon causing objects to be detected closer than they actually are.

The curvature in direction of X-axis can be safely neglected. For example, camera with $\alpha = 23°$ field of view placed at $H = 7\,m$ sees the horizon approximately $9449\,m$ away. The horizon projects as an arc $L = 2 \times 9449 \times \tan\frac{\alpha}{2\times 180} \times \pi \approx 3798\,m$ long (see Figure 2.4b). The height δ of the arc is 30 cm which is negligible error.

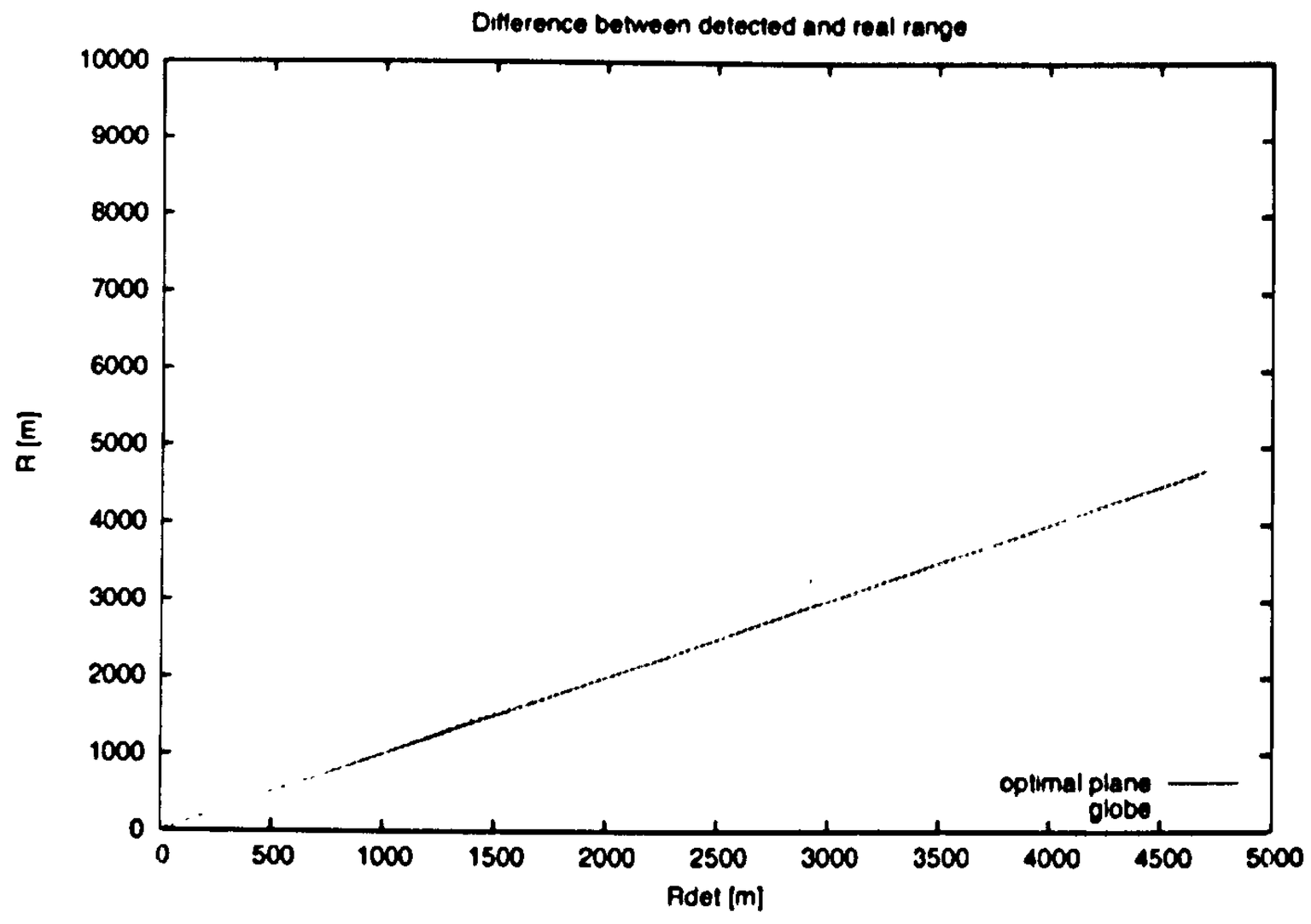


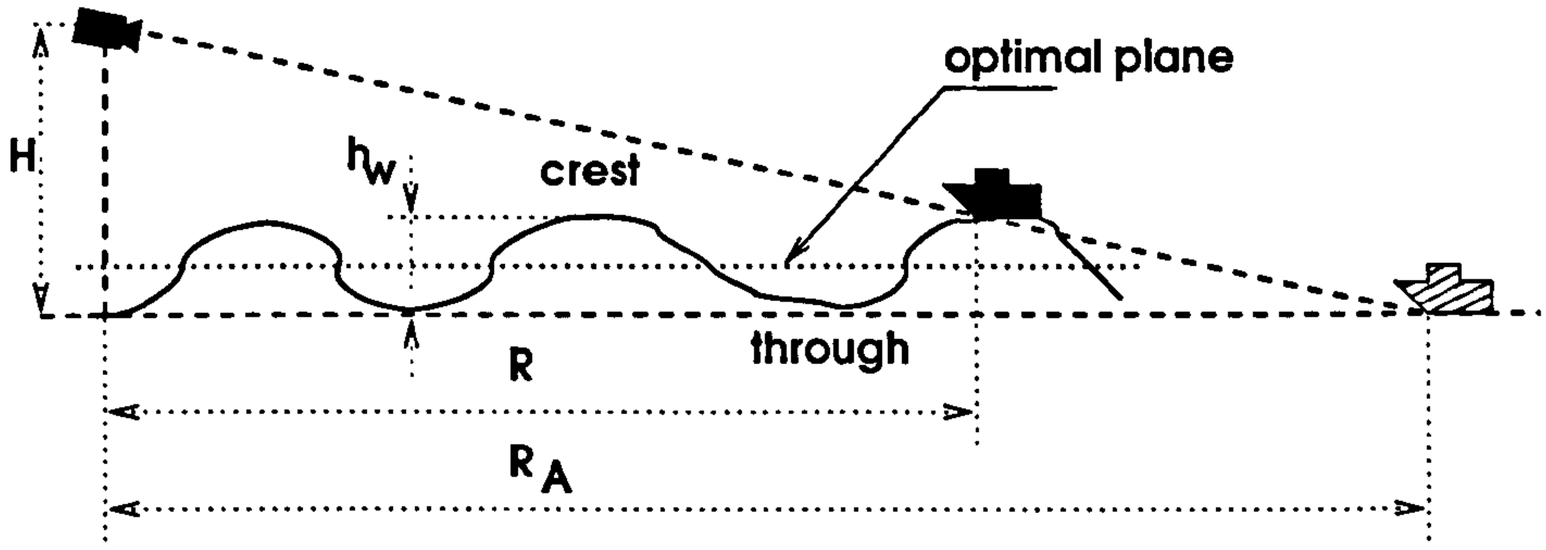
Figure 2.5: Difference between the actual range R and detected range R_{det} caused by the curvature of the Earth.

2.3.2.2 Waves

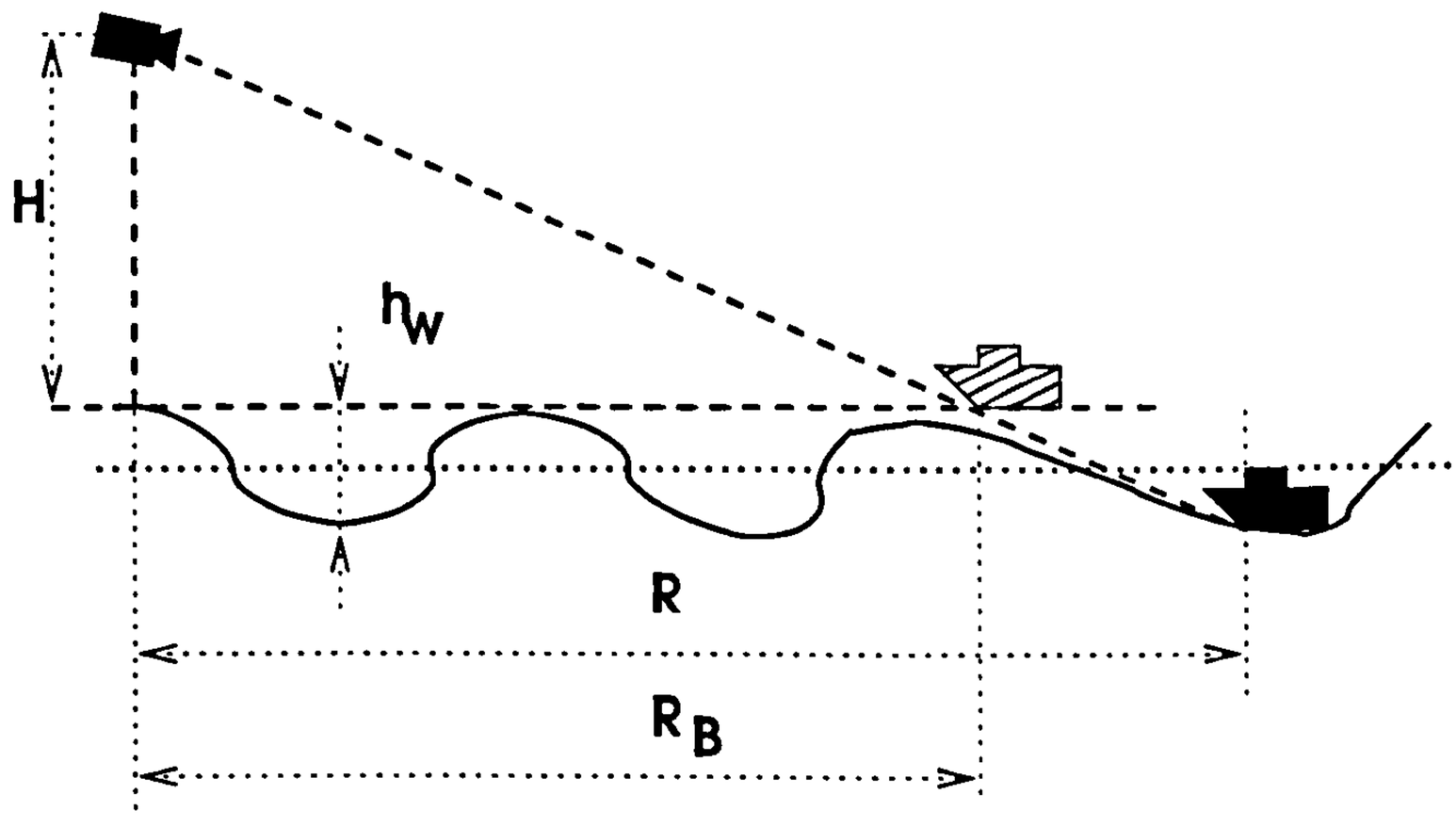
Waves are considered yet another source of deviations from the optimal planar scene model. Their influence is difficult to model analytically as the magnitude of these deviations depends on environmental conditions and location (open sea, coastal areas (Norland and Loberg, 2001)).

The height of waves h_w is measured from through to crest (see Figure 2.6a,b). A measure commonly used for quantifying the sea state is known as significant wave height (SWH). The SWH is defined as the mean value of the highest third of measured waves present. As data from the National Oceanic and Atmospheric Administration website (National Oceanic and Atmospheric Administration, n.d.) show, the average SWH values are between 0.3 to 3 metres depending on location and season of the year. Nevertheless, the SWH for coastal seas and harbours is usually considerably less, (Norland and Loberg, 2001).

Figures 2.6a,b illustrate the influence of waves on the detected range of objects. In the Case A (Figure 2.6a) an object located at range R is elevated by $\frac{h_w}{2}$ above the optimal plane. The camera at height H is lowered by $\frac{h_w}{2}$ below the optimal plane. The perceived range R_A is longer than an actual range R.



(a) Case A: camera at through, object at crest



(b) Case B: Camera at crest, object at through

Figure 2.6: The influence of waves of height h_w on the detected object ranges $R_{A,B}$.

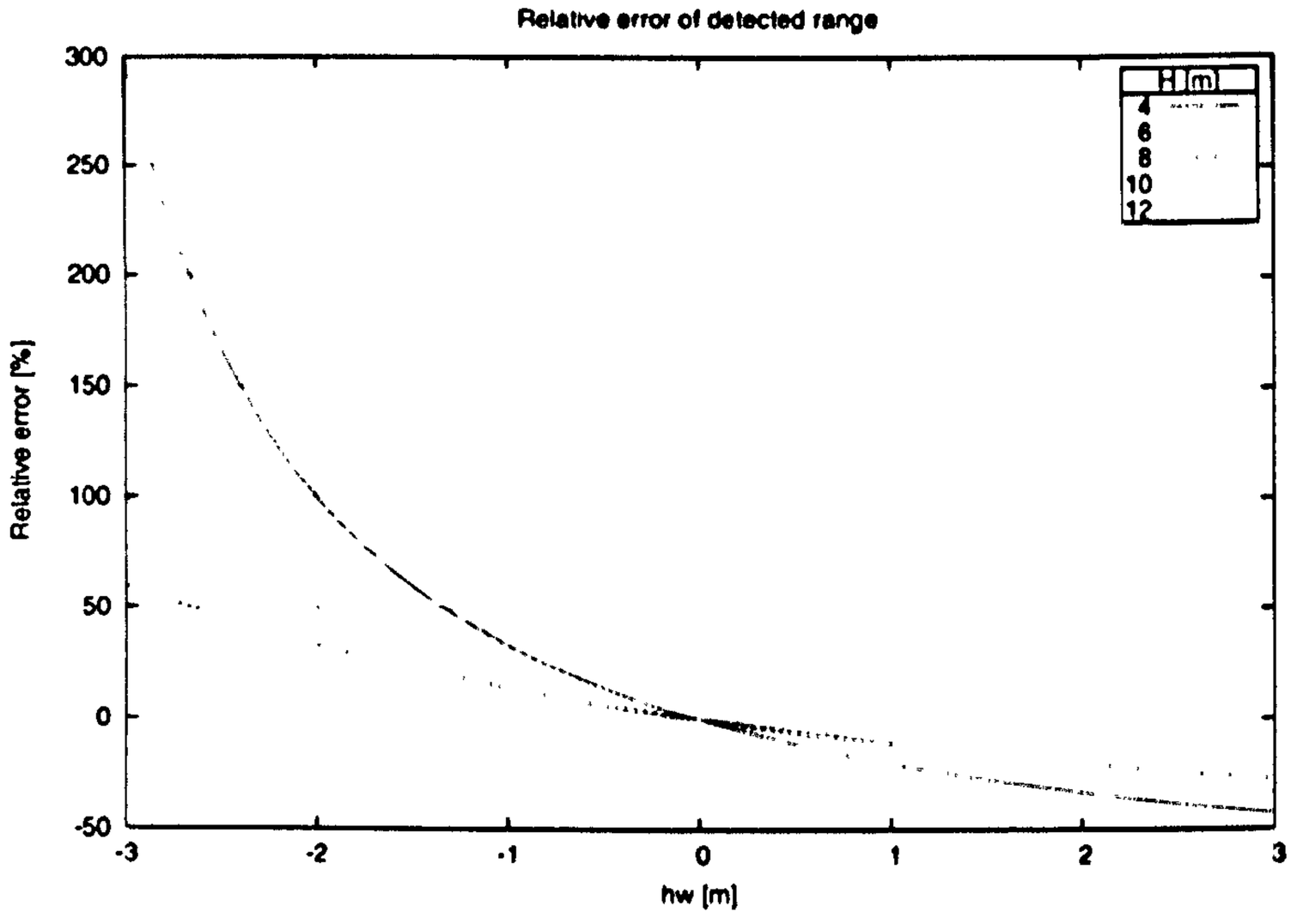


Figure 2.7: Relative error of range detection with respect to wave height h_w for various camera heights H.

The situation is reversed in the Case B, i.e. the object is lowered and the camera is elevated by $\frac{h_w}{2}$. The perceived range R_B is shorter than actual range R.

The difference ΔR between R and R_A or R_B can be derived from the triangulation

$$\frac{R}{H \mp h_w} = \frac{R_{A,B}}{H} \tag{2.6}$$

$$\Delta R = R - R_{A,B} = R\left(1 - \frac{H}{H \mp h_w}\right) = R\left(\frac{\mp h_w}{H \mp h_w}\right) \tag{2.7}$$

where wave height h_w is negative for the Case A and positive for the Case B. Equation 2.7 shows that the difference is directly proportional to the range R of the object in the scene. The relative error of detected range with respect to wave height h_w and various camera heights H is plotted in Figure 2.7. The plot indicates that the waves considerably contribute to uncertainty of the detected range.

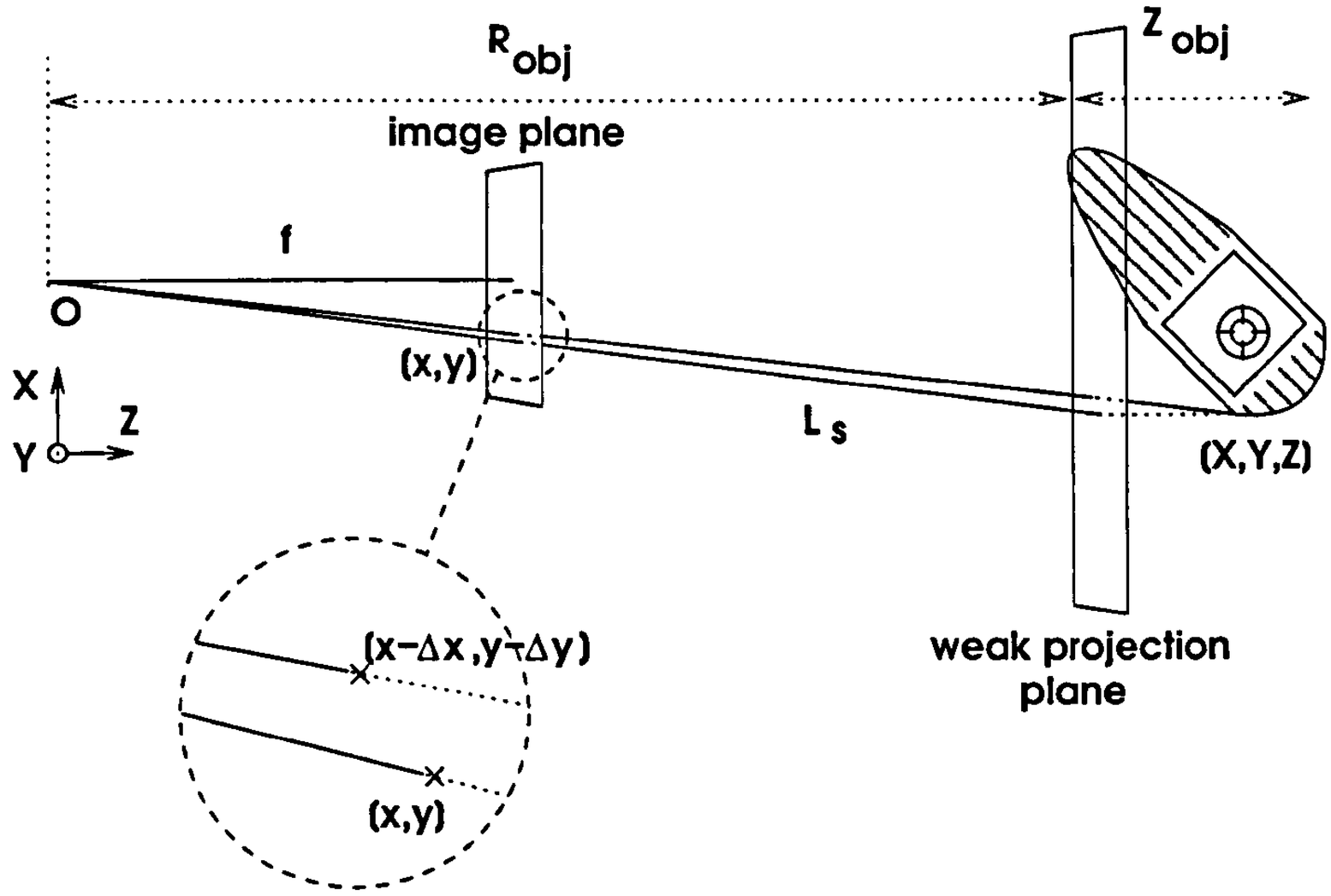


Figure 2.8: The error $(\Delta x, \Delta y)$ in image projection of the scene point (X, Y, Z) under a weak perspective projection.

2.3.3 Object Model

Optimally, the objects in the scene would be recognised and treated as three-dimensional. This would, however, require a recovery of unknown 3D structures of the objects which is a paramount task beyond the scope of the research presented here. To keep the model mathematically tractable yet adequate a simplified 2D representation called 'weak perspective' (Shapiro, 1995) is considered. The weak perspective assumes that the depth of an object Z_{obj} in the direction of the line of sight L_s is significantly smaller than the length of the line of sight $|L_s|$ (see Figures 2.2,2.3). In such case the object can be collapsed to a single plane parallel to the image plane and located at R_{obj}.

The error caused by the weak perspective projection can be expressed as a displacement $(\Delta x\, \Delta y)^T$ of the image projection of the scene point, (Banerjee, 2002) (see Figures 2.3 and 2.8)

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = -\frac{f_{pix}}{|L_s|}\left(\frac{Z_{obj}}{|L_s| + Z_{obj}}\right)\begin{bmatrix} X \\ Y \end{bmatrix} \tag{2.8}$$

where f_{pix} is the focal length of the camera in pixels and $(X\, Y)^T$ are the first two coordinates of the point in the scene. Assuming that the object is projected near the centre of the image so that $X \ll R_{obj}$ then $|L_s|$ can be approximated

from triangulation

$$|L_s| \approx \sqrt{H^2 + R_{obj}^2} \tag{2.9}$$

Equation 2.8 indicates that small ratios of $\frac{f_{pix}}{|L_s|}$ (long range of objects in the scene compared to focal length), $\frac{X}{|L_s|}$ and $\frac{Y}{|L_s|}$ (small field of view) contribute to the validity of the model.

For example, if a 5 metres long vessel ($Z_{obj} = 5\,m$) detected at 150 metres ($R_{obj} = 150\,m$) faces the camera with focal length $f_{pix} = 1928\,pix$ overlooking the scene from height $H = 7\,m$ then the projection displacement caused by weak perspective approximation is

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = -\frac{1928}{\sqrt{7^2 + 150^2}} \left(\frac{5}{\sqrt{7^2 + 150^2} + 5} \right) \begin{bmatrix} X \\ Y \end{bmatrix} = 0.4137 \begin{bmatrix} X \\ Y \end{bmatrix} \tag{2.10}$$

The value of f_{pix} is given by the physical dimensions of the pixels. CCTV cameras used in industrial and surveillance applications typically operate with square pixels with size s_{pix} (see Figure 2.2) between 4 and 12 μm (Hitachi Denshi, n.d.) and lenses with focal lengths f between 1 to 75 mm. The value of f_{pix} used in the example corresponds to $f = 16\,mm$ and $s_{pix} = 8.3\,\mu m$. The displacement error as a function of the detected range R_{obj} of the vessel in the example above is is plotted in Figure 2.9.

The approximation given by Equation 2.8 is rather pessimistic. Most points belonging to an object in the real maritime scene are located closer to the weak perspective plane than the maximum object depth Z_{obj}. Distant objects monitored by an elevated camera at a relatively low tilt angle ω are partially self-occluded and only fractions of their structures are visible. For example, less than a half of the width of the small fishing boat in Figure 2.1d is visible.

2.4 Temporal Context

Temporal context analysis looks at the dynamics of maritime scenes. The changes of structure and appearance of maritime scene are due to several kinds of motion and environmental factors. Four kinds of motion in maritime scenes can be recognised:

- independent motion of objects in the scene,

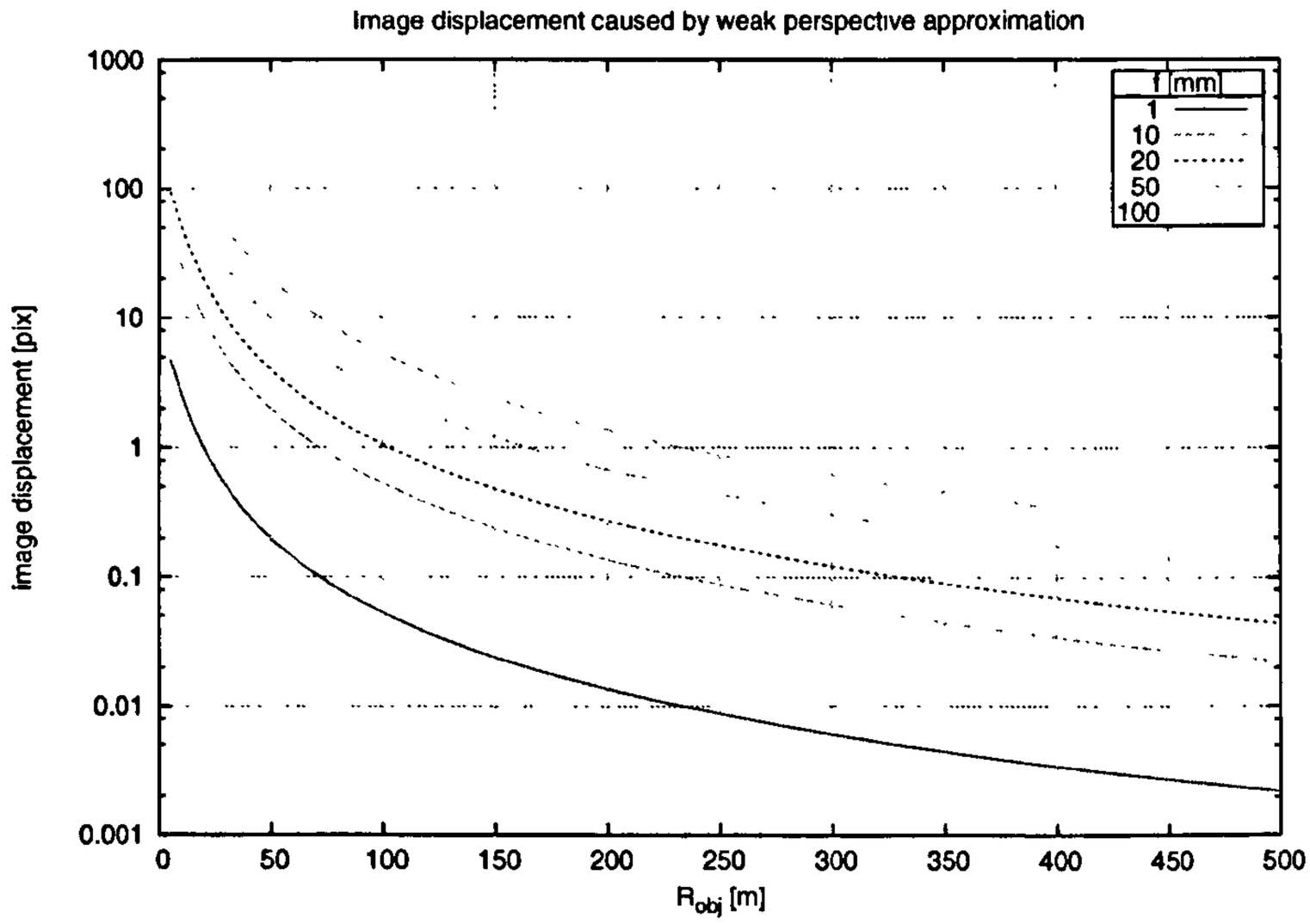


Figure 2.9: Projection displacement caused by the weak perspective approximation as a function of the detected object range R_{obj} and various focal lengths f of the camera.

- motion of objects due to interaction with waves,
- motion of the sea surface due to sea waves,
- global displacement of the scene due to camera self-motion.

Environmental factors such as clouds are locally or globally changing the illumination of the scene over time. Cloud shadows can locally reduce the illumination of the scene causing darker patches to appear on the sea surface. These darker patches travel approximately at the speed of the clouds. The size and shape of the patches depend on the weather conditions.

The objects are assumed to obey laws of physics, namely rigid body mechanics. Their velocity and acceleration are assumed finite and smooth functions of time, their mass non-zero and approximately constant throughout their motion. The sea, however, does not obey rigid body mechanics and it's motion is modelled by fluid dynamics.

The interaction of rigid objects with the sea is characterised by non-linear models, (Kim et al., 1987), as fluid dynamics interacts with rigid body mechanics. The resulting motion described as 'rolling' can be approximated as an oscillatory motion with time-varying amplitude and frequency. The

amplitude and frequency of these oscillations is given by the frequency of the waves, amplitude, and weight, dimensions and geometry of the object. The direction of the oscillations depends on these factors as well as on the pose of the object with respect to the direction of the propagating waves.

2.4.1 Motion of the Sea

Sea surface undergoes permanent motion that consists of vertically oscillating waves that propagate horizontally in all directions. Waves can be divided into two categories with respect to their prime source. Natural waves are generated by interaction of the water surface with wind. Artificial wakes are generated by interaction of the water with either static or moving objects.

Among many environmental factors that contribute to generation of natural waves the most important are: wind strength, depth and shape of pool and location (coastal seas, open seas). The wave motion is difficult to model in a deterministic way, even in a controlled environment and for a one-dimensional case, (Capitao and de Carvalho, 2000), as the numerous factors mentioned above interact in complex ways. Due to its periodic nature, this motion is typically modelled as a Fourier series with stochastic time variant parameters, (Belmont and Morris, 1994; Capitao and de Carvalho, 2000; Kim et al., 1987).

Section 2.3.2 described regular patterns such as wakes generated by the hulls of moving vessels, underwater streams and other sources. These wakes propagate along the surface away from their source before they gradually lose energy and fade away or blend into the surrounding background. A typical size and life span of the wake depends on the size and speed of the object that caused it and the state of the sea. Wakes can last from a couple of seconds up to several minutes. The wakes propagate for longer on a calmer sea as there is less attenuation due to interferences with natural waves. The propagation speed, magnitude and direction of the wakes are influenced by physical properties of the objects.

2.4.2 Motion of Objects

Categorisation of maritime craft is administered according to the intended purpose and functionality by official bodies such as Den Norske Veritas (DNV, 2004). Following categories are typically recognised:

- ships

- high speed, light craft and naval surface craft
- fixed offshore installations
- other objects

For the purpose of tracking and surveillance, a classification with respect to activity of maritime craft is more appropriate. The main objective of such classification is to distinguish objects by their motion dynamics rather than by their purpose or appearance. If the motion is determined relative to a static point of observation then three classes of objects can be identified:

- *Static objects* do not exhibit any kind of motion. These are, for example, piers, poles and other man-made constructions, either embedded into the sea floor or stretching from the shore. Natural objects such as rocks or shores belong to this class as well.
- *Fluctuating objects* exhibit a short-term semi-periodical motion due to interaction with waves. Actual position of these objects is constrained to a certain limit around a fixed or drifting position. These are, for example, buoys, marine craft in an inactive state at a mooring or tied up to other static objects; sea birds sitting on the sea surface; floating objects such as debris, natural objects such as floating seaweed. The common factor of these objects is that their motion is purely by interaction with the waves.
- *Moving objects* can undergo any kind of motion in any direction. For maritime scenes the motion is constrained by the GPC (Worral et al., 1995) to two dimensions along the sea surface. Activities such as a hydroplane taking off are considered as special cases and they are beyond the scope of the research presented here as they violate the GPC. The speed, acceleration and maneuverability of various maritime craft are usually given by their physical properties such as mass and geometry.

The above classification of objects is not strict and objects can change categories over time. For example, a stationary vessel apparently motionless in calm seas belongs to a *static* category even though it is not embedded into the sea floor. Under different weather conditions it might actually start to fluctuate around a fixed point and thus be classified as *fluctuating*. If it starts to move on its own it will become a *moving* object.

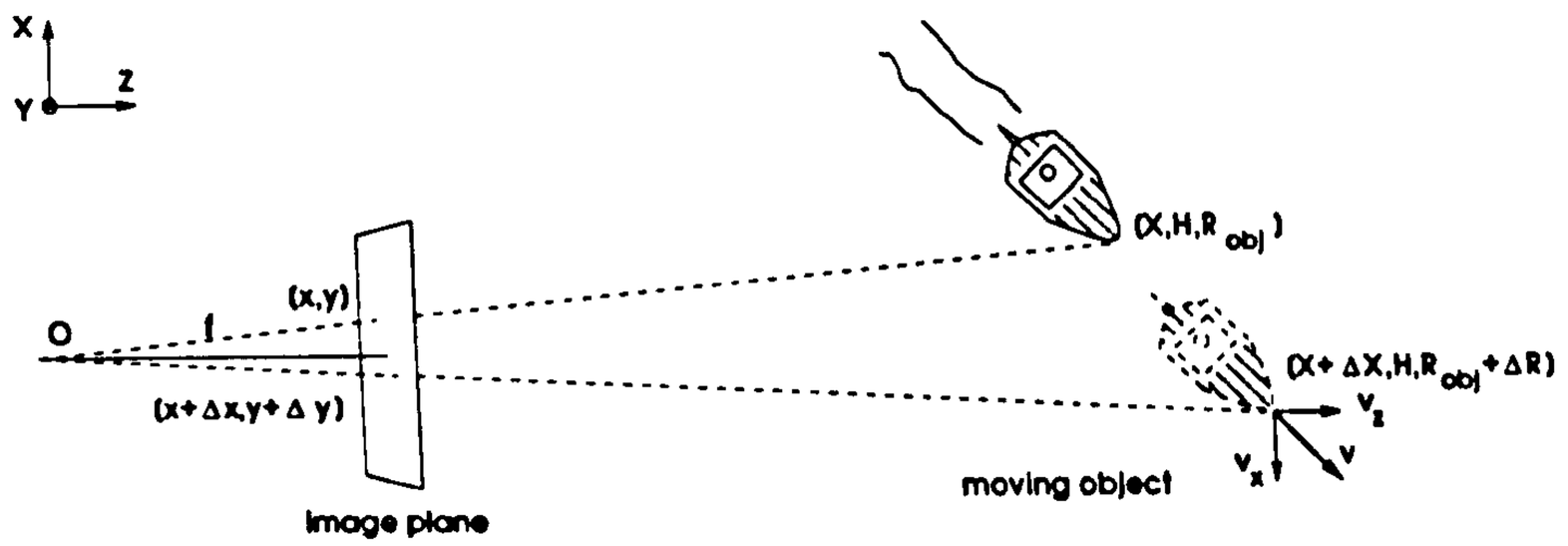


Figure 2.10: An object in the scene travels from (X, H, R_{obj}) to $(X + \Delta X, H, R_{obj} + \Delta R)$ between two consequent frames of the sequence. The motion projects onto the image as displacement $(\Delta x, \Delta y)$.

2.4.3 Projected Displacement

Assuming that maritime objects are rigid, their motion projects as a displacement of their images between two consequent frames in the sequence. Most motion detection algorithms assume a limited projected displacement, (Smith, 1998; Zhang and Lu, 2001), in order to simplify the detection and to reduce computational overhead. Limited projected displacement in combination with the frame rate defined as reciprocal value of time between two frames in a sequence stipulate the maximum detectable speed of objects in the scene.

The relation between the projected displacement and the speed of an object in the scene is obtained by the following analysis. A vessel moving in the scene is monitored by a camera at height H above the scene with a focal length f tilted by an angle ω (see Figure 2.3). Two components of the vessel motion are considered - parallel and perpendicular to the image plane (see Figure 2.10). Motion in other directions is obtained by linear combination of these two components.

The point at location $[X, Y = H, Z = R_{obj}]$ lying on the sea plane Π projects onto the image plane at position

$$x = \frac{f_{pix} X}{H \sin\omega + R_{obj} \cos\omega} \tag{2.11}$$

$$y = f_{pix} \frac{H \cos\omega - R_{obj} \sin\omega}{H \sin\omega + R_{obj} \cos\omega} \tag{2.12}$$

The projection is derived in detail in Chapter 8.

The vessel moves between two frames from location $[X, H, R_{obj}]$ to $[X + \Delta X, H, R_{obj} + \Delta R]$ in the scene. Its displacement $[\Delta X, 0, \Delta R]$ in the scene

projects as a displacement $[\Delta x, \Delta y]$ in the image. The vessel moving at velocity $\mathbf{v} = (v_x, v_y)$ travels distance ΔX and ΔR in time $t = \frac{1}{q_r}$ reciprocal to the frame rate q_r of the framework

$$\Delta X = v_x t = \frac{v_x}{q_r} \tag{2.13}$$

$$\Delta R = v_y t = \frac{v_y}{q_r} \tag{2.14}$$

The corresponding projected displacement can be determined from 2.11 and 2.12 as

$$\Delta x = f_{pix} \frac{\Delta X}{H \sin\omega + R_{obj} \cos\omega} \tag{2.15}$$

$$\Delta y = f_{pix} \left[\frac{H \cos\omega - (R_{obj} + \Delta R) \sin\omega}{H \sin\omega + (R_{obj} + \Delta R) \cos\omega} - \frac{H \cos\omega - R_{obj} \sin\omega}{H \sin\omega + R_{obj} \cos\omega} \right] \tag{2.16}$$

The image displacement in terms of vessel's range R_{obj}, velocity $\mathbf{v}$ and frame rate q_r is obtained by substituting from Equations 2.13 and 2.14 into 2.15 and 2.16

$$\Delta x = \frac{f_{pix} v_x}{q_r (H \sin\omega + R_{obj} \cos\omega)} \tag{2.17}$$

$$\Delta y = \frac{-f_{pix} H v_y}{(H \sin\omega + R_{obj} \cos\omega)(H q_r \sin\omega + (v_y + q_r R_{obj}) \cos\omega)} \tag{2.18}$$

Both equations can be simplified under assumptions that angle ω is relatively small and that $v_y \ll q_r R_{obj}$. The displacements can be approximated as

$$\Delta x \approx \frac{f_{pix} v_x}{q_r R_{obj}} \tag{2.19}$$

$$\Delta y \approx \frac{-f_{pix} H v_y}{q_r R_{obj}^2} \tag{2.20}$$

2.4.3.1 An Example Scenario

Parameters of the motion detection can be determined by evaluating the above approximations for a required maximum detectable velocity $\mathbf{v}$ at a specific minimum range R_{min}.

Parameter	Value
Imaging area size	1/2"
Imaging area width	823 pix
Imaging area height	592 pix
Pixel width s_x	8.3 μm
Pixel height s_y	8.3 μm
Horizontal multiplier w	6.4
Vertical multiplier h	4.8

Table 2.1: Parameters of the Hitachi Denshi (n.d.) KPF1E camera used in the example scenario. The values of horizontal and vertical multipliers for 1/2" imaging area are obtained from RMA Electronics Inc. (2005)

The following scenario serves as an example. The framework monitors a harbour entrance 400 m wide and 1 km long. The camera overlooking the entrance is mounted on a pole at height $H = 7\,m$ above the sea and tilted by $\omega = 2°$. The framework is required to capture objects moving at speeds up to 50 $knots$ (approx. 28 m/s) inside the harbour entrance in order to capture small recreational craft which are often capable of such high speeds.

A standard machine vision camera such as Hitachi Denshi (n.d.) with a fixed focal length lens is used. The selected camera can capture 25 frames per second. The focal length of the lens depends on the size of the camera's imaging area and the required field of view at a specified range. The parameters of the camera are summarised in Table 2.1.

The focal length f is given as, (RMA Electronics Inc., 2005)

$$f = w\frac{R}{W_{fov}}\,[mm] \tag{2.21}$$

where w is a horizontal multiplier (see Table 2.1 for a specific value), R is the specified range and W_{fov} is the required width of the field of view. For the above mentioned harbour entrance the values are $R = 1000\,m$, $W_{fov} = 400\,m$ and the focal length is $f = 6.4 \cdot \frac{1000}{400} = 16\,mm$.

In order to establish the minimum detectable range R_{min} of the framework the height of the field of view must be obtained by rearranging Equation 2.21 and replacing horizontal multiplier w with the vertical one, h (see Table 2.1 for a specific value)

$$H_{fov} = h\frac{R}{f} = 4.8 \cdot \frac{1000}{16} = 300\,m \tag{2.22}$$

The value of R_{min} is obtained from triangulation illustrated in Figure 2.11. The vertical field of view γ is determined as

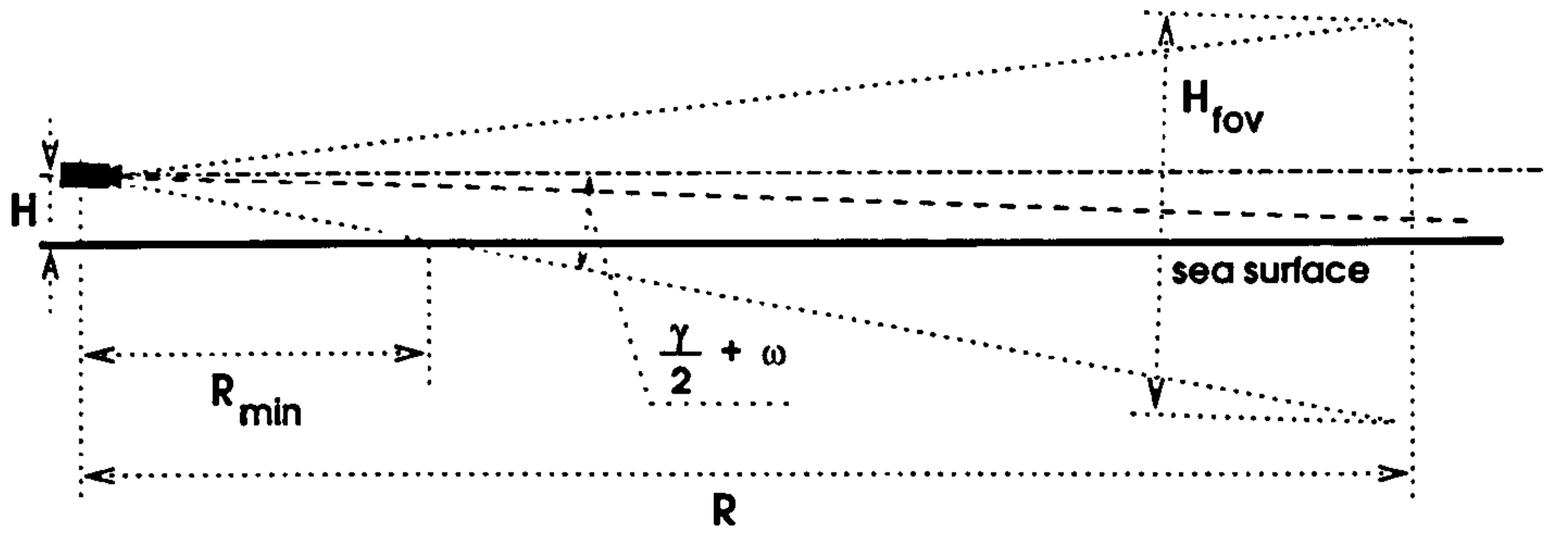


Figure 2.11: Minimum detectable range R_{min} is determined from the height H of the camera and the angle $\frac{\gamma}{2} + \omega$, where $\frac{\gamma}{2}$ is the half of the vertical field of view and ω is the camera tilt

$$\gamma = 2\arctan(\frac{H_{fov}/2}{R}) = 2 \cdot \arctan(\frac{300/2}{1000}) = 17° \tag{2.23}$$

and R_{min} as

$$R_{min} = \frac{H}{\tan(\frac{\gamma}{2} + \omega)} = \frac{7}{\tan(\frac{17}{2} + 2)} \doteq 38\,m \tag{2.24}$$

The maximum projected displacement are obtained by substituting corresponding values into Equations 2.19 and 2.20

$$\Delta x \approx \frac{\frac{f}{s_x} v}{q_r R_{min}} = \frac{\frac{16 \cdot 10^{-3}}{8.3 \cdot 10^{-6}} \cdot 28}{25 \cdot 38} \doteq 57\,pix \tag{2.25}$$

$$\Delta y \approx \frac{-\frac{f}{s_y} H v}{q_r R_{obj}^2} = \frac{-\frac{16 \cdot 10^{-3}}{8.3 \cdot 10^{-6}} \cdot 7 \cdot 28}{25 \cdot 38^2} \doteq 11\,pix \tag{2.26}$$

The results show that in order to detect craft moving at maximum speed $50\,knots$ anywhere inside the monitored harbour entrance the expected projected displacements are up to $57\,pix$ in horizontal and $11\,pix$ in vertical direction.

2.5 Research Objectives

The contextual analysis provided in previous sections supplies information necessary for the formulation of the objectives addressed in this thesis. The objectives of the research are to deliver and evaluate a machine vision based framework that

- detects any objects in a maritime scene from a sequence of images captured by a camera monitoring the scene,
- locates these objects with respect to a fixed point in the scene,
- tracks any motion of the detected objects,
- estimates location, velocity and their uncertainties in units that are related to the scene, i.e. location in metres, velocity in knots,
- based on the estimates, identifies any events in the scene that by definition require attention of a human operator, i.e. collisions, threats, intrusions,
- gives an early warning to the human operator if any such activity is identified.

Automated early warning also known as cueing is an essential feature of many surveillance systems. It alerts the human operator of a situation requiring his attention and it gives them time to fully assess it and make competent decisions.

A study by Hitchcock et al. (2003) concerns itself with the influence of cueing on the vigilance and decision making processes in humans. As the first study of its kind (Hitchcock et al., 2003) also investigates the effect of the reliability of the cueing. The results confirm that the ability of correct assessment is directly related to the quality of the cueing. If a vigilance task is to be automated then cueing must have the highest possible reliability to make it relevant and effective. Cueing might otherwise have adverse effects, making a chance of correct assessment by the operator less likely.

Based on these conclusions the following two requirements are formulated regarding the early warning capability of the framework:

- the ratio of relevant to irrelevant information in the output should be as high as possible and
- the relevant information should be as reliable as possible.

Apart from principal functionality the emphasis of the research is on following properties of the framework:

- the performance of the framework is independent on the scene and object appearances as well as environmental factors,

- performance does not significantly change with time,
- systematic errors identified at any processing stage are compensated,
- stochastic errors identified at any processing stage are either compensated or minimised.

2.6 Constraints and Assumptions

The following constraints and assumptions arising from the contextual analysis are applied to the problem domain in order to arrive to a suitable and effective solution:

- the maritime scene is represented by an infinite horizontal plane corresponding to the sea surface with objects located on it,
- both systematic and stochastic deviations from the planar model are identified to allow for their compensation or reduction,
- the Ground Plane Constraint holds for the objects, i.e. all objects detected are assumed to lie on the plane representing the sea surface,
- a single camera monitors the plane from an elevated point in such a configuration that the largest possible area of the plane is projected onto the image,
- the camera is fully calibrated, i.e. all the necessary intrinsic parameters are obtained in advance; the height H of the camera above the sea surface is known,
- the background of the maritime scene is represented by the sea; the land or sky are excluded from the processing,
- the majority of the scene structure is the sea, objects occupy minority of the scene,
- the appearance of objects is not uniform and it varies considerably in numerous aspects as there are several kinds of objects to be encountered,
- a weak perspective planar model can be used to represent objects in maritime scene,

- the camera is either static or moving in any direction parallel to the scene plane,
- the sequence consists of fixed size intensity frames taken at a constant frame rate; processing of colour is avoided.

An assumption of a single camera input into the framework is based on the fact that the Ground Plane Constraint (GPC) resolves ambiguity of perspective projection where a single point in the image represents infinite number of points on a ray coming through the optical centre and the point in the image (see Figure 2.2). The GPC specifies that objects lie on the sea plane for which $Y = 0$ (see Figure 2.3). If the height H and intrinsic parameters of the camera are known, then it is possible to unambiguously estimate the location of the object with respect to the camera from a single view. No multiple views generated by, for example, stereo camera rig (Li, 1994; Li and Lavest, 1995) are necessary in principle.

Many machine vision applications targeting natural outdoor scenes (Buluswar and Draper, 1994; Campbell and Thomas, 1996; Skarbek and Koschan, 1994; Lucchese and Mitra, 2001) operate on colour images. There is, however, a number of issues associated with the use of colour in image processing.

The pixel value in the intensity image is proportional to the intensity and wavelength of the incident light. There are typically three values per pixel in colour images corresponding to intensities of red, green and blue regions of the spectra of the incident light. The colour image theoretically contains three times more data than an intensity image. The increase in volume of image data has to be considered in time-critical applications. Some applications reduce the amount of data in colour images by lossy compressions, (Murray and van Ryper, 1994), pp. 456-464. Such reduction of image data can have a serious impact on image quality, (Wang et al., 2004), namely precise location of edges, resolution of region boundaries, etc.

Reflections of objects on the sea surface are common in maritime scenes. Reflections usually have the same colour as the reflected objects. Segmentation of a scene with reflections can present a challenge for colour-based methods. The colour attributes of the background and the objects are strongly influenced by the illumination which is a dynamic, constantly changing process in outdoor scenes, (Buluswar and Draper, 1994).

In addition, there are numerous limitations to technology used in colour imaging devices as identified by Martinkauppi (2002):

- *clipping* - occurs when one or more colour channels becomes null or saturated due to a presence of dark or bright objects in the scene,

- *non-linearity of the sensor* - the response of the sensor is non-linear, (Shafique and Shah, 2004; Tsin et al., 2001) it can be different for each channel and it can be influenced by features such as Automatic Gain Control commonly used in applications with time-varying illumination conditions,

- *white balance adjustment* - is necessary in order to avoid bias of sensor response caused by diverse illumination conditions.

Infra-red and intensifying imagery used in maritime night vision applications normally provide monochrome intensity based images, (Vistar Night Vision Limited, 2004a; Vistar Night Vision Limited, 2004b; Vistar Night Vision Limited, 2004c). This offers an opportunity of straight-forward integration of the framework within such systems as the data representations are alike.

2.7 Methodology

A research methodology suitable for the addressed problem is specified in order to deliver the proposed objectives. The methodology is discussed from four aspects: research design, architecture of the framework, sample sequences used for development and methods of evaluation. Research design outlines the procedures necessary to obtain a plausible solution to the problem. The architecture of the proposed framework follows a bottom-up control model (Morris, 2004; Batlle et al., 2000) which is a feasible architecture for the type of the problem addressed in this thesis. Video sequences used for development of the framework are presented. Conditions under which the sequences were obtained are detailed. Finally, the methods of evaluation are described.

2.7.1 Research Design

The research deals with real open-world maritime scenes which are essentially of stochastic nature influenced by phenomena such as weather, daylight, etc. The stochastic properties of open-world maritime scenes are often difficult to model precisely as illustrated by Capitao and de Carvalho (2000) and Preetham et al. (1999) with many factors to be considered.

It would be rather complex to encompass all these factors into a single mathematical model of the problem as their interactions are often unknown. Any simplifications to the model would have to be revised in order to preserve the ability of the framework to operate on real scenes.

An alternative approach is to obtain a representative sample of the problem domain and derive the solution using the sampled data. In the case of maritime scenes the sampled data are represented by a set of video sequences captured at conditions similar to those in the eventual application of the framework.

In order to make the research more manageable the problem is split into a series of linked sub-tasks that are solved separately. The specifications of input and output of each sub-task are outlined in advance. The solution is obtained by searching out, developing and assembling methods that are candidates for the solution to a partial sub-task. The best performing candidate is then selected using either relative or ground-truth based experimental evaluation.

Once all the partial sub-tasks are solved, the whole framework is assembled and cross-validated in order to verify that the objective of the independence on object and scene appearance is delivered.

2.7.2 Framework Architecture

There are numerous system architectures available in machine vision and their choice depends on the underlying problem, (Morris, 2004) pp. 213-216. Batlle et al. (2000) review these architectures using an example of a system for image understanding of natural outdoor scenes using colour information. They categorise systems into three groups depending on a control strategy these systems employ: top-down, bottom-up and hybrid.

- *Top-down* architecture starts with a hypothesis about possible objects in the image. A set of features, attributes and relationships that support the hypothesis is generated. The hypothesis is then verified by checking whether a same set obtained from the image supports it. The hypothesis is either accepted or rejected depending on the verification result. The top-down architecture is suitable for problems with a limited number of well-defined hypotheses to be tested. Such approach is of limited use in maritime scenes as the number of objects, their variety and, therefore, number of hypotheses to test is virtually unlimited.

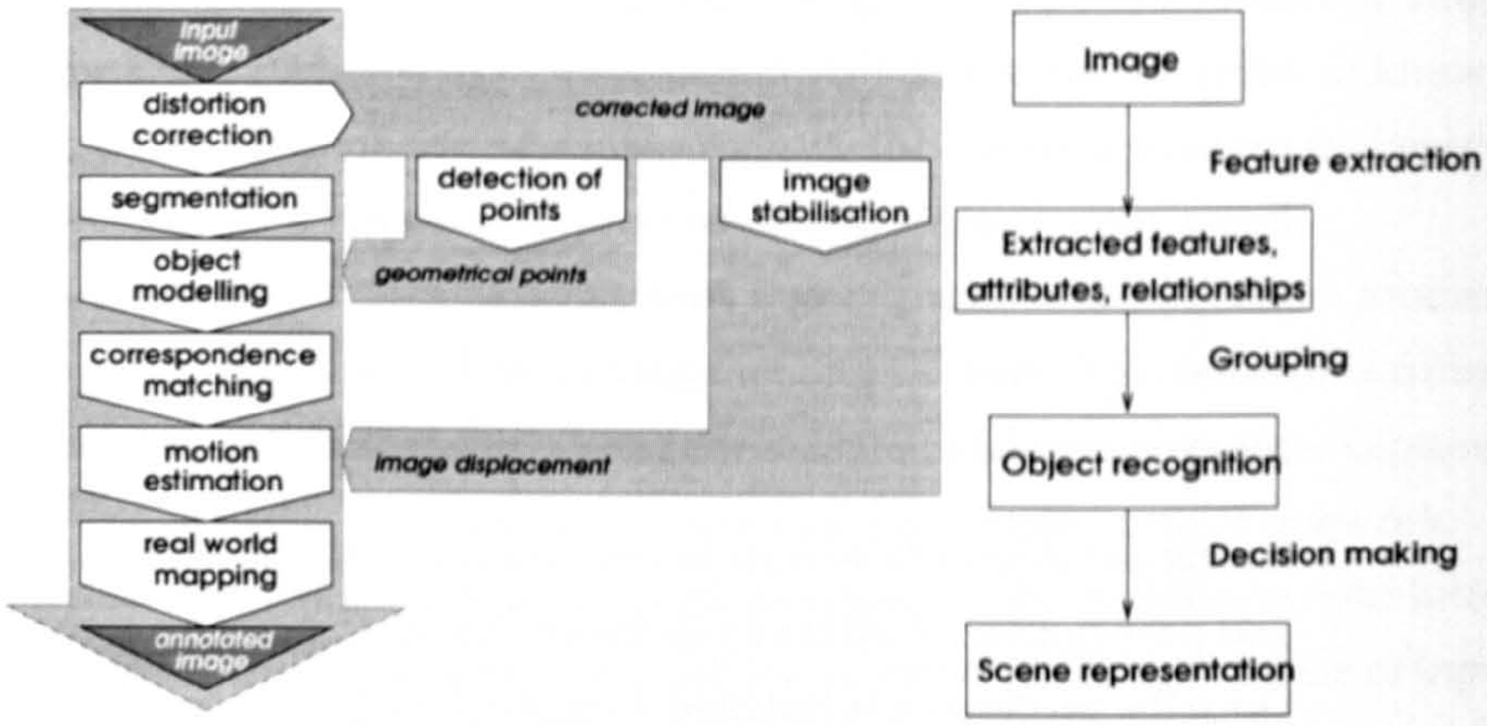
- *Bottom-up* architecture follows on Marr's vision model (Marr, 1982). Marr

formalised an architecture of a machine vision system that is coherent with a structure of the human visual system. In Marr's vision model primary features called tokens are extracted from the image, assembled into a more complex set of compositions called the primal sketch. The compositions are grouped into surfaces to provide object description called $2\frac{1}{2}D$ sketch used in recognition. The model follows a logical progression: data are processed and refined, decisions are made.

- *Hybrid* architecture attempts to overcome limitations such as inflexibility and propagation of errors from which the other two architectures suffer. Once the processing is initiated it does not stop until a result is obtained whether it is correct or not. Errors occurring at any levels of processing will propagate and influence the outcome. Hybrid architecture mixes both top-down and bottom-up principles. Features, attributes and relationships extracted from the image in a bottom-up branch are used in hypothesis generation. The hypothesis is then tested on the data from the image in a top-down branch.

The proposed framework is derived from the bottom-up data processing hierarchy. Generic models of objects are first obtained by grouping together geometric primitives detected in the image and projecting them onto weak-perspective planes. The locations and dynamic characteristics of the objects are then established using their generic models. Final decisions are based on the characteristics obtained in previous steps. The architecture of framework is shown in Figure 2.12a. The framework adheres to the bottom-up hierarchy (see Figure 2.12b) with processing path divided into multiple modules. Each module takes data from one or more outputs of previous modules, processes them and passes the results as inputs to following modules.

In principle, the information about the maritime scene flows in one direction, from the visual sensor through the framework to the operator. The framework operates as an information filter that detects and represents relevant information about activity of objects and suppresses redundant information such as the motion of the sea background. The bottom-up hierarchy matches this type of one-directional processing control.



(a) Proposed framework

(b) Bottom-up architecture

Figure 2.12: The architecture (a) of the proposed framework adheres to the bottom-up structure (b), (Morris, 2004; Batlle et al., 2000).

2.7.3 Development Sequences

2.7.3.1 Absence of Ground Truth

The research is using video sequences obtained at conditions matching those expected in the targeted applications of the framework. The sequences represent a sample of the real maritime scenes including objects and activities likely to be encountered in the proposed applications.

The drawback of using real maritime sequences is the absence of ground truth. Ground truth can be defined as the actual facts of a situation, without errors introduced by sensors or human perception and judgement. Limited research resources did not allow to obtain maritime sequences with ground truth such as true scales, orientations, locations and velocities of objects or state of the sea. Two workarounds are applied in cases where ground truth is essential for evaluation.

The first workaround is applied when only the presence or absence of an object has to be established. The sequence is interactively surveyed by a human and occurrences of objects and their activities in every frame are marked down. The obtained ground truth provides a rough characterisation of the activity in the scene.

The second workaround is applied when precise locations of objects in each frame are necessary. An artificial sequence is generated by superimposing

images of objects with a known scale and geometry onto either real or artificial backgrounds. Different scenarios of object activity are obtained by a controlled placement of objects in each frame. Each generated frame is blended with Gaussian noise that approximates various types of noise present during image acquisition. Real backgrounds consist of sea surfaces with no objects in the scene. Various states of the sea are captured, ranging from calm to rough seas. Artificial backgrounds consist of two types of noise that are typically encountered at a pixel level - Gaussian and uniform. The obtained artificial ground truth sequences allow to evaluate object detection as well as precisions of location and velocity estimations.

Use of artificial sequences is not novel in development of algorithms targeting maritime scenes. Messer and Kittler (2000) and Messer et al. (1999) evaluate their segmentation algorithm using a ray tracer generated artificial sequences. Despite the best effort to make the scenes look natural the scenes are too regular compared to real scenes which is reflected in over-optimistic results of the segmentation.

2.7.3.2 Real-World Sequences

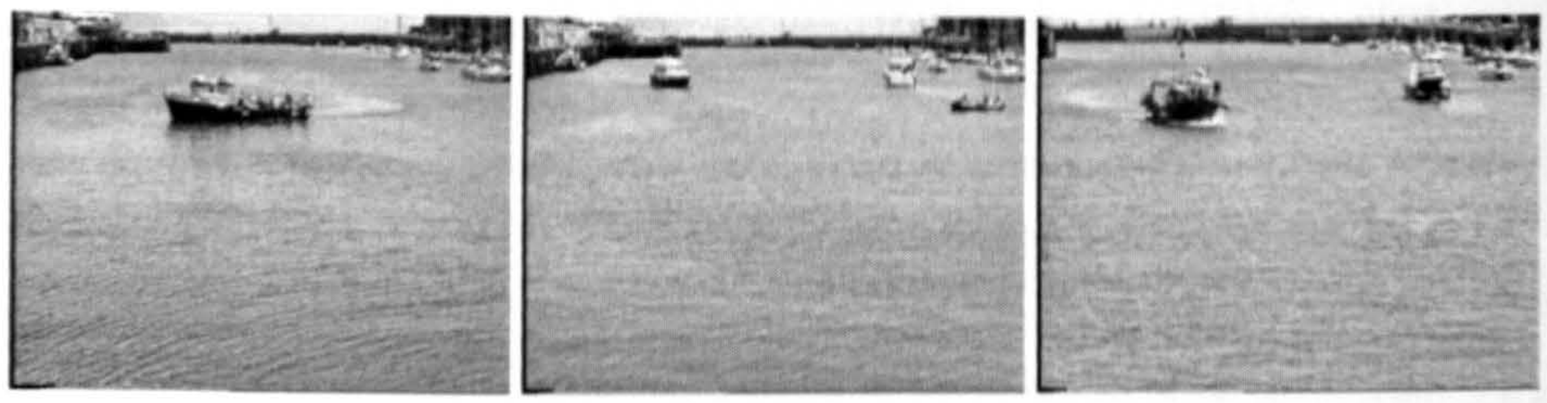
The development sequences were captured by an analogue off-the-shelf Panasonic NV-M10 camera. The focal length of the camera was set to two values - 842 and 940 pixels. These values were obtained by an off-line camera calibration detailed in Section 8.3.3.

Captured sequences were digitised using general purpose Fast Multimedia AVMaster v1.3 frame grabber. In order to retain maximum detail in the image, the highest resolution of 768×576 pixels provided by the frame grabber was used. The colour was quantised at 8 bits per colour component. Digitised colour frames were converted to intensity frames by averaging red, green and blue components at every pixel, (Morris, 2004), pp. 34-35, and normalising the result to integer values between 0 and 255.

The frame rates of the sequences are either 12.5 or 5 frames per second (fps). These values correspond to a half and a fifth of the standard 25 fps frame rate stipulated by PAL TV norm. These frame rates are sufficient for detection of all moving objects present in the scenes. Equations 2.19 and 2.20 are used to validate whether the chosen frame rates are realistic. The validation is done by determining the projected displacements for $v_x = v_y = 1\,knot$, a fixed detection range $R = 10\,m$ and a given frame rate. Excessively large displacement values

Sequence	Length [frames]	q_r [fps]	H [m]	ω [°]	f_{pix} [pix]	Δx_n [pix/knot]	Δy_n
2A	1045	12.5	4.5	2	842	6.7	3
2D	1418	12.5	4.5	2	842	6.7	3
2E	918	12.5	4.5	2	940	7.5	3.4

Table 2.2: Settings for Weymouth sequences.



(a) 2A (frame 574) (b) 2D (frame 496) (c) 2E (frame 627)

Figure 2.13: Sample frames from Weymouth development sequences.

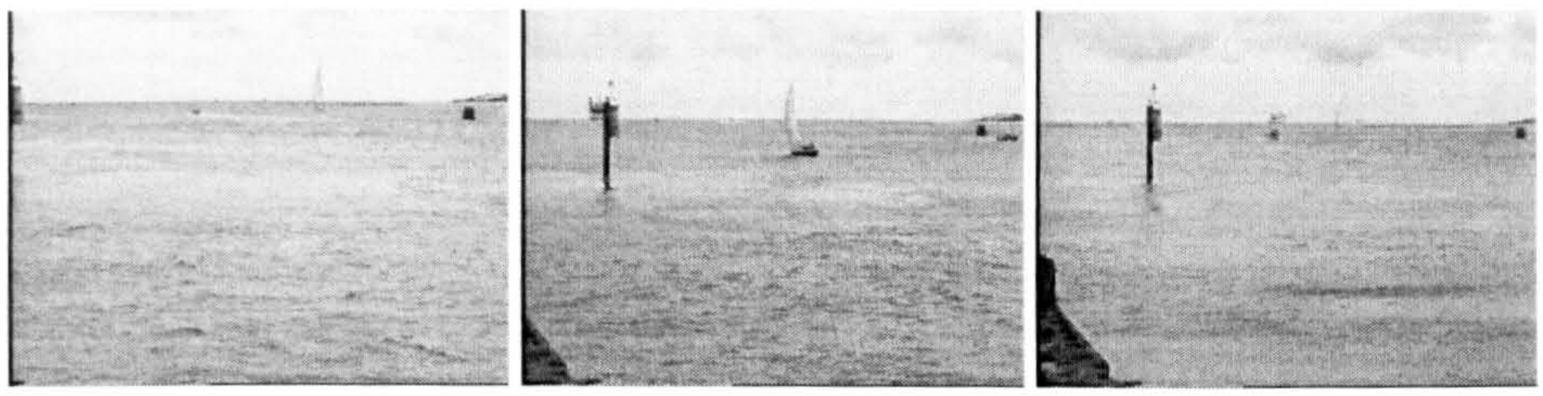
would indicate that the selected frame rate is too low for given velocity and detection range requirements.

Weymouth Sequences The first set of sequences was acquired in Weymouth (Dorset, UK) overlooking the harbour entrance during a sunny and calm day. There are multiple vessels entering and leaving the harbour while some of them are maneuvering along the way. A small ferry crosses the middle of the scene at regular intervals. A top left part of the scene contains a pier which represents a static structure. The pier is about 100 metres from the camera. There is also a group of mooring vessels on the right side of the scene. These sequences illustrate scenarios of a threat and collision as many of the objects move straight towards the camera. Table 2.2 shows the settings at which the sequences were obtained. Sample frames from each sequence are shown in Figure 2.13.

Sandbanks Sequences The second set of sequences was acquired at Sandbanks (Poole, Dorset, UK) near chain-ferry crossing on a windy and overcast day. The sequences contain multiple craft moving mostly from side to side of the sequence. The appearances and motion characteristics of objects vary throughout the sequence. There is a channel marker buoy visible on the right

Sequence	Length [frames]	q_r [fps]	H [m]	ω [°]	f_{pix} [pix]	Δx_n [pix/knot]	Δy_n
2M	1352	5	3	0	940	9.4	2.8
2Q	517	5	3	0	940	9.4	2.8
2R	245	5	3	0	940	9.4	2.8

Table 2.3: Settings for Sandbanks sequences.



(a) 2M (frame 800) (b) 2Q (frame 374) (c) 2R (frame 112)

Figure 2.14: Sample frames from Sandbanks development sequences.

in the scene. The camera monitors an open sea and it is positioned about 3 metres above the water. Table 2.3 shows the settings at which the sequences were obtained. Sample frames from each sequence are shown in Figure 2.14.

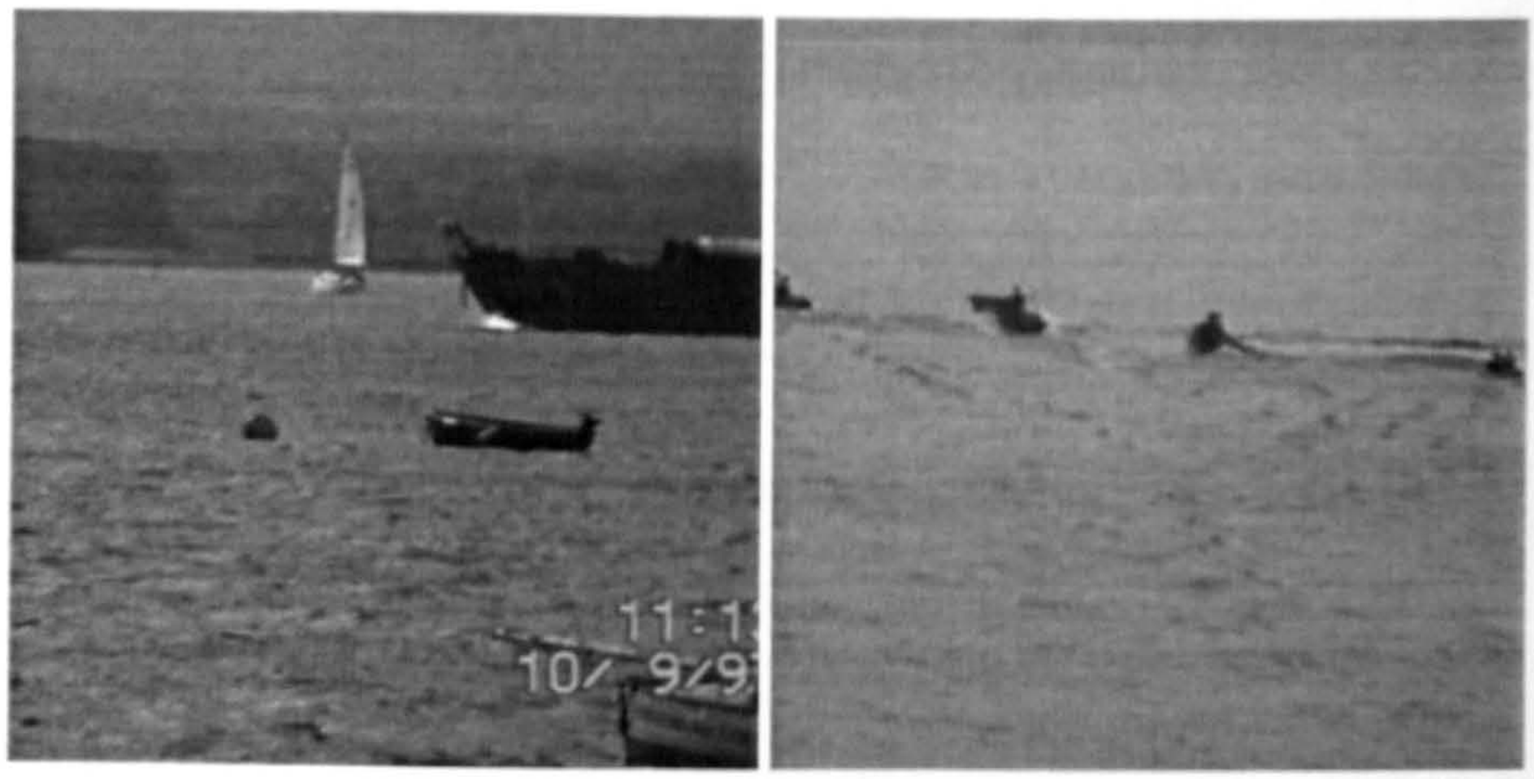
Poole and Portsmouth Sequences These two sequences were obtained in Poole (Dorset, UK) and Portsmouth (Hampshire, UK). The sequences are used only in a development of the segmentation algorithms due to unknown settings. The POOLEHARBOUR1 sequence shows three objects mooring near the observation point. A distant yacht sails slowly towards the left side of the image. A large dark vessel crosses the scene from right to left near the horizon. The frame size is 512×512 pixels. The sequence is 627 frames long. The frame rate is unknown.

The PORTSMOUTH5 sequence shows a group of small dark speedboats moving in a formation on a calm sea from left to right. The size of the frame is 512×512 pixels. The sequence is 1200 frames long.The frame rate is unknown.

Sample frames from both sequences are shown in Figure 2.15.

2.7.3.3 Artificial sequences

A ground truth is necessary in evaluation of many design steps. Because of the complexity in obtaining the ground-truth from real maritime sequences



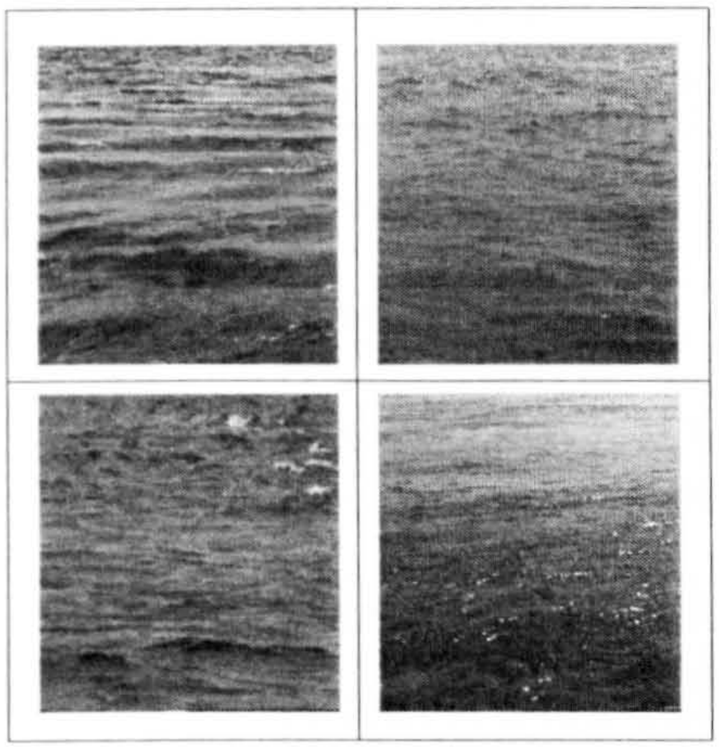
(a) POOLEHARBOUR1 (frame 150) (b) PORSTSMOUTH5 (frame 451)

Figure 2.15: Sample frames from POOLEHARBOUR1 and PORTSMOUTH5 development sequences.

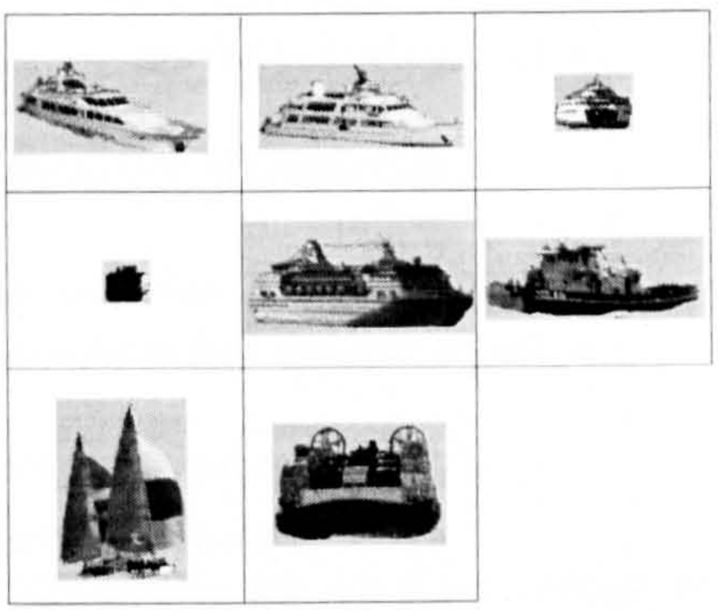
artificial sequences are generated by controlled superimposing. Sequences containing the sea at various states were acquired under conditions similar to the real-world sequences. Images of various maritime craft were acquired by 'cutting out' the silhouettes of the objects. The ground truth sequences were then obtained by superimposing the object images onto the sea frames at locations given by a precise mathematical model. The resulting frames are blended with Gaussian noise to emulate noise generated during the process of image acquisition. Figures 2.16a,b show sample background frames and images of maritime craft used in production of artificial sequences. Figure 2.16c shows a sample artificial frame.

2.7.4 Evaluations

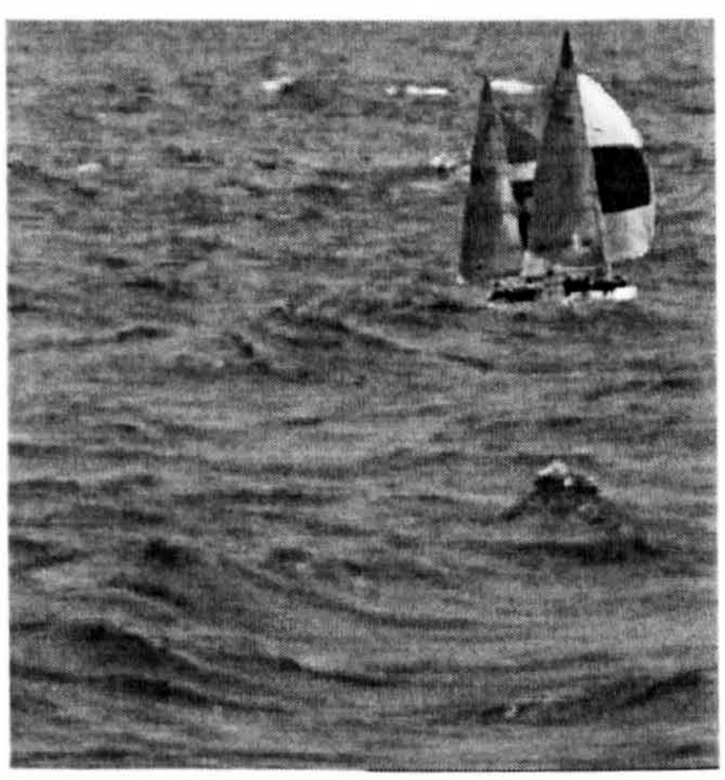
Numerous experiments conducted during the research and framework design presented here use either relative or goal-driven evaluations. The relative evaluation is done in cases where the ground truth cannot be obtained. The goal-driven evaluation is used in cases where the ground truth is available. A final cross-validation of the whole framework is conducted by processing previously unseen maritime sequences in order to test whether the attributes of development scenes did not bias the performance of the framework.



(a) Backgrounds



(b) Objects



(c) Artificial frame

Figure 2.16: Backgrounds (a) and objects (b) used in production of artificial sequences. A sample artificial frame (c).

2.7.4.1 Relative Evaluation

Relative evaluation is used in experiments where ground truth for data is absent. The evaluation consists of the following steps:

- sample data that represent typical and exceptional cases are collected,
- evaluation criteria based on relative quantifications are defined,
- for method selection,
  - all methods evaluated in the selection are applied to the sample data,
- for parameter value adjustment,
  - a set of values that methodically covers the whole numerical range of the parameter is determined,
  - the sample data are successively processed with the evaluated parameter set to each value from the set,
  - relative indications that quantify the performance of the method or the influence of the parameter value are determined for the results,
- the method or parameter value that maximises or minimises the evaluation criteria is selected.

Minimisation of variance of the results for samples from the same group or maximisation of variance between results for samples from two different groups are two examples of evaluation criteria. Inclusion of exceptional cases in sample data allows to investigate stability and robustness of the evaluated subjects.

The main drawback of relative evaluation is that it provides only relative assessment of the performance. The performance of the best candidate cannot be quantified in absolute terms as the ground truth is absent. Relative evaluation is therefore used only for selection of methods and parameter values that are not crucial to the outcome of the processing.

2.7.4.2 Goal-driven Evaluation

Goal-driven evaluation is applied when ground truth is available. The process is similar to the relative evaluation. The difference is that the ground truth

specifies the goal to be reached by the evaluated subjects. The evaluation criteria is based on minimisation of discrepancy between the obtained results and the goal. The evaluation consists of the following steps:

- sample data that represent typical and exceptional cases are collected together with the ground truth,
- evaluation criteria based on minimisation of discrepancy between results and the goal are established,
- for method selection,
  - all methods evaluated in the selection are applied to the sample data,
- for parameter value adjustment,
  - a set of values that methodically covers the whole numerical range of the parameter is determined,
  - the sample data are successively processed with the evaluated parameter set to each value from the set,
- a function that quantifies the discrepancy between the results and ground truth is evaluated for all results,
- the method or parameter value that minimise the discrepancy criteria are selected.

An example of evaluation criterion is minimisation of average distance between detected and actual geometric features. The evaluation allows to select the best candidate and also to quantify its performance with respect to the ground truth. The uncertainty associated with the selection of the best candidate can be determined from the value of the criterion function. For example, uncertainty associated with a selection of a specific corner detector can be expressed as an average error of corner localisation.

2.7.4.3 Cross-Validation

A realistic and manageable data sample used in experiments typically represents only a fraction of the whole problem domain. Experimental results can become biased due to the limited magnitude of data samples. The bias

could have an adverse impact on the performance of the framework. For example, the framework could be less reliable when detecting dark objects if the development sequences contained only bright objects. Most such issues are avoided by careful sampling and experimental design. Nevertheless, a suitable cross-validation is still essential.

The cross-validation of the proposed framework is conducted by presenting it with maritime sequences previously unused for the development. The sequences were obtained at conditions different to those for the development sequences, namely locations, weather conditions, object types and geometric setup. The performance of the whole framework is evaluated in various categories including detection of objects, motion estimation and identification of threats. A robust framework should provide results consistent with those obtained during its development.

2.8 Summary

A detailed contextual analysis of the problem domain is conducted prior to the specification of research objectives. The analysis looks at optical, geometric and dynamic contexts of maritime scenes.

The analysis of optical context concludes that appearances of the maritime scenes and objects are generally varying to such an extent that any modelling based purely on appearance attributes would be complex.

Geometric context analysis derives a scene projection model from the general pinhole camera projection. The model complies with a plane-to-plane projection where the first plane corresponds to the sea surface and the other plane corresponds to the image. The objects are assumed to obey the Ground Plane Constraint, i.e. they are all located at the level of the sea surface. A planar-based representation of objects in the scene is inferred from a weak perspective projection. The weak perspective model assumes that the depth of an object is significantly smaller than its distance from the camera. All possible deviations from the optimal geometric models are identified as either systematic or stochastic errors contributing to the overall uncertainty of the processing outcome. Mathematical models of errors are provided.

As a part of dynamic context analysis, the relation between an object projection displacement and the speed and location of the object in the scene is obtained. The relation provides a constraint to the solution in terms of

maximum detectable speed of objects at a given frame rate.

Research objectives based on the conclusions of contextual analysis are outlined together with constraints imposed on the problem domain. The research is based on analysis and evaluation of real-world video sequences that provide a representative sample of the problem domain. The problem is split into sub-tasks that are solved individually. The sub-tasks correspond to various components in the architecture of the proposed vision-based framework. Such a structure adheres to the bottom-up architecture which is an adequate solution for the type of problem addressed by this thesis.

The sequences used in the development have been acquired at conditions similar to those of the intended applications of the framework. The sequences include various maritime scenes, objects and activities assumed to be typical samples of the maritime domain.

Three categories of evaluations used in the development are introduced - relative, goal-driven and cross-validation. Relative evaluation is applicable in experiments where ground truth is absent. Goal-driven evaluation uses the ground truth as criterion. Cross-validation is a final evaluation of the assembled framework that checks whether the obtained solution is robust and unbiased.

Chapter 3

Literature Review

Three main areas of research and technology are reviewed in the following Chapter that are relevant to the subject of the thesis. These are:

- *Vision-based technology in maritime navigation and surveillance.* Vision-based applications such as night-vision systems used nowadays to aid maritime navigation and surveillance are described and discussed.

- *Land-based surveillance and tracking applications.* Common approaches to the problem of detection and tracking of objects in land-based applications are categorised and discussed. Their limitations to the detection of objects in maritime scenes are pointed out. Two frameworks, VSAM (Collins et al., 2000) and ASSET-2 (Smith, 1998), are described in details.

- *Image processing in maritime sector.* Recent research studies and works concerned with image processing in maritime sector are reviewed. These works are categorised with respect to the type of imagery they utilise - infra-red and visible range.

3.1 Vision-based Technology in Maritime Navigation and Surveillance

Most vessels, ports and harbours are equipped with numerous navigation aiding devices that facilitate the complex task of secure maritime navigation. These are mainly radars operating at multiple wavelengths for detection of

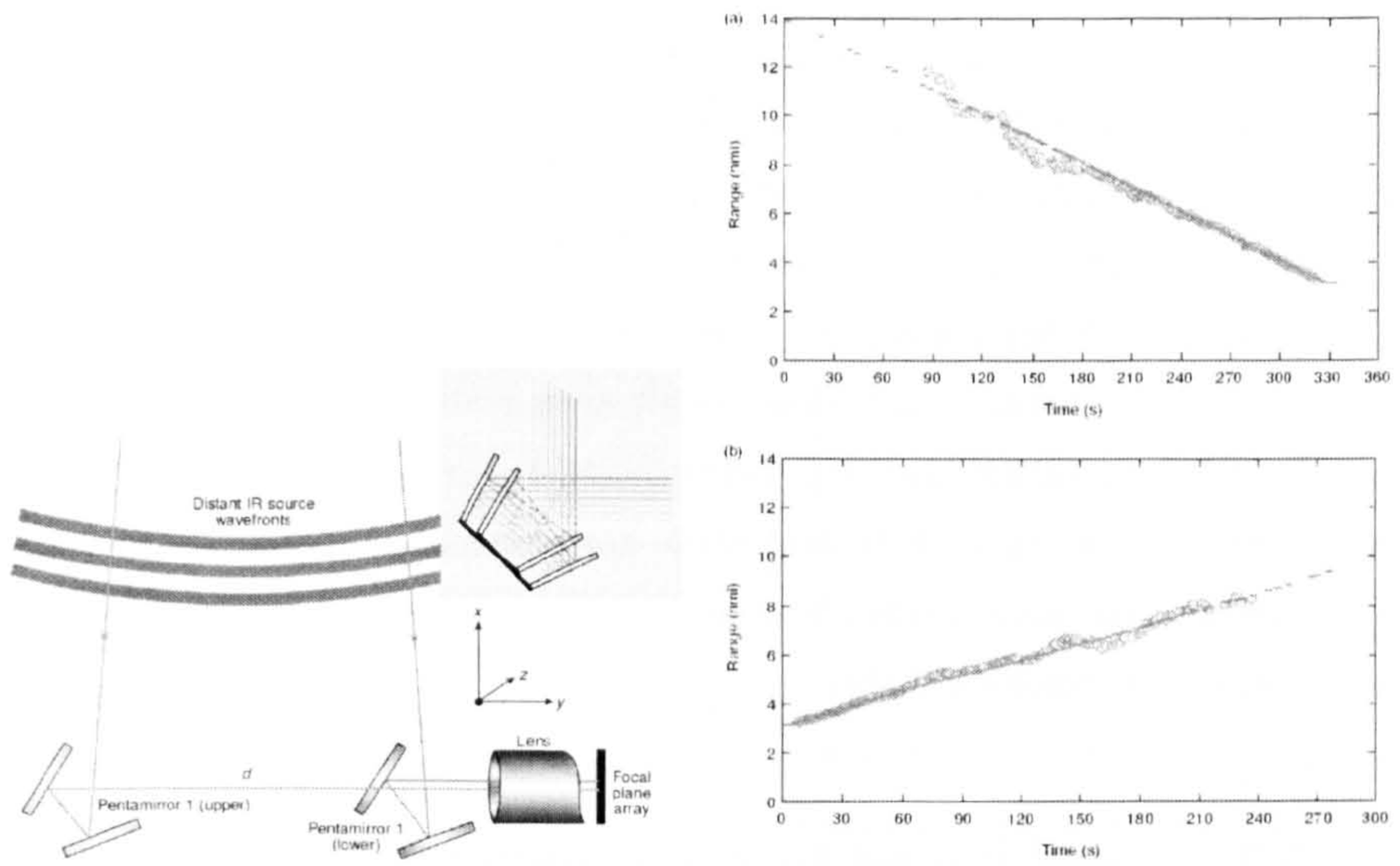
(a) marinised camera (b) operator's console (c) image provided by the system

Figure 3.1: Components (a),(b) of the Vistar IM405 Multi-Sensor Surveillance system (Vistar Night Vision Limited, 2004c). The image of the scene (c) provided by the image intensifying sensor of the system.

objects, the Global Positioning System (GPS) for precise localisation of detected objects, Automatic Identification System (AIS) for identification of objects, VHF radio links for communications and electronic maps and charts for navigation, (Nera GmbH, 2004; Raymarine Limited, 2004; Raytheon Marine GmbH, 2001). In addition, light intensifying (Vistar Night Vision Limited, 2004a; Turn Ltd., 2001) and infra-red cameras (Vector Developments Ltd., 2004; Vistar Night Vision Limited, 2004b) are often used as additional devices assisting the navigation during night time or in a bad weather.

A typical maritime vision system consists of one or more cameras in weather-proof (marinised) casing mounted on an articulated point of the vessel structure. An analog output from the camera is connected to a control panel with monitor on the bridge. The cameras provide the operator with either infra-red or light intensified images, (Vistar Night Vision Limited, 2004c). The operator can control the pitch, yaw and zoom of the camera from the cabin as well as essential settings of intensity or contrast of the image. Some infra-red based maritime vision systems (The Current Sales Corp., 2004) are equipped with narrow beam infra-red reflectors that illuminate the scene and objects under observation improving the contrast of the acquired image. Figure 3.1 shows the components and output image of a standard night vision system, (Vistar Night Vision Limited, 2004c).

Nevertheless, the system's functionality is limited to a simple provision of intensified or infra-red images from the outside camera to the monitor inside



(a) Optical design

(b) Range detection - results

Figure 3.2: Optical infra-red ranger described by Reilly et al. (1999). The optical design (a) utilises two pentamirrors that reflect all rays in 90° towards the single lens. Results (b) of two encounters of the ranger with a small aircraft. The circles correspond to the estimates provided by the ranger, dashes correspond to ranges detected by a radar.

the bridge. The system still requires the continuous attention of a human operator. The human operator has to constantly monitor and adjust the system parameters such as camera position, zoom or image quality in order to reliably detect any objects in the scene. So far, there is no commercial application capable of automated detection and tracking of objects in sequences obtained by these maritime vision systems.

A notable exception is a passive infra-red ranger presented by Reilly et al. (1999). The system employs the principle of stereopsis in order to detect the range of objects using their infra-red signature. High precision of the system is achieved by a novel optical design shown in Figure 3.2a that employs a single camera and a set of so-called 'pentamirrors' that reflect all incoming rays in 90° angles. Both mirrors reflect the incoming rays towards the lens of the camera. A single image contains both projections of the object. The range is determined from the displacement of the two projections.

The experimental results shown in Figures 3.2b indicate that the error in the

estimation of the range of the objects increases linearly with the range and at 20 nautical miles the error is below 10%. The major advantage of the system is that it is a passive sensor resistant to detection by active sensor detectors used in the military sector. It , however, demands a high precision optical design that is costly. Also, the issue of protection of the device against harsh weather conditions at the sea remained unresolved at the time of publishing.

The situation in maritime sector is in great contrast with the situation in land-based surveillance applications where automation of the detection and identification of objects and their activities is a well established subject of extensive research with successful commercial applications already emerging, (Dick and Brooks, 2003).

3.2 Land-based Surveillance and Tracking Applications

A typical land-based surveillance and tracking application consists of one or more cameras mounted on a static or moving platform overlooking the scene. Depending on the purpose of the application, objects that are either static or moving are detected and located in the scene. Their motions and activities with respect to the rest of the scene are identified and assessed. The applications can be roughly divided into two categories with respect to the dynamics of the camera, (Morris, 2004).

3.2.1 Static Camera Moving Objects

The first category can be characterised as a 'static camera, moving objects' (SCMO) problem. The principal assumption is that subjects of processing in these applications usually undergo permanent or transitory motion while the background remains either static or slowly changing, (Lee and Hedley, 2002; Tornieri et al., 2003), its change is due to the noise that can be either suppressed (Rosin and Ellis, 1995), or statistically modelled (Elgammal et al., 2002; Magee, 2004). The assumption enables moving objects to be detected as localised changes between two consequent frames or between current frame and the reference frame representing the background (Lee and Hedley, 2002; Collins et al., 2000), or as outliers to the background and noise models (Elgammal et al., 2002; Magee, 2004).

Detected objects are tracked by finding correspondences of their signatures in consequent frames of the sequence. Signatures uniquely characterise the objects in terms of location, geometry and other attributes such as intensity or colour. Typical signatures include coordinates of the centre of gravity of the blob representing the detected object (Fuentes and Velastin, 2001), coordinates of the rectangle bounding the blob (Black and Ellis, 2002), a rectangular template including the object projection (Collins et al., 2000). More sophisticated signatures involve statistical distributions of intensity or colour and location of the objects (Sheikh et al., 2004), 2D curves (Tai et al., 2004) or 3D wire-frame models (Worral et al., 1995; Remagnino et al., 1997) fitted to the object projection. The correspondence search is based on minimisation of a specific error function that quantifies the difference between candidates for correspondence. An example of such a quantification is the correlation coefficient in a template-based tracking (Collins et al., 2000). Kalman filtering is often employed in order to support reliable tracking (Tai et al., 2004; Magee, 2004; Dellaert and Thorpe, 1997).

Examples of typical SCMO applications include monitoring of public areas such as metropolitan undergrounds (Cupillard et al., 2003), car parks (Micheloni and Foresti, 2003), railway crossings (Sheikh et al., 2004), highway traffic surveillance (Remagnino et al., 1997; Tai et al., 2004; Worrall et al., 1994), etc. A substantial level of the automation has been already achieved in these applications including analysis of behaviour of individuals and groups of people (Cupillard et al., 2003), classification of interactions between individuals and other static or moving objects in the scene (Collins et al., 2000; Micheloni and Foresti, 2003; Haritaoglu et al., 2000), estimation of various attributes of traffic scenes such as density, jam developing, detection of stalled vehicles, (Tai et al., 2004; Smith, 1998).

3.2.2 Moving Camera Moving Objects

A generalisation to SCMO problem can be characterised as 'moving camera, moving objects' (MCMO) problem. A common assumption in MCMO algorithms is that the structures and appearances of the background and moving objects in the scene do not change substantially between the reference and current frames in the sequence and that the inter-frame displacements in the scene are limited, (Barron et al., 1994; Lipton et al., 1998; Lipton, 1999; McCane et al., 2002; Galvin et al., 1999a; Galvin et al., 1999b; Smith,

1998). The methods addressing MCMO problems are often based on estimation of an optical flow in the image which is a 2D projection of the 3D motion in the scene. The numerous methods of optical flow estimations available are overviewed by Beauchemin and Barron (1995). Their choice depends on the purpose of the application. An extensive area of the research is dedicated to a so-called 'structure-from-motion' problem of determining the 3D structure of the objects in the scene from the 2D motion in the image which is an ill-conditioned task (Shapiro, 1995; Torr and Murray, 1993; Torr, 1998).

The MCMO surveillance and tracking applications can be typically found in autonomous navigation of vehicles (Dellaert and Thorpe, 1997; Kastrinaki et al., 2003; Broggi, 1995), detection of motion from an airborne platform (Cohen and Medioni, 1998), image stabilisation (Irani et al., 1994) and others.

ASSET-2

Smith (1998) presents a typical example of an MCMO application that estimates the optical flow in sequence using correspondences between detected geometrical features (see Figure 3.3).

Two-dimensional features using either the SUSAN corner detector (Smith and Brady, 1995) or the Harris corner detector (Harris and Stephens, 1988) are detected in each frame of the video sequence. The features are matched across the frames using a similarity measure based on properties of the detected corners such as intensity at the corner location. This is in contrast with traditional matching methods which use correlation of small image patches located at the corners. Smith (1998) argues that there is a little justification for such an approach as usually less than a half of the patch area covers object's structure, the rest of the area containing the changeable background of the scene. They show that their alternative scheme drops the amount of correct matches only by 10% (from the original 95%) while significantly reducing the computational overhead that is associated with correlation methods, (Lewis, 1995).

A constant velocity motion model is initiated for every pair of corners matched in the first two frames of a sequence. The search for the consequent matches is simplified as the projected position of each corner is calculated from the motion model. A list of flow vectors is estimated from the matched corners that characterises an optical flow in the scene.

The list of flow is segmented into separate clusters by fitting an affine

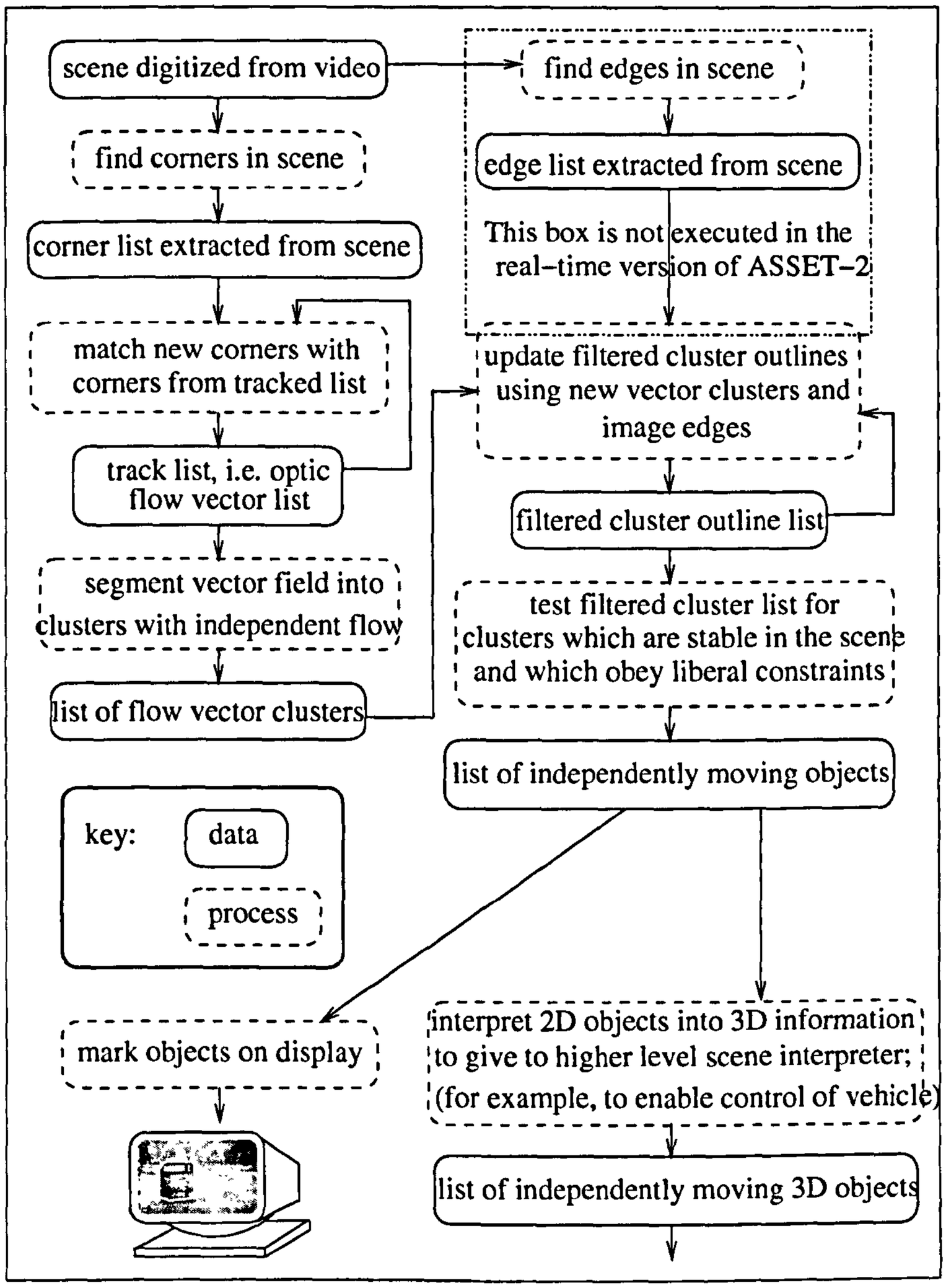
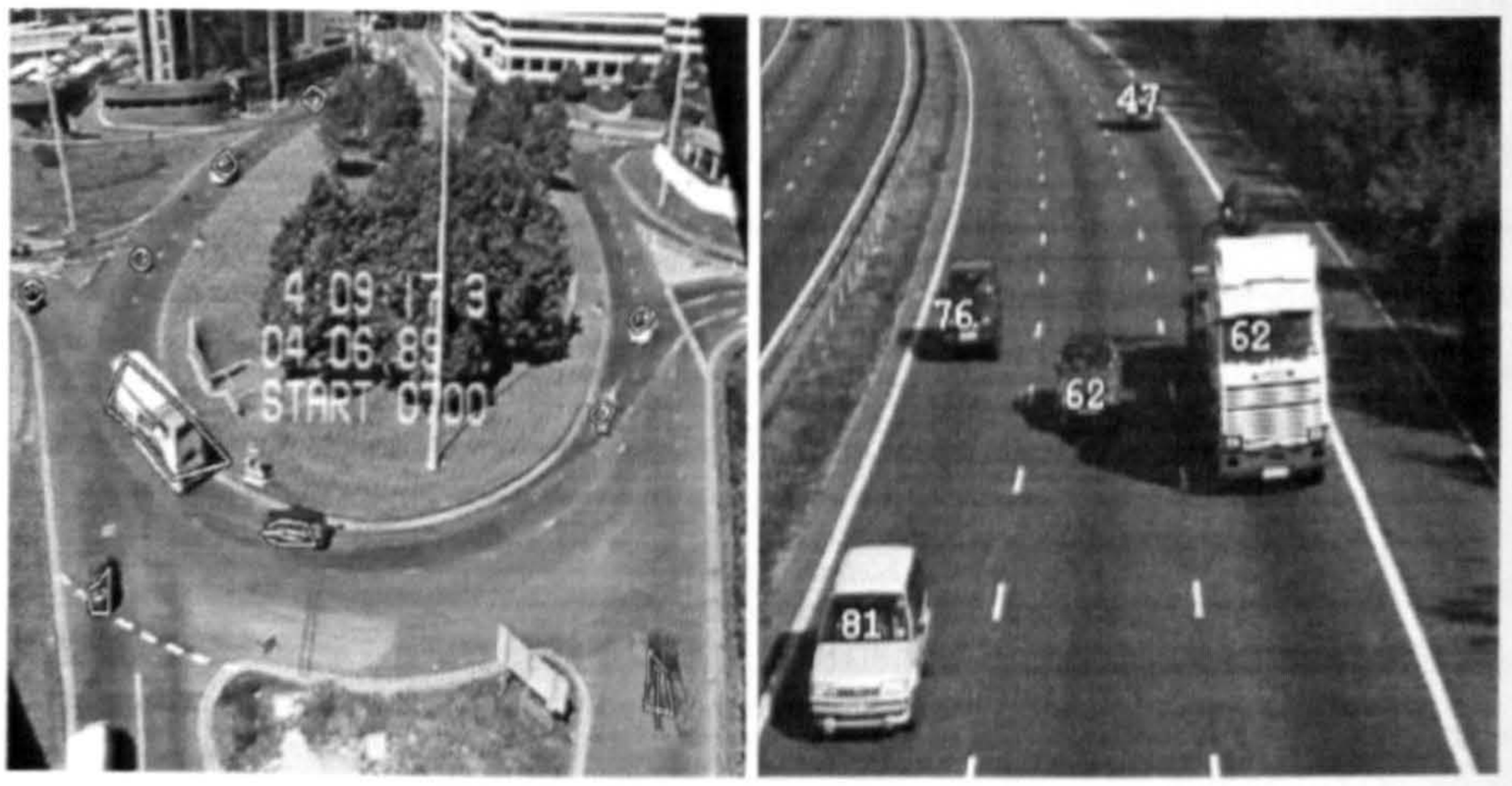


Figure 3.3: The structure of the ASSET-2 tracker proposed by Smith (1998)



(a) traffic surveillance (b) estimation of velocities

Figure 3.4: The results of traffic surveillance by the ASSET-2 tracking system: (a) - detection and motion estimation of objects in a typical traffic scene; (b) - estimation of velocities of vehicles on a highway.

motion model to sets of displacement vectors by minimum spanning tree method also used by Shapiro (1995). Each independent cluster is assigned a centroid and a boundary. The clusters are matched across the sequence using a time-symmetric matching, i.e. both, previous and current clusters must prefer the proposed candidate. The matching is based on the motion parameters and the contours of the clusters. The contour of the cluster is encoded using a convex radial map which is a concept similar to active contours. A contour enhancement method that iteratively shifts the cluster contour towards detected edges is also proposed. Matched clusters are tracked by a simplified Kalman filter.

Contour tracking enables to resolve two types of occlusions - object by object and object by background. The first one is based on the amount of overlap between two clusters and an assumption that occluded object is higher in the scene. The second occlusion is detected by a rapid change in numbers of features belonging to the cluster.

The results in (Smith, 1998) show the ability of the ASSET-2 to track objects in both SCMO and MCMO problems, to deal with occlusions between objects and between object and background. In addition, a traffic surveillance application is presented where the system overlooks a highway and estimates

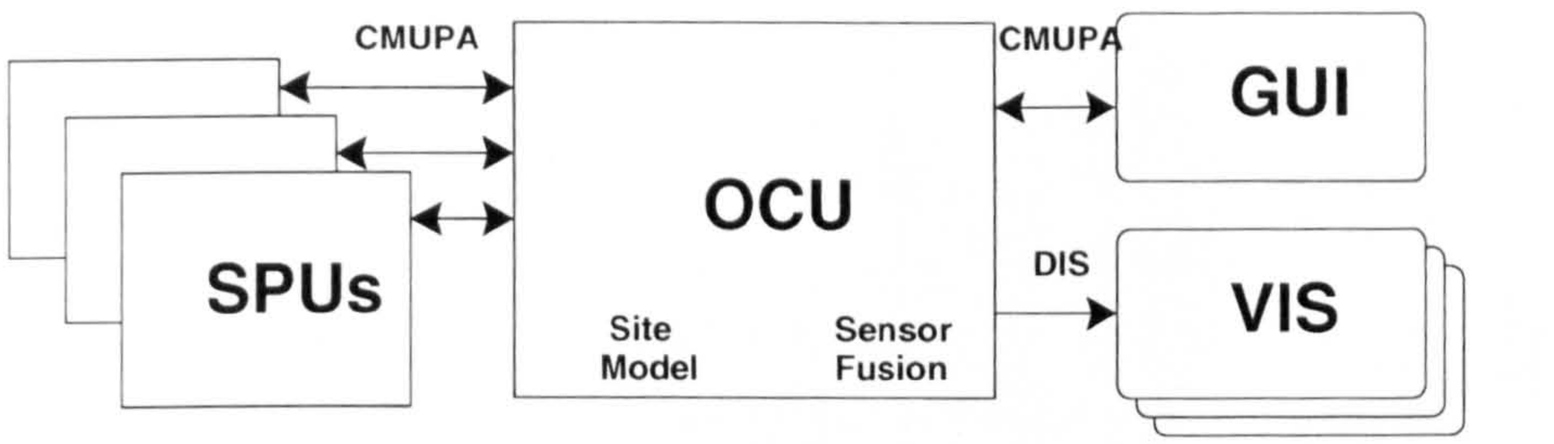


Figure 3.5: The structure of the surveillance, tracking and monitoring testbed VSAM at Carnegie Mellon University, (Collins et al., 2000)

velocities for an oncoming traffic (see Figure 3.4).

VSAM

VSAM (Video Surveillance and Monitoring) project at Carnegie Mellon University (Collins et al., 2000) is a complex testbed for an outdoor surveillance, monitoring and visualisation of objects including cars and people and their activities and interactions to support battlefield awareness. The system consists of an extensible hierarchical architecture shown in Figure 3.5 that can control and process data from multiple distributed sensors such as monochrome, colour and infra-red cameras. The sensors can be static, panning, tilting and zooming, omnidirectional, mobile or airborne. Geo-locations of the detected targets are obtained by combining processed visual data with detailed digital maps of surveyed estates. Targets such as people, cars and interactions between them are recognised and logged into a database. Techniques such as multi-sensor tracking, occlusion detection, target classification into multiple categories according to appearance and dynamic behaviours are all integral parts of the system.

Three algorithms for moving object detection are used in the VSAM. The first is used with static sensors and is a combination of a three-frame differencing and an adaptive background subtraction. The motion is first detected from the frame differencing, the background subtraction is used to fill

in the missing pixels that belong to the moving object. A layered representation of objects in the scene and analysis of pixel dynamics enables the system to identify occlusions and objects that suddenly stop.

The second algorithm is a background subtraction method modified for pan and tilt cameras. A complete set of all background images for varying camera position settings is obtained and stored. New images are registered to the nearest background image using salient features. The registered image is processed the same way as an image from a static camera.

The third algorithm is used in airborne surveillance where a compensation for motion of the camera is necessary. The incoming frames are warped to an initial image which is updated at regular time intervals. The objects are then detected in the warped image by means of the first algorithm.

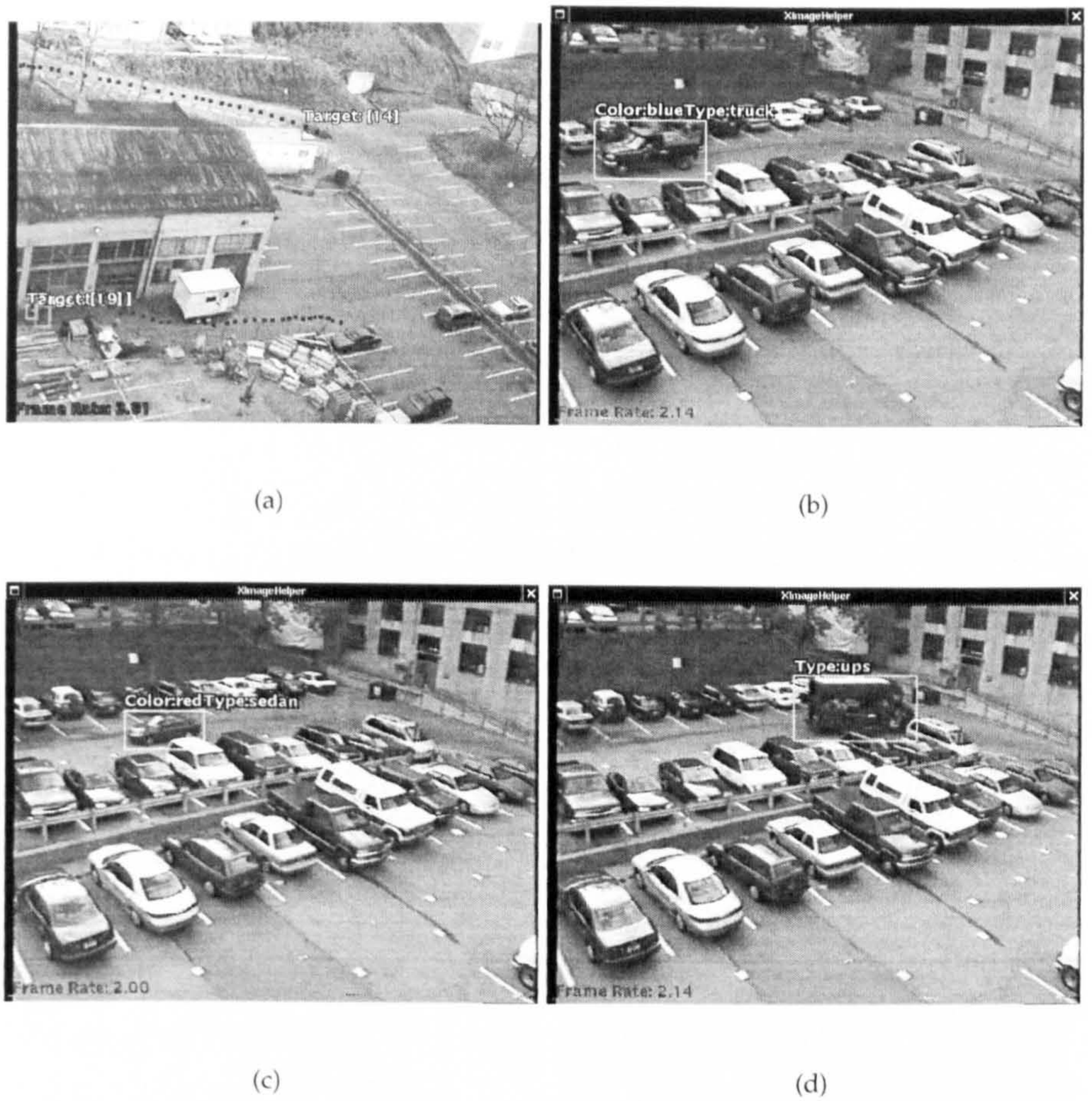
An object detected by any of the three algorithms is matched in the following frame by a weighted correlation of an image patch containing the object. The correlation weights are generated by a linear radial function with centre located at the centre of the patch. A new location of the object is used in estimation of position and velocity. A hypothesis tracking with confidence values enables it to resolve ambiguous cases of objects grouping or parting.

Tracked objects are then classified by a neural network based on geometrical features such as dimensions and rigidity into humans, human groups and cars. Furthermore, cars are classified into various groups such as trucks, sedans, vans, etc. Finally, the interactions between objects are recognised by a gait analysis and Markov models.

The VSAM functionality and performance in various surveillance tasks are well-documented through video sequences available at the project website, (Collins et al., 2000). Figure 3.6 shows sample results of tracking and target classification performed by the VSAM.

3.3 Image Processing in Maritime Sector

The challenge of object detection and tracking in maritime scenes lies in the fact that the background of maritime scenes represented by the sea does not obey the assumption of being static that is essential to methods typically used in land-based SCMO systems. In addition, the background of maritime scenes does not satisfy any of the three conditions required for calculation of optical flow in MCMO applications stated by Beauchemin and Barron (1995):



(a) (b)

(c) (d)

Figure 3.6: The tracking and classification results provided by the VSAM monitoring the warehouse entrance and the parking lot

the illumination of the maritime scene is often directional and, therefore, not uniform, the reflectance of the sea is specular and, therefore, not Lambertian and the motion of the sea is not pure translation parallel to the image plane.

This is reflected by the fact that majority of the image processing research in maritime environment concentrates on a primary task of spatial segmentation of the scene. The complexity of the task is illustrated by the diversity of the methods proposed. The methods address very specific, substantially constrained problems that do not go beyond basic scene segmentation.

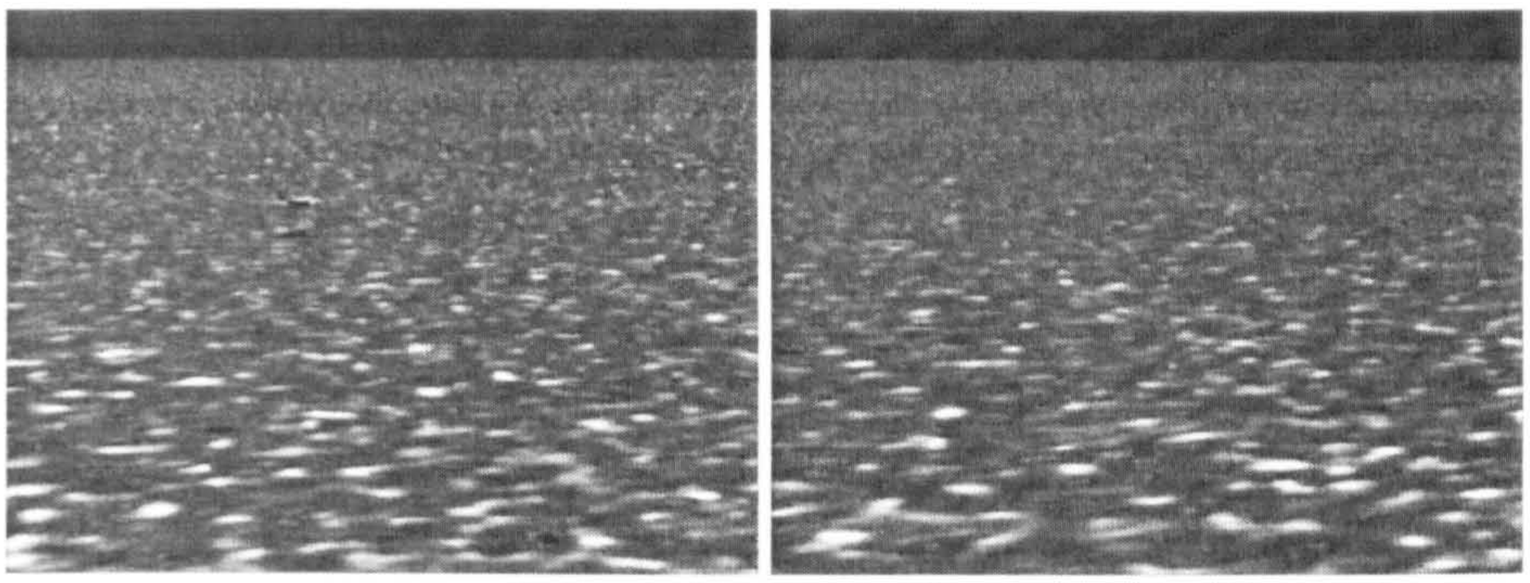
3.3.1 Infrared Images

The majority of the image processing methods deal with the segmentation of infra-red sequences due to the fact that most maritime night vision systems are based on the infra-red sensory. Infra-red imaging is also deployed in military sector for detection of vessels, missiles and airborne targets. A selection of available segmentation techniques for detection of naval objects in infra-red images is presented and discussed in the following section.

Messer et al. (1999) propose a texture representation method for Automatic Target Recognition (ATR) applications in maritime environment. The method works by learning the statistical model of the background and targets in maritime scenes. Randomly sampled small patches in a training frame are rearranged to form vectors. A set of features characterising the vectors is obtained by either Principal Component or Independent Component Analyses (PCA, ICA). The PCA is preferred as it provides results similar to ICA while being simpler. A subset of features is selected that maximises the distance between target and background representing vectors. Finally, a threshold is specified that separates the vectors representing the target from those representing the background. The classification of a new frame is done by segmenting the frame into patches of the same size as those used in training, obtaining the vectors, calculating the selected features and thresholding the results.

The method is applied to a ray-tracer generated artificial sequences of missiles and aeroplanes flying over the sea as shown in Figure 3.7. Authors claim that the method is able to detect single pixel targets. The method is also compared with segmentation based on wavelet transform. The results show that PCA outperforms wavelets by finding all targets in the testing sequence (see Figure 3.8 for the segmented sequence B). The authors, however, acknowledge that the rolling waves and the perspectiveness of the scene cause the target detection to oscillate in time. Despite the best effort, the ray-tracer generated scenes shown in Figure 3.7 are considerably regular which makes the results biased.

The authors extended the algorithm in their interrelated work (Messer and Kittler, 2000) by incorporating a temporal component into the PCA analysis. The vectors entering the PCA and feature selection are obtained by rearranging small 3D volumes that are randomly sampled across the spatial and temporal dimensions of the sequence. This integrates the dynamics of the



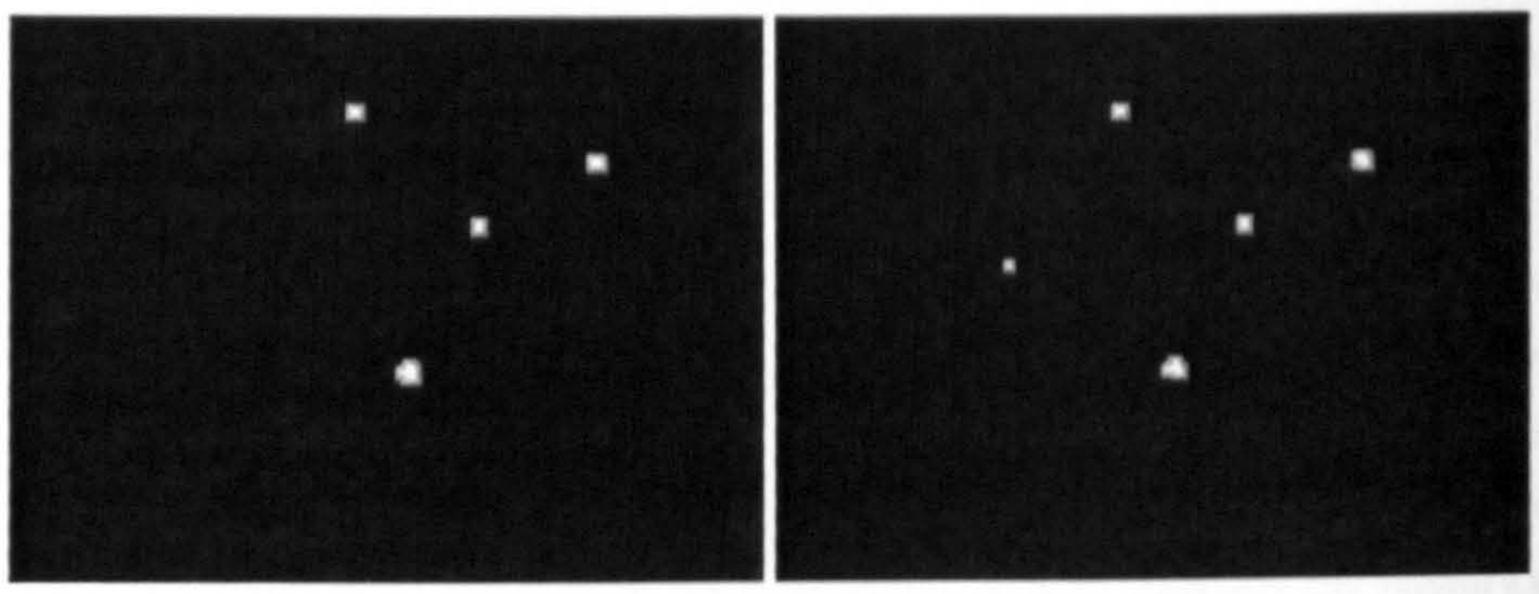
(a) sequence A - sample frame

(b) sequence B - sample frame

(c) sequence A - ground truth

(d) sequence B - ground truth

Figure 3.7: Artificial sequences used in evaluation of segmentation framework proposed by Messer et al. (1999)



(a) (b)

(c) (d)

Figure 3.8: The results of the segmentation by Messer et al. (1999) applied to the sample sequence B shown in Figure 3.7b,d

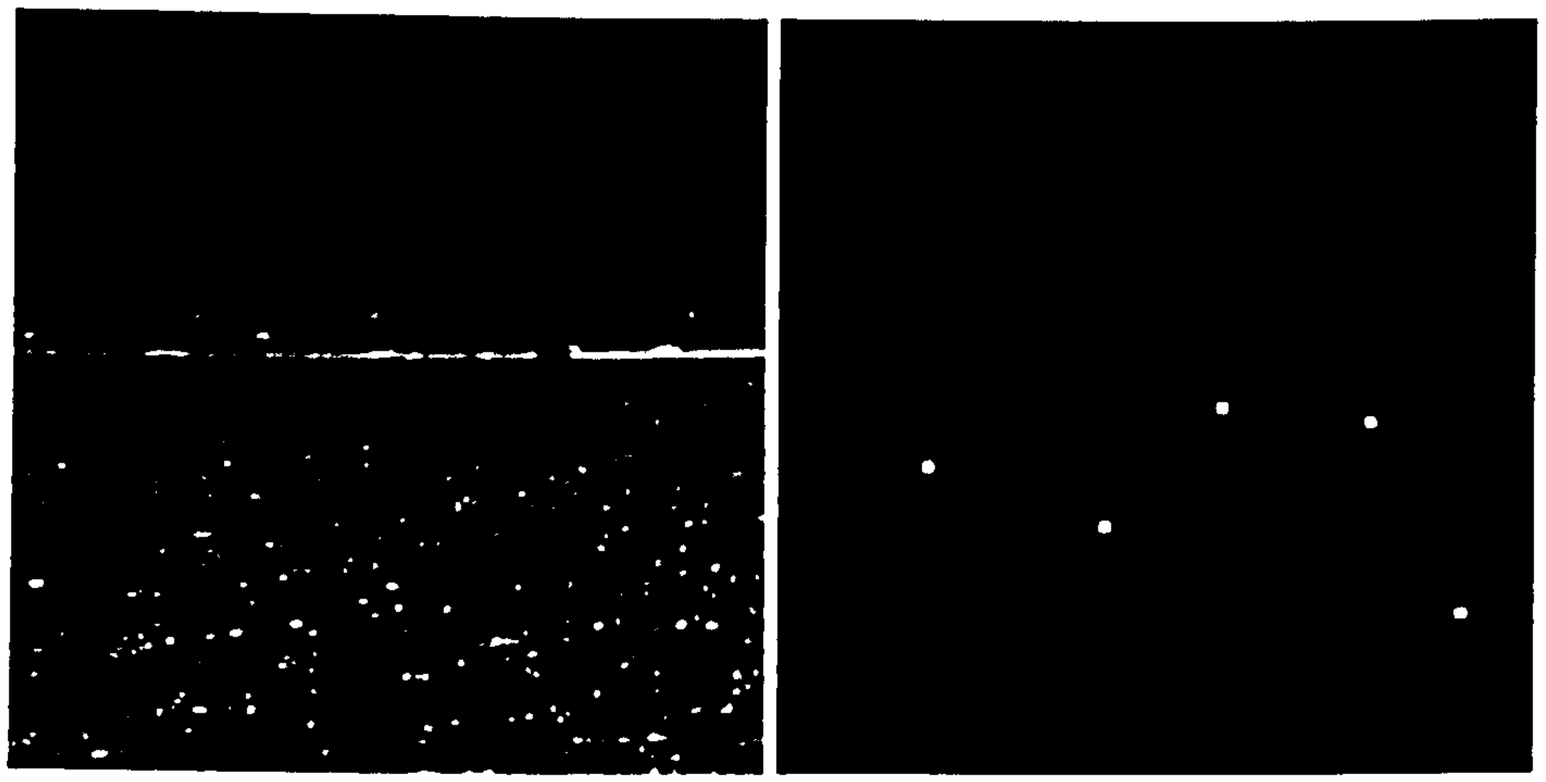
sea background into the statistical model.

The method proceeds by thresholding and morphologically dilating each frame in the sequence. A temporal averaging across five frames filters out short term noise while enhancing the target signature.

The modified method is tested on the same artificial sequence used in (Messer et al., 1999) and, additionally, on a real sequence with real targets shown in Figure 3.9. The extended method is compared with the original version. The evaluation criteria are average numbers of true and false positives in each sequence. When applied to the artificial sequence, the modified method reduces the average number of false positives from the original 18.06 to 4 while preserving the correct number of true positives. When applied to real sequence, the modified method again reduced the number of false positives from the original 14.29 to 4.86 while preserving the number of true positives. The results clearly indicate that the dynamics of the sea background is a significant factor contributing to the structure of the scene.

A similar method for background clutter removal in infra-red images based on PCA is proposed by Diani et al. (2003). They construct the vector space using the image columns instead of rectangular blocks in order to compensate for the horizontal striping noise typical for infra-red imagery. The method is also insensitive to the transition between sea and sky. The vectors of dimension M equal to image height span a vector space. The space is split into a subspace $\mathcal{U}$ of dimension M_B corresponding to the background structure and clutter and an orthogonal residual subspace $\bar{\mathcal{U}}$ of dimension $M - M_B$ containing random vectors with zero-mean Gaussian distribution. The optimal dimension M_B is determined iteratively by a χ^2 and correlation tests applied to the residual subspace $\bar{\mathcal{U}}$ which is assumed to contain uncorrelated random vectors with zero-mean Gaussian distribution. By projecting the image data onto the residual subspace $\bar{\mathcal{U}}$, the problem of target detection is reduced to a standard null hypothesis testing.

In real applications the target is often present in the image used in PCA. The authors investigate the influence of the target presence on the results of the PCA (so-called target leakage) and conclude that the method is feasible for detection of weak targets with a signal-to-clutter ratio below 24 dB. Stronger targets leak into the background structure and, therefore, adversely influence the separation of the vector space. Unfortunately, any rigorous numerical evaluation of the method is missing as only visual results for a single image



(a) Evaluation sequence

(b) Ground truth

(c) segmented sample frame

(d) segmented sample frame

Figure 3.9: The results of the segmentation by Messer and Kittler (2000) applied to the sample real sequence.

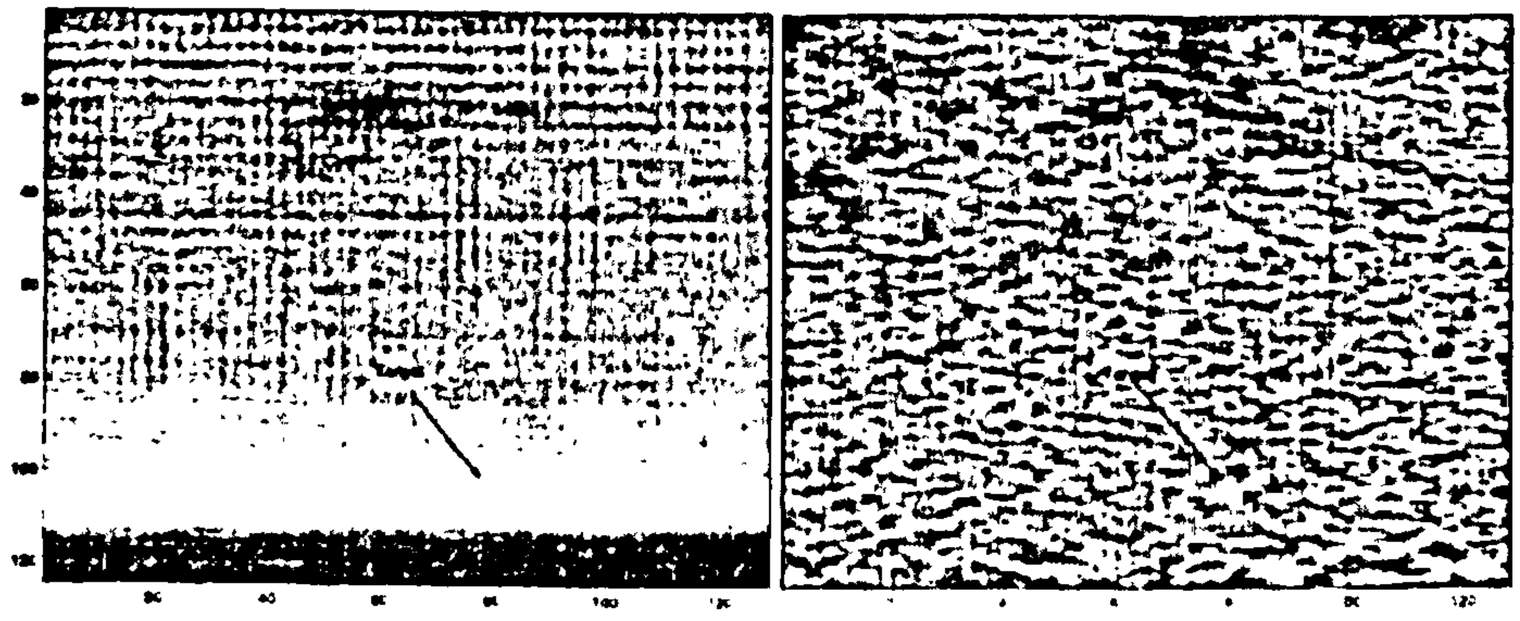
with a superimposed artificial target are provided as shown in Figure 3.10.

Toet (2002) provides a less complex approach to detection of small and dim objects in maritime scenes based on morphological operators. The algorithm operates on multi-spectral infra-red images of small targets at the sea captured from a position simulating a large vessel bridge. The images were obtained by two infra-red cameras operating at 3-5 and 8-12 μm wavelengths and a visible range camera. All three cameras were mounted at the same platform close to each other and the capture was synchronised. The images from infra-red cameras were aligned by using fiducial markers in the scene in order to obtain pixel-to-pixel correspondences. The visible range images serve for reference and comparison. The scene contains small and slowly moving objects distant from the camera at a calm sea.

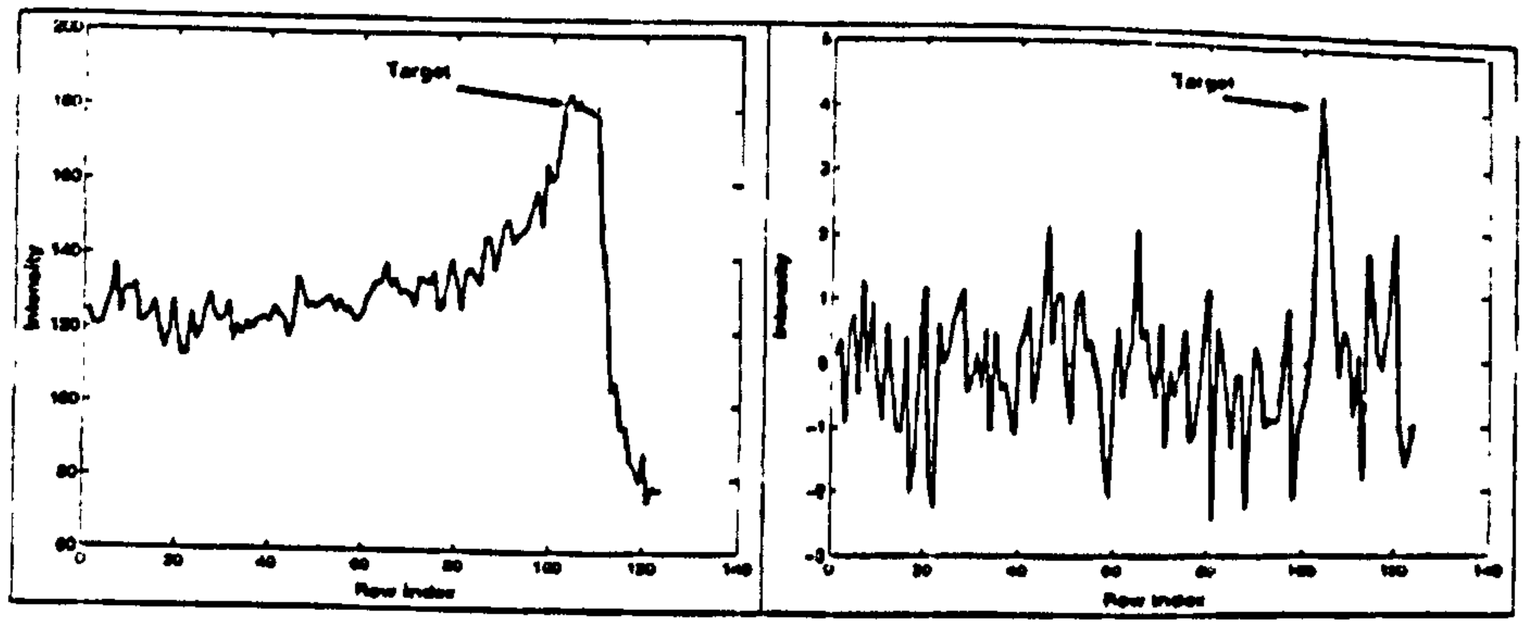
Both infra-red images are subjected to a 'top-hat' morphological filtering. The filtering consists of two steps. Bright regions smaller than the processing element are removed in the first step of morphological opening. A residual image is obtained in the second step by subtracting the opened image from the original one. The residual image contains only the bright regions removed by the opening. Both filtered images contain a considerable level of noise. The hypothesis is that noise is mostly restricted to a single particular band. By taking the intersection of the filtered images the targets are retained and noise is suppressed.

The algorithm is tested on numerous images obtained using the setup of multiple cameras described above. The method successfully detects small targets a couple of pixels in size as shown in Figure 3.11. Noise is significantly suppressed by the intersection of the 'top-hat' filtered images confirming the hypothesis of noise being restricted to a single band. The method, however, fails to detect larger objects composed of multiple parts. Unfortunately, only visual results for images of a similar nature are presented. Any further evaluation of the method is absent.

Methods by (Messer et al., 1999; Messer and Kittler, 2000; Diani et al., 2003) indicate that PCA has a potential for detection of small and dim targets in infra-red maritime scenes in presence of significant background clutter. Disadvantages of the approaches are the need of prior training and the fact that an unaccounted presence of the target in the training data significantly reduces detection capabilities of the trained feature sets. Method by Toet (2002) does not require any training, it is simple and effective. It, however, requires infra-

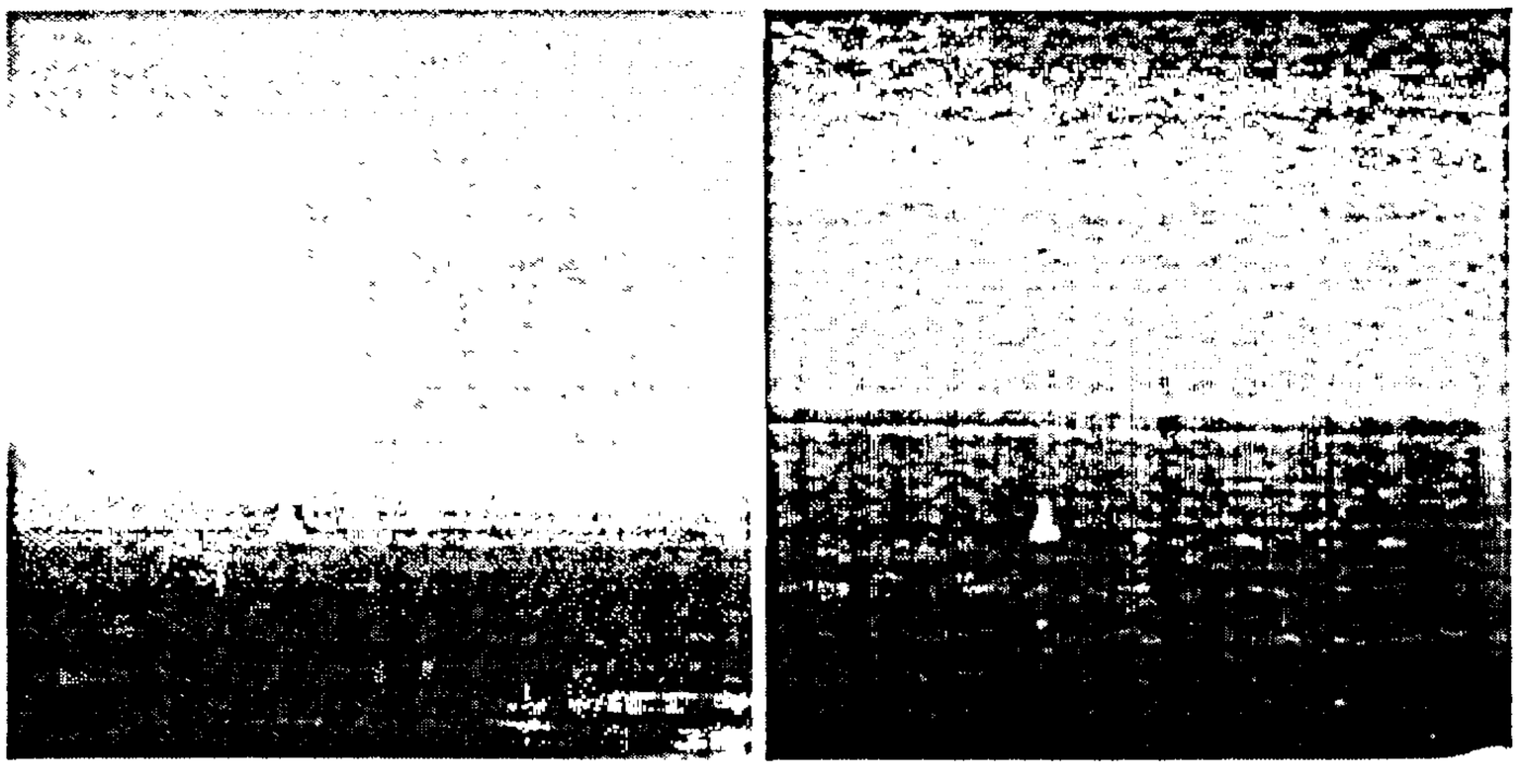


(a) The original image with superimposed target before and after clutter removal

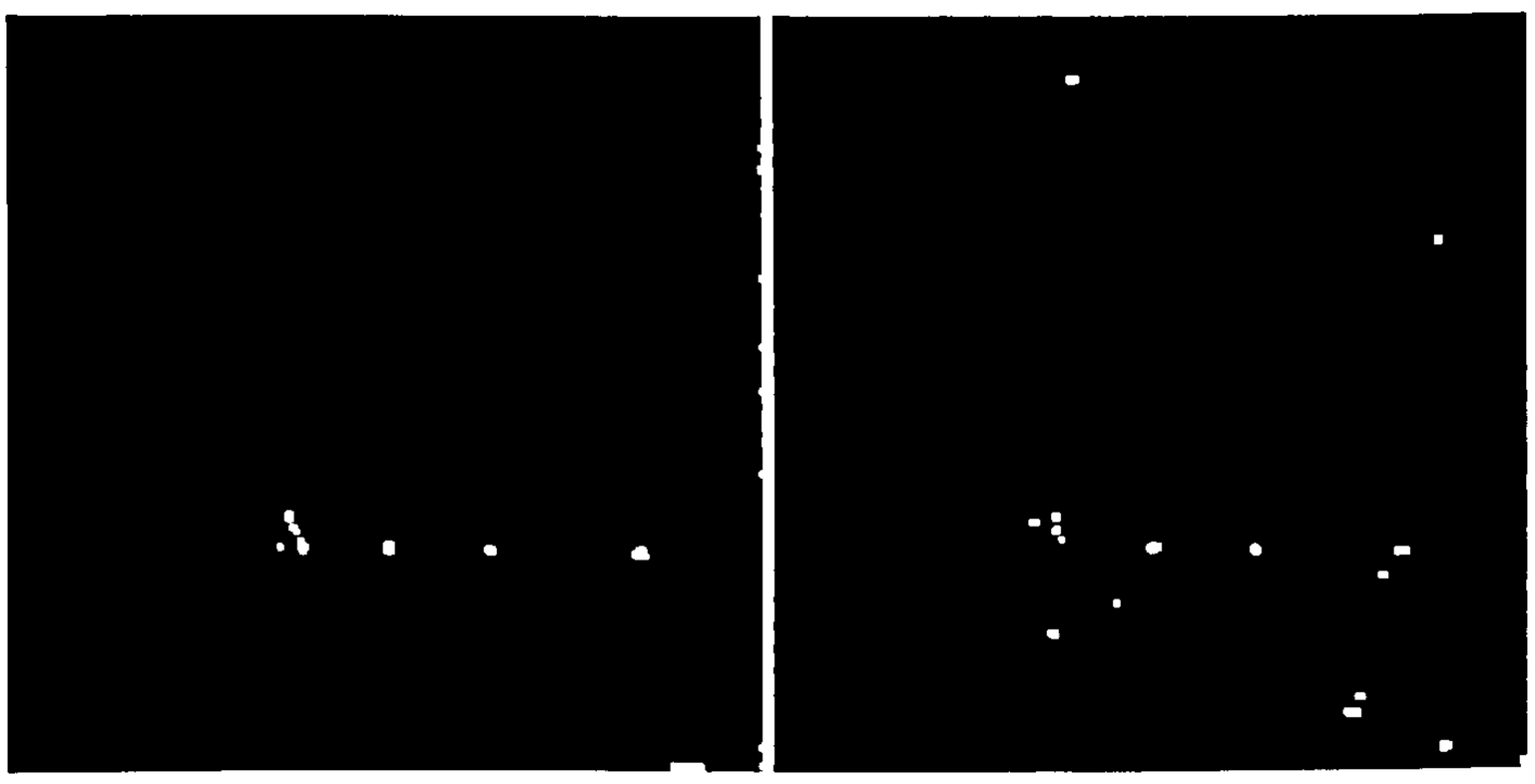


(b) The vertical cross-sections of the above images at the position of the target

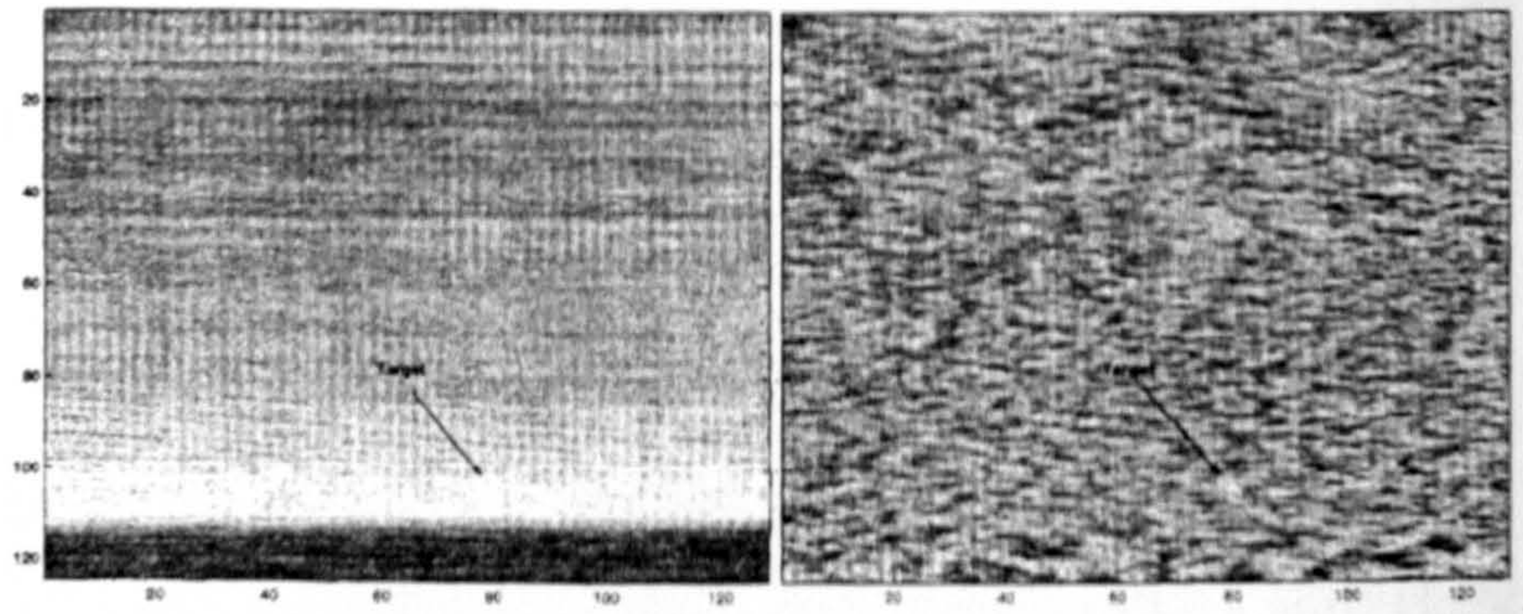
Figure 3.10: The results of the enhancement method for infra-red maritime images proposed by Diani et al. (2003)



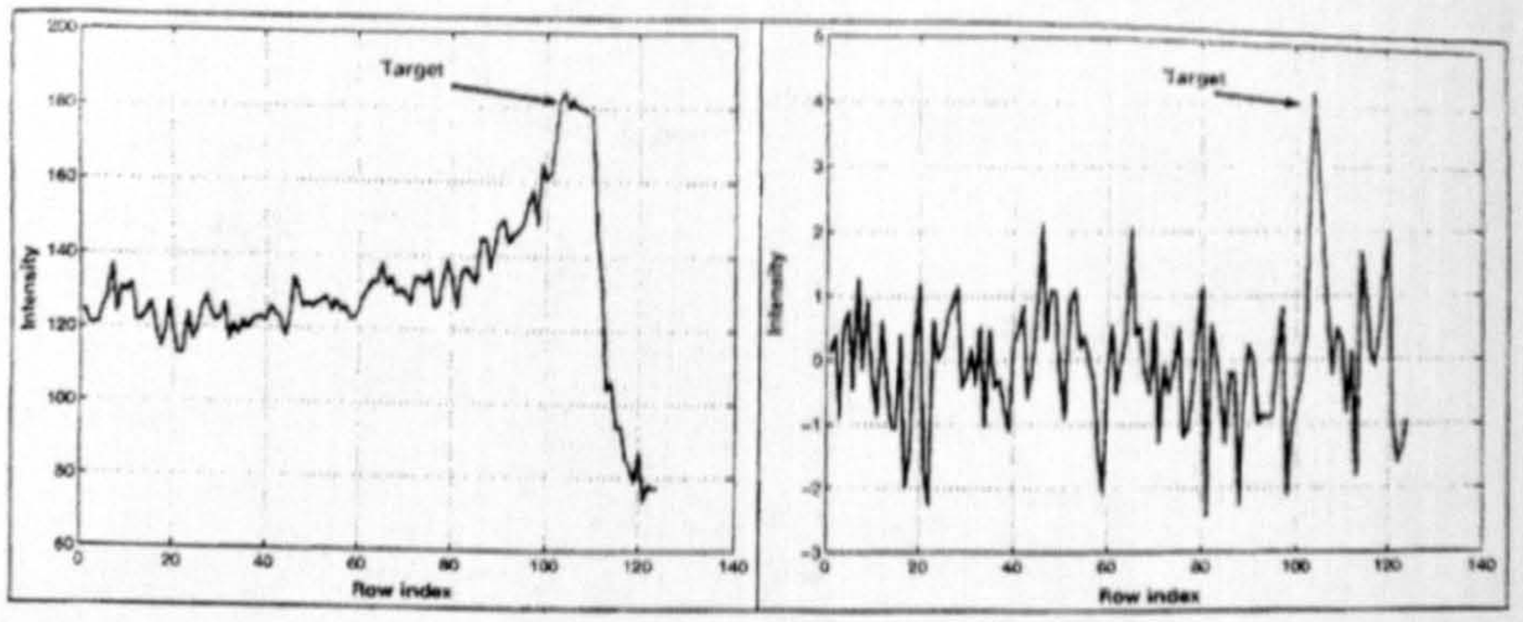
(a) 3-5 μm (b) 8-12 μm



(c) 3-5 μm alarms (d) 8-12 μm alarms

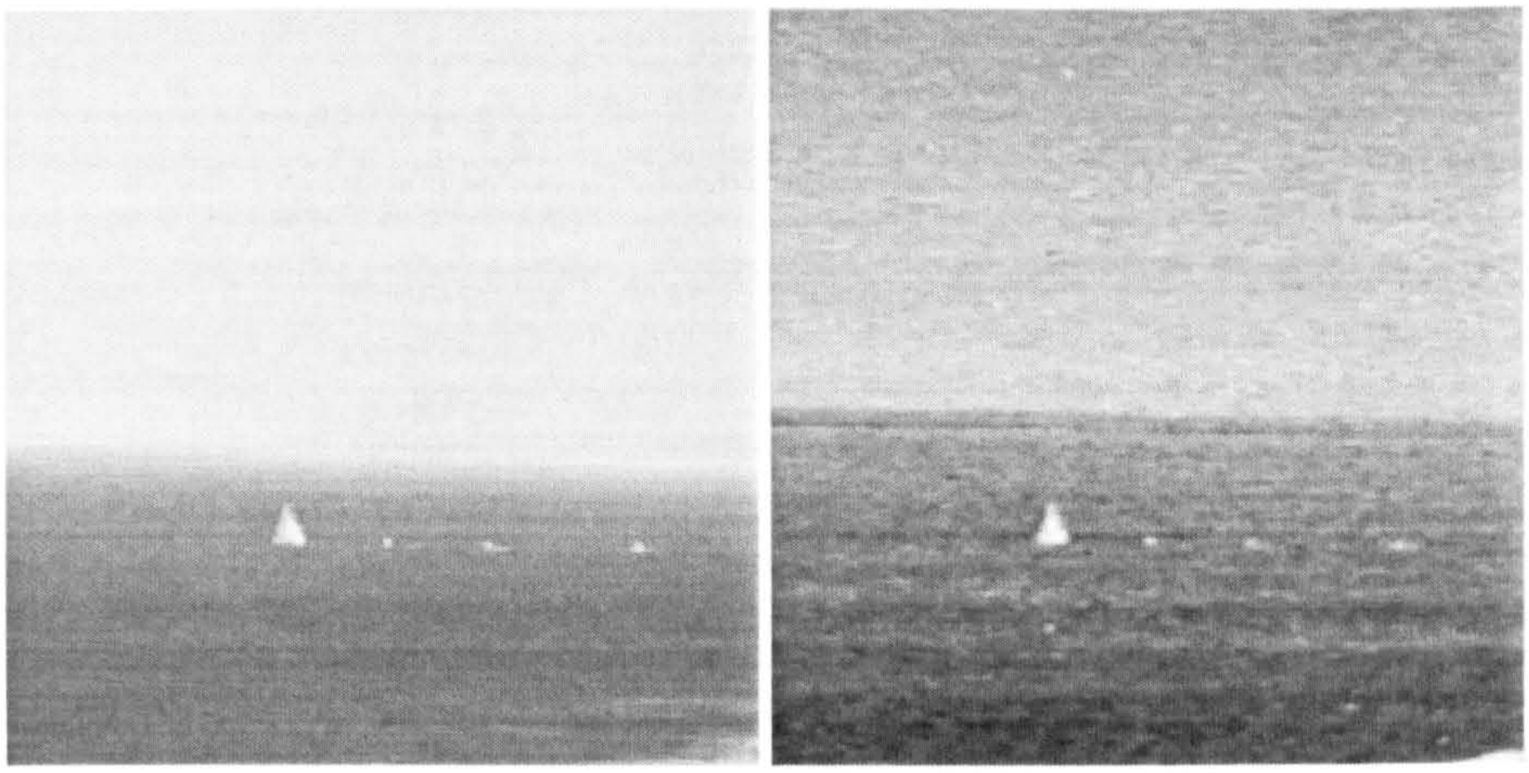


(a) The original image with superimposed target before and after clutter removal



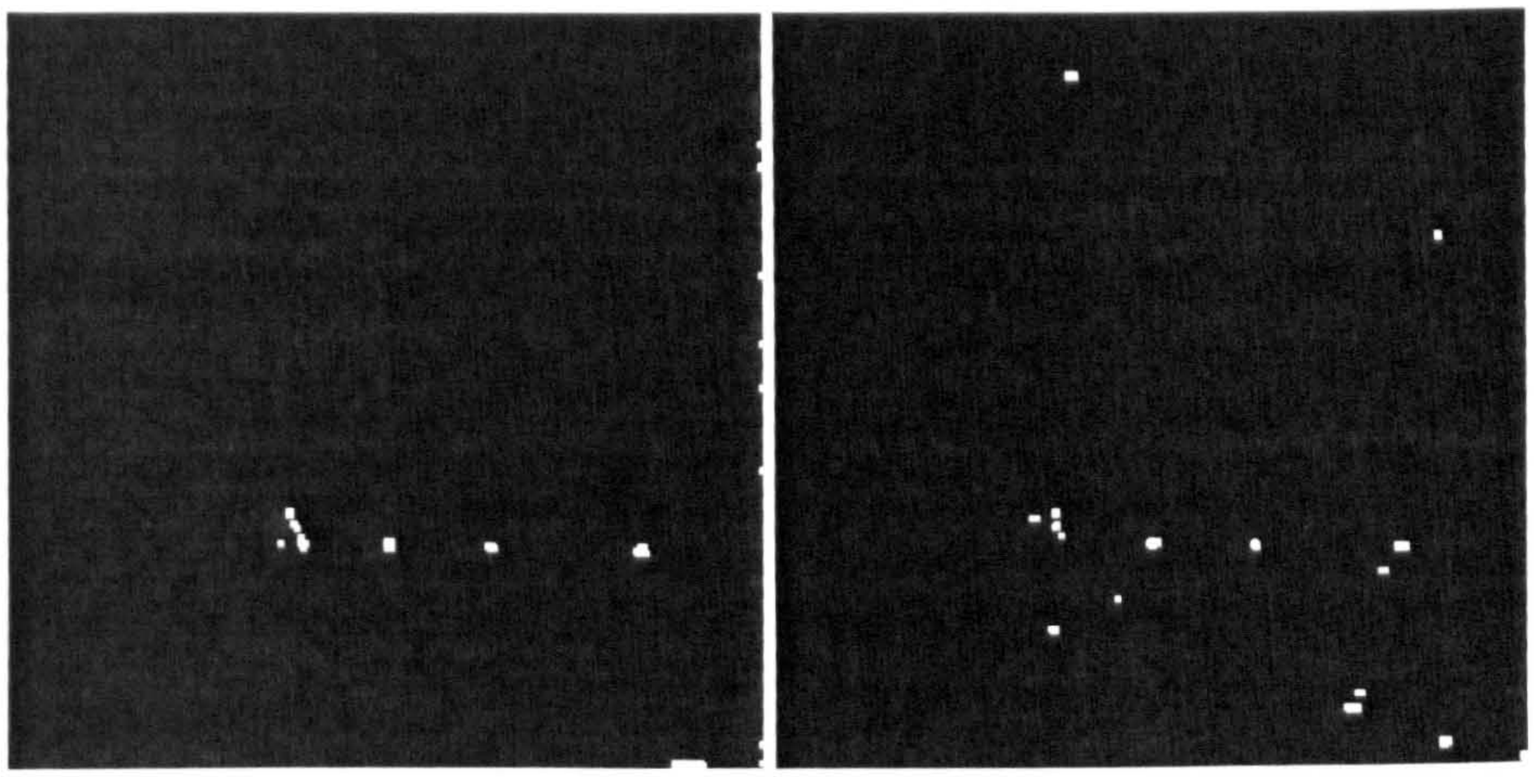
(b) The vertical cross-sections of the above images at the position of the target

Figure 3.10: The results of the enhancement method for infra-red maritime images proposed by Diani et al. (2003)



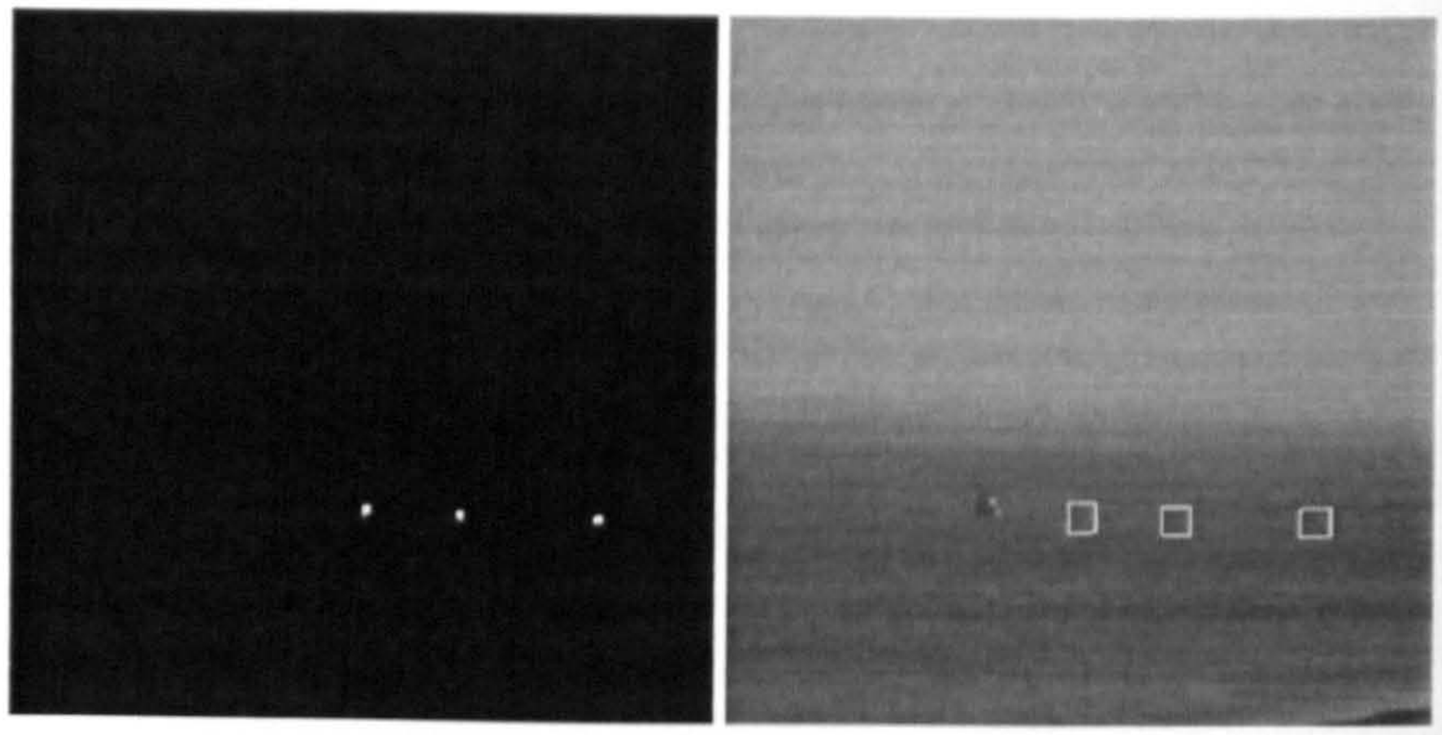
(a) 3-5 μm

(b) 8-12 μm



(c) 3-5 μm alarms

(d) 8-12 μm alarms



(e) multiband alarms (f) CCD image with potential targets

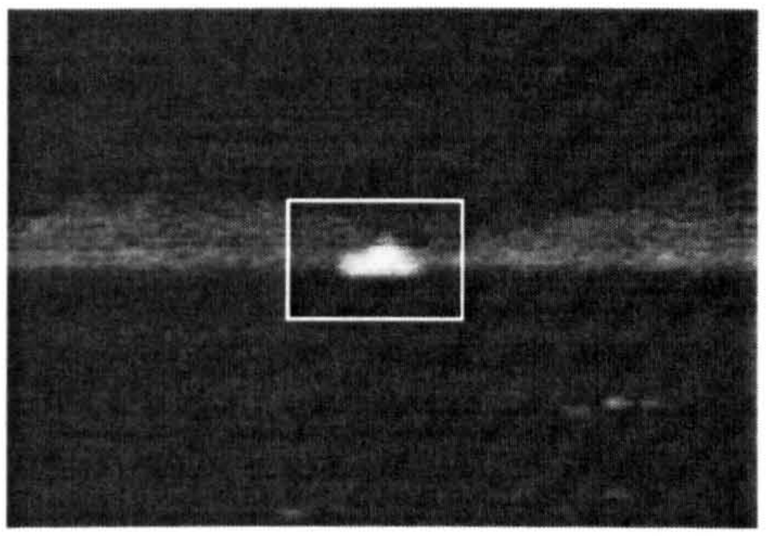
Figure 3.11: The detection of objects in multi-band maritime images by morphological filtering and fusion proposed by Toet (2002).

red images obtained in two different wavelength ranges that are synchronised and aligned. Neither of the methods have been throughly tested on more than two scenes with more complex targets.

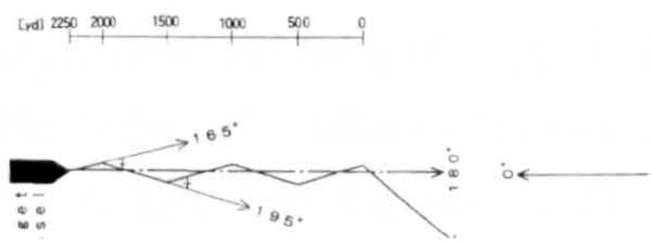
Sato and Ishii (1998) present a machine vision system that complements nautical radar. The purpose of the system is to instantly determine the pose of a vessel detected by the radar as the pose cannot be determined directly from the radar echo. The pose indicates the intended course of the vessel which does not always correspond with the direction of the trace on the radar. Unambiguous course is essential in collision avoidance especially if large vessels with slow response to a course change are involved.

The system uses an infra-red image of the detected vessel obtained by pointing the camera with a narrow field of view onto the vessel detected by radar. The vessel projects as a bright homogeneous region on a dark background. A rectangle enclosing the vessel projection is specified interactively. The area within the rectangle is filtered by median and segmented using thresholding algorithm based on area information, similar to (Sezgin and Sankur, 2004), pp. 154.

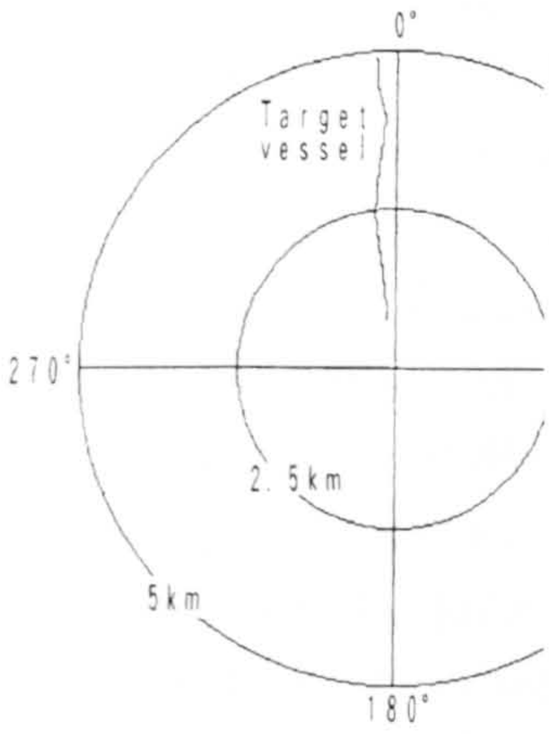
The initial length of the vessel is estimated from the width of the binarised projection and the bearing provided by the radar. An aspect angle between the course direction of the vessel and line of sight is determined from the projection width and initial vessel length. The aspect angle is then updated by using



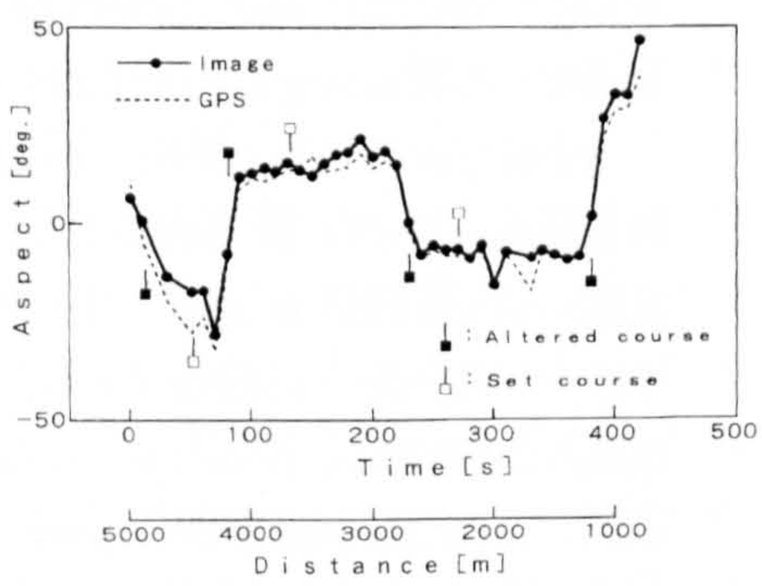
(a) Vessel projection



(b) Sample scenario



(c) Position on radar



(d) Calculated aspect

Figure 3.12: A detection of course alterations based on the infra-red projection (a) of the vessel by Sato and Ishii (1998). A sample scenario (b) shows a vessel heading straight towards the observation point changing its course along the way. The radar trace (c) does not indicate any major course alterations. The calculated aspect (d) in comparison to the GPS ground truth.

the projection width obtained from consequent frames. A relative centroid of the binarised projection is calculated. Relative horizontal position of centroid indicates the direction of the vessel's bow. The pose of the vessel with respect to its course is determined from the aspect angle and relative position of the centroid.

The system was evaluated using a real world scenario of two 400 gross ton vessels approaching at 45° and 0° bearings initially 4 km apart. The course of one of the vessels was changed by 15° to either side of the course at regular intervals. The system was capable to instantly recognise the change of the pose of the vessel. The standard deviation of the estimated aspect angle was 4.5% in the first case and 15.4% in the second case. The setup and results of the second scenario are illustrated in Figure 3.12b-d. Despite promising results there remain unresolved problems of initial object segmentation as the rectangle enclosing the vessel projection is specified manually. The method assumes that all vessels have similar geometry with the tower located at the back of the hull. The evaluation experiment was done during the winter and the vessel projects as very bright, homogeneous and well connected region against very dark and smooth background as shown in Figure 3.12a. These are rather constraining assumptions and idealistic conditions.

Withagen et al. (1999) propose a segmentation method as a part of their evaluation study of methods and features for classification of vessels from airborne infrared images. The method consists of the following steps. Shading due to non-uniform illumination is removed by fitting and subtracting quadratic surface from the original image. A top-hat transform is applied to the result that detects hot-spots corresponding to funnels. The whole vessel is detected by applying two thresholds onto the region surrounding the hot-spot. The threshold values are specified as $\mu - 2\sigma$ and $\mu + 3\sigma$ where μ and σ are mean and standard deviation of the pixel values corresponding to the surrounding sea. Asymmetry of the threshold values compensates for bright caps that appear on the sea surface. Morphological closing fills in the gaps caused by imperfect segmentation. A Hough transform is applied to find the waterline of the vessel. The vessel image is spatially transformed so that the waterline appears horizontal. The skewed image of the vessel is used to determine the features to be used in classification.

The segmentation method proves to be robust as it is a part of extensive evaluation with many various images of vessels involved. The drawback is

that the method is specifically designed for airborne images where the vessel is distant from the camera and the background structure does not significantly change with the scene depth. The method also relies on detection of the hot-spot which is available only for objects with temperatures significantly higher than the surrounding sea.

3.3.2 Visible Range Images

The research into the processing of visible range images of maritime scenes is less extensive than research into infra-red sequences. One reason is that visible range images are available only during the daytime which severely limits their use for some applications. Another reason is that water under outdoor illumination is in general perceived as complex entity with appearance that is difficult to characterise and model. This section looks at a selection of methods that either attempt to segment maritime scenes or characterise and model the water surface in outdoor environment.

Yamamoto et al. (1999) describe an airborne maritime surveillance system used in rescue operations that detects maritime craft, namely life-rafts, floating on the sea. The system consists of two cameras mounted on an airplane that overlook the sea surface from approximately 5,000 feet (see Figure 3.13). One camera is a high resolution infrared camera that detects any signal coming from emergency life flares. The other one is a colour camera that allows life-rafts to be recognised by their typical orange colour. The outputs from both cameras are filtered for noise by median filter. Filtered images are fused and a detection algorithm based on shape and intensity variance comparison is applied to identify the life-rafts in the fused image. Only visual results on a single image are provided, any further analysis is missing.

Sumimoto et al. (2000) present a similar airborne system for detection of orange life-rafts on the sea. The system operates on visible range colour images. The processing is divided into four stages depending on the environmental conditions. In favourable conditions when the sea is calm and appears homogeneous and the colour of the raft is distinguishable the detection uses difference between red and green components of the image. If the scene is dim then there is a bias between the red and green. The bias is cancelled by offsetting the values of one of the colour components. The offset is done line by line as it is assumed to vary vertically in the scene due to outdoor illumination conditions where light is coming from the top of the scene. The

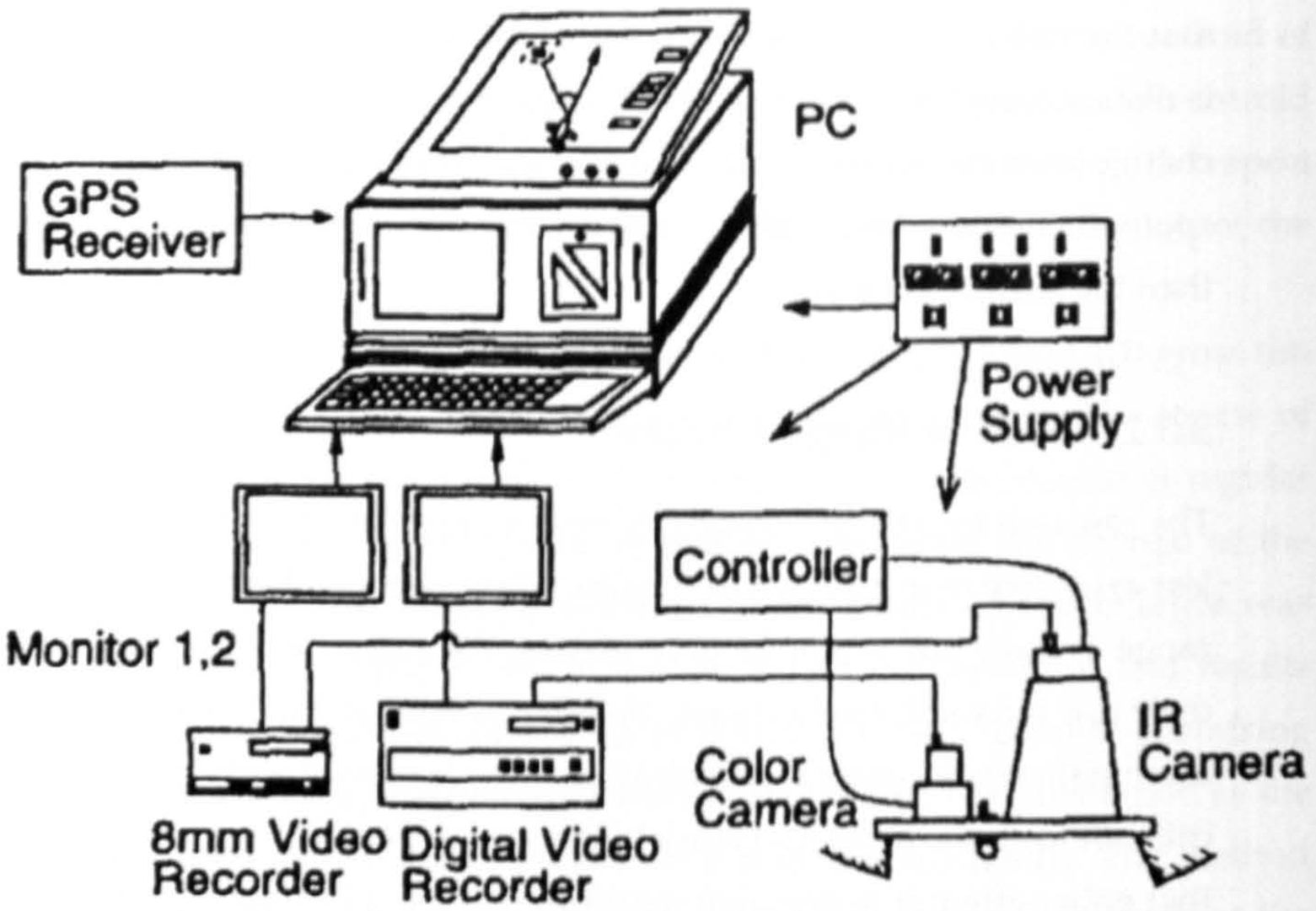


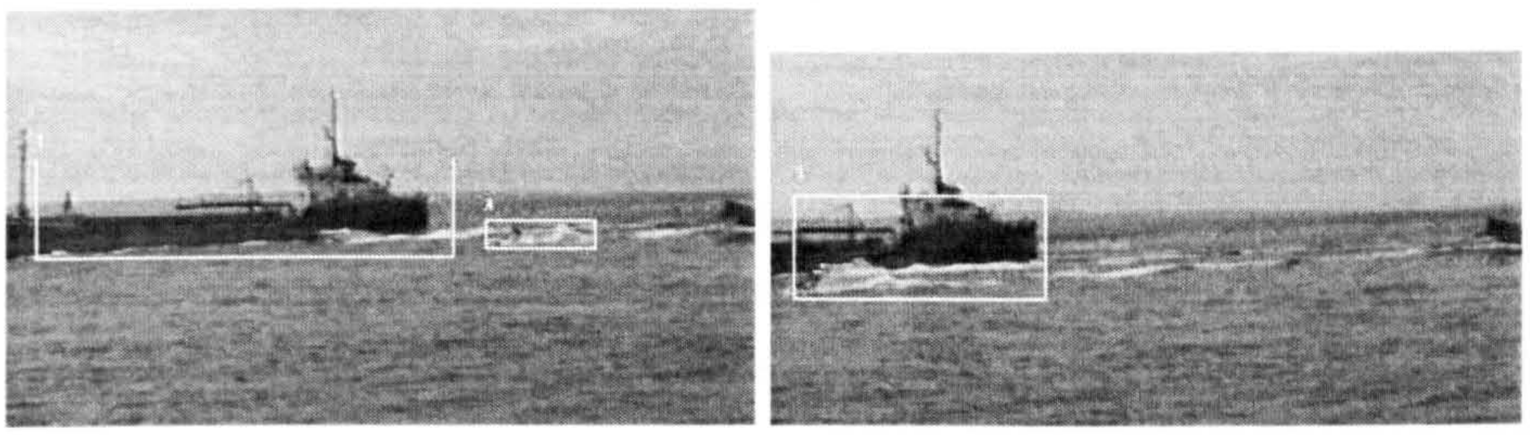
Figure 3.13: An airborne system for detection of life-raft on the sea proposed by Yamamoto et al. (1999). The system consists of colour and infra-red cameras. The life-rafts are detected in the images fused from both cameras.

colour information is unavailable during the night time. In such a case, a sensitive camera provides an image of the sea that appears homogeneous and brighter than the target. The segmentation is done by histogram thresholding. Finally, if the image is captured when the camera is facing the sun, no colour information is available and histogram based segmentation is applied.

A detailed description of the histogram-based segmentation is missing, as well as crucial details about criteria for the selection of each method. Only visual results are provided, any further analysis is missing.

Sanderson et al. (1997) present a two-stage algorithm for detection and tracking of objects in maritime scenes. The algorithm operates on visible range monochrome sequences captured from a point elevated above the sea surface by a couple of metres. The objects tracked are entering and leaving the port entrance.

The first stage provides so-called motion cues by finding regions in the scene where motion is likely to occur. A six-level pyramid representation is built by partitioning the image at each level into regular segments and determining the mean and standard deviation of intensities in each segment. A standard t-test with significance level of 5% compares the statistical pyramids



(a) (b)

Figure 3.14: Maritime scene segmentation method by Sanderson et al. (1997) based on a hierarchical statistical characterisation of image segments. The frames show segmented vessels leaving the port.

of two consequent frames at each level. A significant difference in statistics indicates change in the image due to motion. Segments with significant change at each level of the pyramid are labelled and 8-neighbourhood connectivity is applied. Each resulting blob is assigned a unique label, it's centroid and area calculated and enclosing rectangle determined. All the values are arranged in a feature vector characterising the blob.

The motion cues stored in the pyramid are evaluated at the following stage. The algorithm starts at the initial level of the pyramid at the coarsest resolution. Any change at this level corresponds to the motion of the largest objects in the scene. Any overlapping or connected blobs at lower levels are removed. If there remain any blobs at lower levels the process is repeated, starting at the actual level. A measure of 'edginess' of each remaining region in the pyramid is added to the feature vector. The measure indicates whether the motion is likely due to the presence of a rigid object. Only regions that comply with motion constraints specifying maximum acceleration and orientation change are considered. The objects are tracked by finding correspondences between regions in consequent frames by evaluating an Euclidean distance between feature vectors of each correspondence candidate.

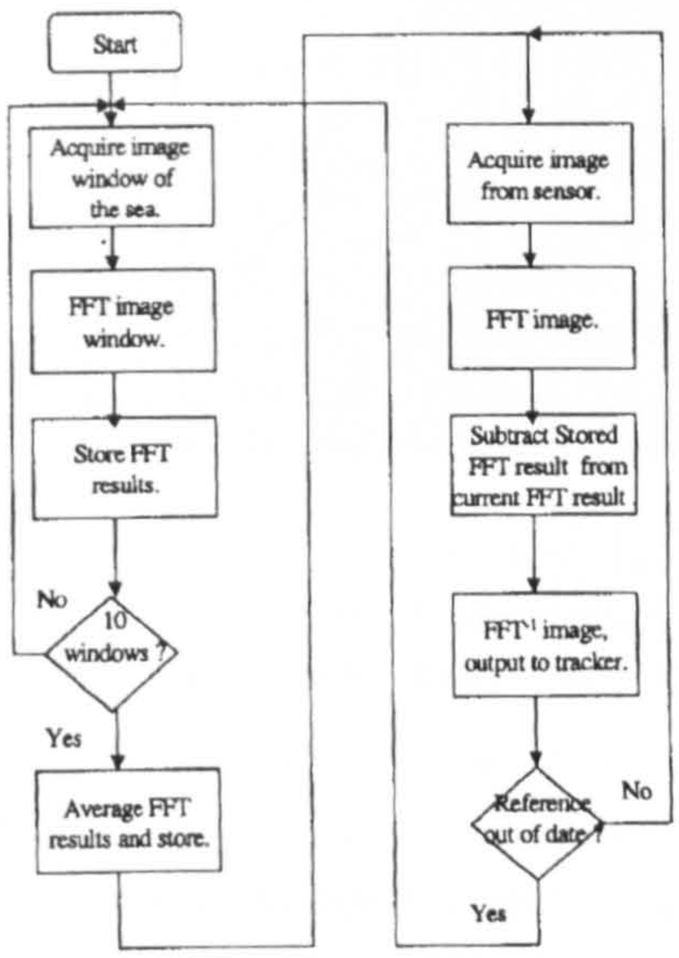
The system is evaluated on a single sequence showing a large vessel surrounded by smaller craft moving across the scene in same direction. Figure 3.14 shows that all moving objects are detected. Any further analysis of the algorithm performance is missing. Though the results indicate robust performance when detecting large moving targets the method cannot detect static objects.

Sanderson et al. (1999) modify the above algorithm by introducing an alternative motion cue generation stage of the algorithm. It is based on Fourier analysis of the sea background. In the first step a Fast Fourier Transform (FFT) is applied to 10 randomly selected regions of 32×32 pixels in a reference frame that contains sea only. The Fourier spectra obtained for each region are averaged. The resulting average spectrum characterises the current sea state.

In the second step each input frame is split into 32×32 pixel tiles. The FFT is applied to each tile and previously determined average spectrum is subtracted. Each modified tile is transformed back to spatial domain and the filtered frame is reassembled. Frame differencing is applied to filtered frames in order to obtain motion cues. Motion cues enter the tracking process described in (Sanderson et al., 1997). In case the number of motion cues increases over a specified threshold the average spectrum is updated using a recent frame. The structure of the segmentation method is shown in Figure 3.15a.

The system is evaluated on a single sequence showing two small rubber inflatable boats moving across the image (see Figure 3.15b). Both boats are detected as shown in Figure 3.15c. Due to the absence of any further analysis it is not clear how does the depth of the scene influence the average spectrum characterising the sea. Main drawback of the method is the reliance on randomly sampled reference images that do not contain any objects. The computational complexity of the method is another notable factor. The method works on 512×512 pixels images. Each image requires $16 \times 16 = 256$ FFT transforms. The complexity of 2D radix-2 FFT transform is $\frac{N_R N_C}{2} \log_2(N_R N_C)$ where N_R and N_C is number of pixels in row and column of the tile. Processing of a single image would therefore require more than 1.3 million operations.

Smith et al. (2003) employ a statistical characterisation of the sea in localised regions in order to obtain a segmentation of maritime scenes into objects and background. The algorithm consists of two steps. The range of intensity levels corresponding to the sea is determined in the first step. The sea region in the scene is divided by 2×2 grid into four areas. A set of five 32×32 pixel tiles are placed inside each area so that four tiles are located near the corners and the fifth tile is in the centre of each area as shown in Figure 3.16. Mean μ_i and standard deviation σ_i are calculated for each tile, $i = 1, \ldots, 5$. The values are compared and those that vary greatly are rejected as being contaminated by a possible object. The initial intensity range of the sea is given by $\min_{i=1,\ldots,5}(\mu_i - 2\sigma_i)$ and $\max_{i=1,\ldots,5}(\mu_i + 2\sigma_i)$ of remaining values.



(a) Segmentation scheme

(b) Sample frame

(c) Frame segmented

Figure 3.15: Maritime scene segmentation method by Sanderson et al. (1999) based on a substraction of FFT spectra of the background. The frames show segmented vessels leaving the port.

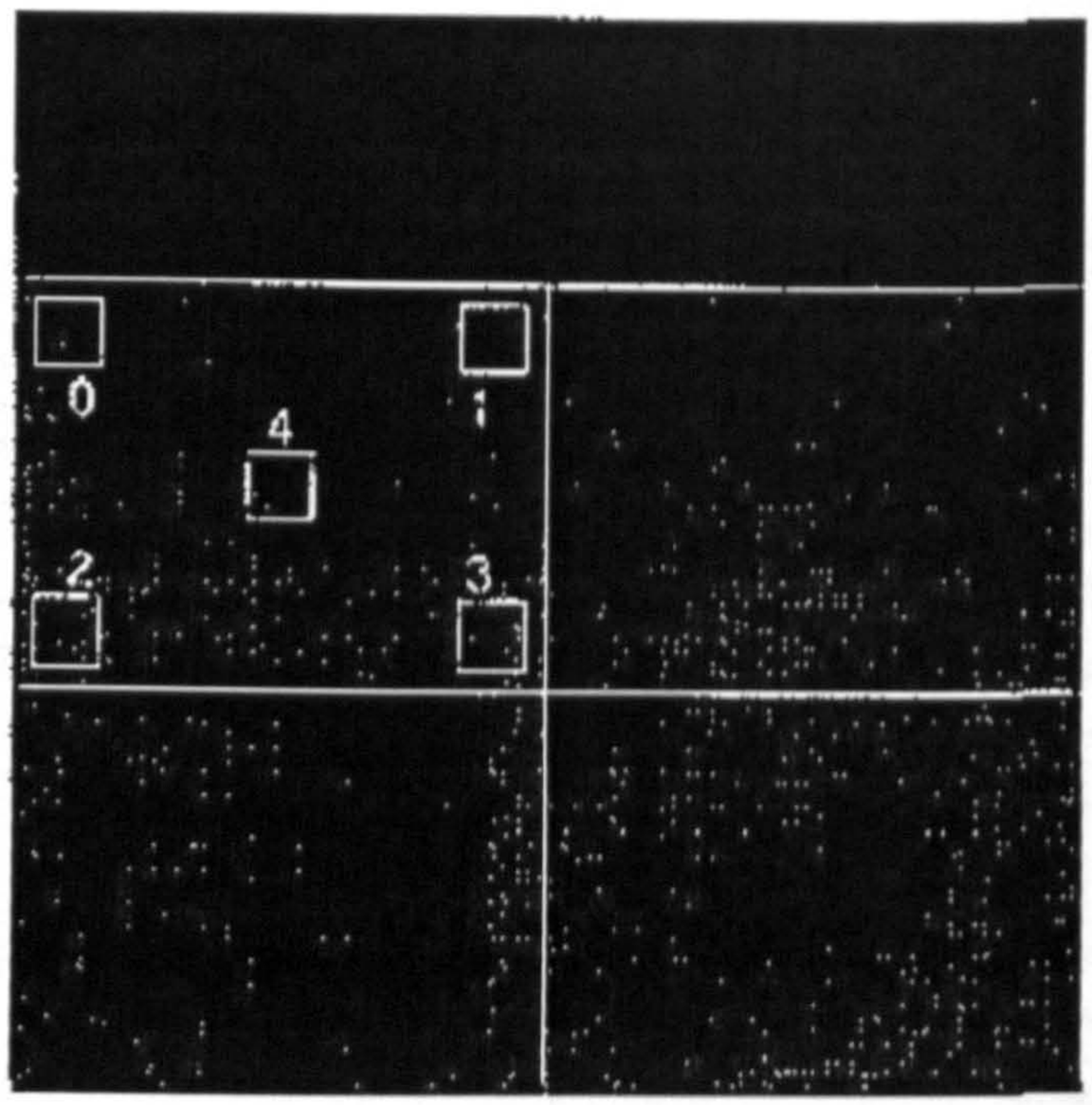


Figure 3.16: The grid used by Smith et al. (2003) to statistically characterise the sea intensity levels. Intensity means and standard deviations are determined for pixels within the marked segments in each quarter of the image.

The process is repeated for each of the four areas.

The sea region is split into 32×32 pixel tiles in the second step. Each tile is labelled as the sea or the object by determining a relative amount of pixels that lie within the sea intensity range estimated in the previous step. If the amount is more than 90% the tile is labelled as the sea. If the amount is less than 10% the tile is labelled as an object. Labelling of individual pixels is applied in case the value is between 10% and 90%. A maximum m and standard deviation σ of the pixels in the tile are determined first. If the maximum m lies inside the sea range then pixels with intensity within $m \mp \sigma$ range are labelled as the sea. Other pixels within the tile are labelled as object. If the maximum m lies outside the sea range then pixels with intensity within $m \mp \sigma$ are labelled as object and remaining pixels are labelled as the sea. The segmentation results are improved by joining object pixels using an N-way connectivity check.

The algorithm is evaluated on two maritime sequences showing static and moving objects. A numerical analysis shows that the number of incorrectly classified tiles is less than 5% in the first sequence and 10% in the second

sequence. The method exhibits a good performance in cases of calm seas and homogeneous objects. It, however, tends to break up more structured objects into multiple separate parts.

Spencer and Shah (2004) introduce a spatio-temporal analysis based on Discrete Fourier Transform (DFT) of the video sequences showing the sea surface. The analysis provides estimates of various parameters that represent the state of the sea, namely wave height, wave period and, consequently, wind speed. The actual scale of objects in the scene can be derived from the knowledge of these parameters.

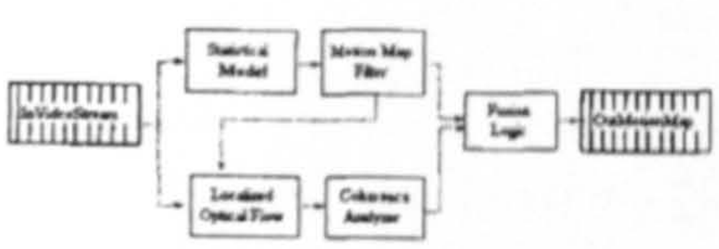
The first step of the analysis consists of the DFT of each frame of the sequence. The Fourier spectrum expressed in polar coordinates encodes the spatial periodic patterns that represent the waves. A profile of the spectrum magnitude is obtained by averaging the spectrum across all angles in the polar coordinates. The position of the peak in the profile is reciprocal to the wavelength in pixels of the most dominant waves on the sea.

The second step involves a temporal analysis of the sequence. The amount of data in the sequence is reduced by Principal Component Analysis. The sequence is represented by a 2D array where rows correspond to frames and columns correspond to PCA coefficients. The DFT is applied to the array and the spectrum values in each row are summed. The sums indicate the energy at each temporal frequency. If the frame rate of the sequence is known the period of the waves in seconds can be determined from the position of the maximum in the spectrum energy sums.

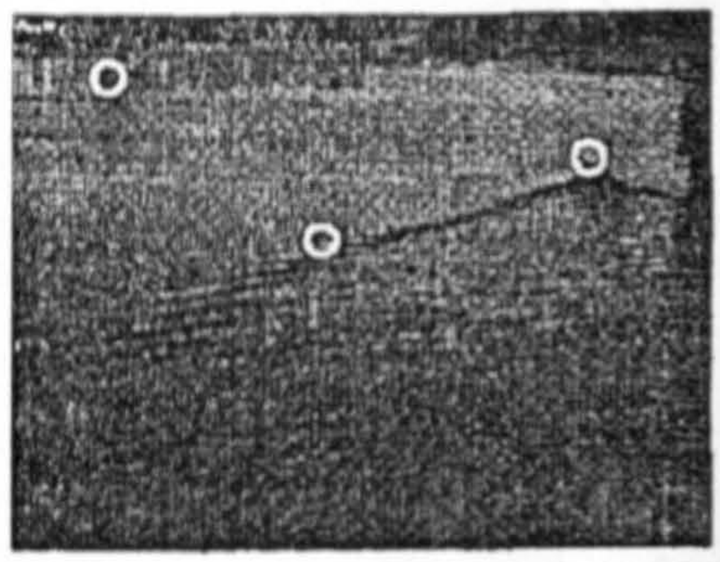
The wavelength in metres is given as $L = \frac{g}{2\pi}T^2$ where g is acceleration due to gravity and T is the period of the waves in seconds determined from the temporal analysis. Finally, the scale of the scene in pixels per metre is obtained by correlating the temporal spectrum with the spectra of the individual frames.

The results of the analysis can be used, for example, to determine wave heights which are given as $0.008L$ to $0.1L$. The wind speed can be approximated from the wave heights using Beaufort scale.

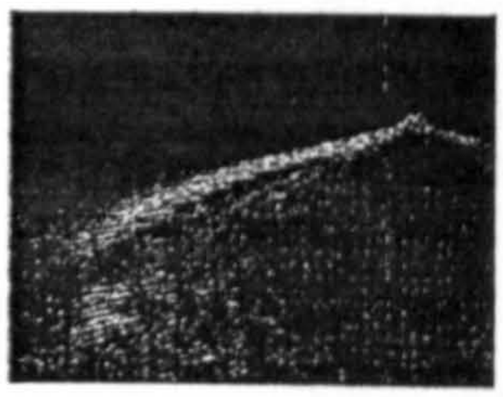
The analysis is tested on three sequences showing various states of the sea. The estimates of the wavelengths, wave heights and wind speeds are realistic and, according to authors correspond to the actual weather conditions at the day of capture. These preliminary results suggest that the method can be used for other purposes such as determining the camera parameters (zoom, tilt, roll) or scene segmentation.



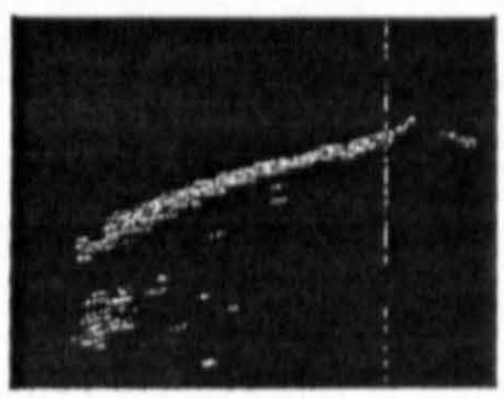
(a) system structure



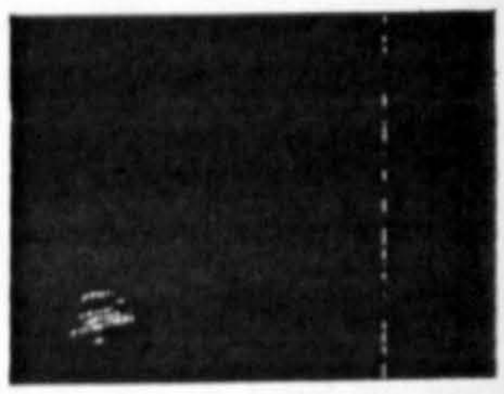
(b) sample frame



(c) initial segmentation



(d) coherence map



(e) fused result

Figure 3.17: Water scene segmentation by Ablavsky (2003): (a) the structure of the segmentation system, (b) a sample frame, (c) initial segmentation of the foreground, (d) coherence map containing the wake, (e) fused result.

Ablavsky (2003) points out that most background modelling techniques based on statistics do not consider spatio-temporal correlations between the background pixel values typical for water surfaces in outdoor scenes. He proposes a background estimation model that combines a local optical flow with a statistical background model similar to one described by Haritaoglu et al. (2000) (see Figure 3.17a).

The regions that violate the statistical background model are selected first. The selected regions form the first foreground likelihood map shown in Figure 3.17c for a sample sequence. A motion map filter designed to extract regions of a high eccentricity is applied to the selected regions. Highly eccentric regions supposedly correspond to wakes, ripples and small moving objects. A localised optical flow calculation is applied to the output of the filter. A coherence of the optical flow directions is determined and the second

foreground likelihood map is obtained that associates the probability of the wakes with the regions in the map as shown in Figure 3.17d. Both maps are fused by Bayesian rule to provide a probability of regions belonging to foreground or background classes (see Figure 3.17e).

Ablavsky states that the method is being tested on several dozen sequences at various environmental conditions. The results presented in (Ablavsky, 2003) are for a single sequence of small boat moving away from camera on a calm water surface. The results show that most of the wakes are successfully removed from the image. It is not, however, clear whether the object has been detected as well or if it has been removed together with wakes. No further analysis is provided. The method assumes that the motion of the wakes is consistent and that their appearance does not significantly change over extended period of time. Such assumptions are valid only on a calm water such as very calm seas or slow rivers and they do not apply to most maritime scenes.

3.4 Summary

The importance of visual information is clearly recognised in maritime sector. Night vision systems based on infra red or light intensifying visual sensors (Vistar Night Vision Limited, 2004c; The Current Sales Corp., 2004) are becoming essential aids to navigation in unfavourable environmental conditions. The systems, however, merely provide the images from outside the bridge. The process of detection and identification of objects in the images relies strictly on the vigilance of the operator.

This is in contrast with land-based surveillance systems where a high degree of automation has been achieved, (Dick and Brooks, 2003). The systems are capable of automated classification of object types (Collins et al., 2000), suspicious activity of individuals in the scene (Cupillard et al., 2003), analysis of the traffic parameters (Tai et al., 2004), etc.

The major obstacle in development of similar applications in maritime sector is the spatio-temporal variability of the maritime scene due to the presence of waves. Machine vision applications in maritime sector concentrate mainly on processing of infra-red sequences. Various infra-red image segmentation algorithms detect small and weak targets (Messer and Kittler, 2000; Messer et al., 1999; Diani et al., 2003) common in military and rescue

operations. Several applications employ a fusion of multi-spectral images in order to improve the detection rate and reduce the noise, (Toet, 2002; Yamamoto et al., 1999).

There is a limited number of methods and applications that detect and track objects in visible light images in maritime sector. The method by Sanderson et al. (1997) based on statistical differences in the hierarchical pyramid and the modified version (Sanderson et al., 1999) using the substraction of the sea spectrum detect only moving objects. The algorithm by Smith et al. (2003) detects both static and moving objects. It, however, requires objects to have intensities outside a specific interval which is not always the case in maritime scenes.

Studies by Spencer and Shah (2004) and Ablavsky (2003) concentrate on modelling of the phenomena of the water surface and wakes caused by moving objects. Both studies provide only limited evaluations of the functionality of the proposed methods. Possible use of the methods for a segmentation of the scene are only suggested in the conclusions of both studies.

The works concerned with infra-red and visible range maritime scenes are either specialised at a particular task such as a detection of life-raft from an airborne platform (Yamamoto et al., 1999; Sumimoto et al., 2000), detection of change in course (Sato and Ishii, 1998), classification of objects (Withagen et al., 1999), etc. or provide a mere detection of objects in infra-red (Messer et al., 1999; Messer and Kittler, 2000; Toet, 2002) or visible sequences (Sanderson et al., 1997; Sanderson et al., 1999; Smith et al., 2003) without any consequent processing.

The review indicates that despite the importance of visual information in maritime navigation and safety applications a robust system that would process and exploit this information in a way similar to land-based surveillance and tracking systems is yet to be developed. The basis of such a vision-based framework is already available in the maritime sector in terms of night-vision systems. It is a matter of providing it with more functionality and automation.

Chapter 4

Segmentation

4.1 Introduction

The nature of the background in the maritime scene is temporally and spatially highly changeable in appearance as discussed in Chapter 2. The waves on the sea typically appear as a nearly regular pattern with apparent directionality due to the perspective projection. Features such as wakes and crests of similar size are scaled down towards the horizon in the image. Despite an evident presence of a perceptible regularity in the pattern of the sea that can be described as a texture, the texture analysis remains unexplored in the maritime related research. Texture is characterised by a spatial distribution of pixel intensities in a fundamental neighbourhood (textons) that is repeated periodically with little or no variation. Texture analysis identifies the texture by characterising the textons and their distribution. Generally, there is an apparent spatial relation between neighbouring elements in textures that are not purely random. This relation can be used to identify texture properties as shown by Chetverikov and Haralick (1995).

In a similar fashion, the objects in maritime scenes can be perceived as being composed of neighbouring texture patches. These texture patches typically differ from textures representing the sea. Usually they are more homogeneous or their structure or intensity range vary. Even though the difference in intensity ranges for sea and objects has been exploited by Smith et al. (2003), the results show that it is not always reliable criterion for segmentation, especially when objects are composed of parts with various intensities.

Instead, segmentation can be achieved by characterising various textures in the scene and separating those corresponding to the objects from the texture of the sea. The segmentation of the maritime scenes by textural characterisation is the subject of the following sections.

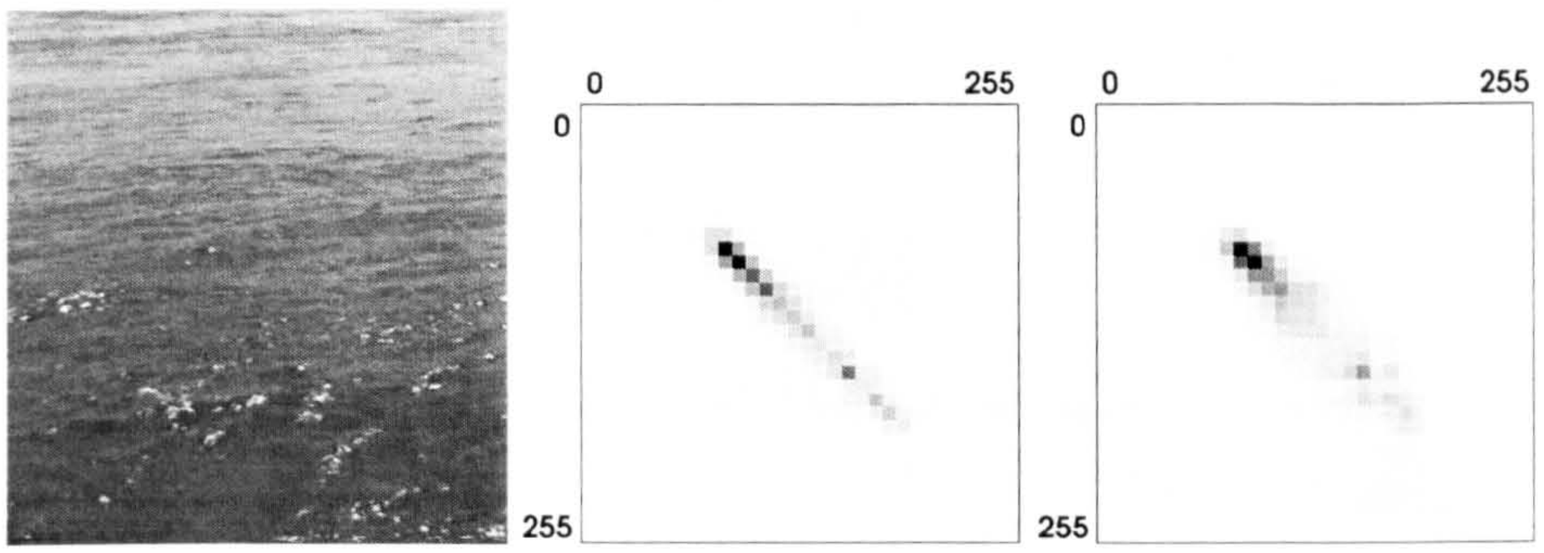
4.2 Statistical Characterisation of Textures

Textures are generally composed of multiple repeating primitives that are spatially transformed. It is possible to characterise textural properties by using either semantic, spectral or statistical approaches, (Schalkoff, 1992). Semantic approaches are well suited for highly regular undistorted textures that are mostly uncommon in natural outdoor scenes. Spectral approaches to texture characterisation of maritime scenes were explored by Spencer and Shah (2004) who apply the Discrete Fourier Transform to image sequences of the sea in order to identify wave parameters. The method, however, does not convey image segmentation. Sanderson et al. (1999) determine the frequency spectrum of the regular sea pattern and filter it out prior to the segmentation of the scene. The method is computationally demanding and it cannot detect static objects. Statistical methods are used in the analysis of outdoor scenes as they capture the stochastic properties of the natural textures.

Various properties and attributes of the texture patterns are described by so-called features. Features have two main purposes: they reduce the amount of data to be processed and they provide more efficient representation of the texture. Features can be determined either directly from the image intensities (Laws, 1980) or as parameters of some functional model that approximates the data by, for example, minimising residual error or maximising likelihood (Chen and Kundu, 1995). The latter approach is inefficient in maritime scenes due to the complexity of the sea texture, as pointed out in Section 2.2. Therefore, the features are inferred directly from the intensity data.

The spatial correlation of the texture pattern can be described by a co-occurrence matrix. The co-occurrence matrix contains the joint probabilities of two pixels having certain intensity values and particular layout. The layout of the pixel pairs must be specified in advance of constructing the co-occurrence matrix. This, however, assumes a prior knowledge about the texture geometry.

In practice, the co-occurrence matrix size is reduced by assigning each single column or row to a range of intensities. For example, instead of



(a) sample image (b) dx=5,dy=0 (c) dx=0,dy=5

Figure 4.1: A sample maritime image (a) and co-occurrence matrices (b),(c) for two layouts of the pixel pairs. In the first layout (b) the pair is 5 pixels apart horizontally (dx=5,dy=0). In the second layout (c) the pair is 5 pixels apart vertically (dx=0,dy=5). Both matrices are 32×32 elements in size where one element corresponds to a range of 8 intensities. Dark elements correspond to high counts of the corresponding pixel pairs.

using all 256 intensities in the image which would lead to a 256×256 matrix, the intensities are quantised by a factor of 8 which gives a 32×32 matrix as illustrated in Figure 4.1. The reduction of the intensity resolution can lead to a situation where low-contrast textures are characterised as homogeneous regions.

Some suggestions how to decrease a computational demand of co-occurrence matrices have been proposed by Argenti et al. (1990). Walker et al. (1995), propose a method for improving the discrimination of co-occurrence matrix features.

The co-occurrence matrix is an intermediate step in the characterisation of the textures. The features characterising the texture are obtained as statistical weighting functions (features) applied to the data in the co-occurrence matrix. These weighting functions employ either the values in the co-occurrence matrix or their positions, (Walker et al., 1995). Four of these commonly used features are:

$$Energy = \sum_{r,c=0,0}^{R-1,C-1} f(r,c)^2 \tag{4.1}$$

$$Entropy = \sum_{r,c=0,0}^{R-1,C-1} f(r,c)\log(f(r,c)+1) \tag{4.2}$$

$$Homogeneity = \sum_{r,c=0,0}^{R-1,C-1} \frac{f(r,c)}{|r-c|+1} \tag{4.3}$$

$$Contrast = \sum_{r,c=0,0}^{R-1,C-1} f(r,c)(r-c)^2 \tag{4.4}$$

where $f(r,c)$ is the element value at the position (r,c) in the matrix with dimensions R, C.

4.2.1 Redundancy of the Co-occurrence Matrix

The ultimate goal of the segmentation is not to characterise the texture of the sea in an absolute quantitative way but to separate varying textures representing the objects from the texture of the sea. It is then possible to apply the features directly onto the image data instead of the co-occurrence matrix.

A co-occurrence matrix can also be regarded as an intensity image of different texture where the values of matrix elements correspond to the intensity values at each pixel. Similarly, an image can be regarded as a co-occurrence matrix determined for some unknown texture. The intensity values in the image segment then represent values of the elements in such co-occurrence matrix.

If there is an object in the segment of the image then the structure in the segment would differ from that of the sea. Similarly, the structure of a co-occurrence matrix and hence the feature values vary for different textures. Regarding the values of the features, it is the relative difference between the features characterising the sea and the objects which is important, not their absolute values. If the features are able to quantify the difference between different textures directly from the image intensities, the co-occurrence matrix would become redundant.

This hypothesis was later confirmed in an experiment where the segmentation described in the following Sections 4.3 - 4.7 was applied using features obtained from co-occurrence matrices. The co-occurrence matrices were generated for three intensity resolutions (8, 32 and 64 levels) and nine different spatial configurations covering all possible directions (see Figure 4.2).

The results of segmentation utilising co-occurrence matrices were compared against ground truth obtained by applying the same segmentation on evaluation sequences using features calculated directly from image intensities

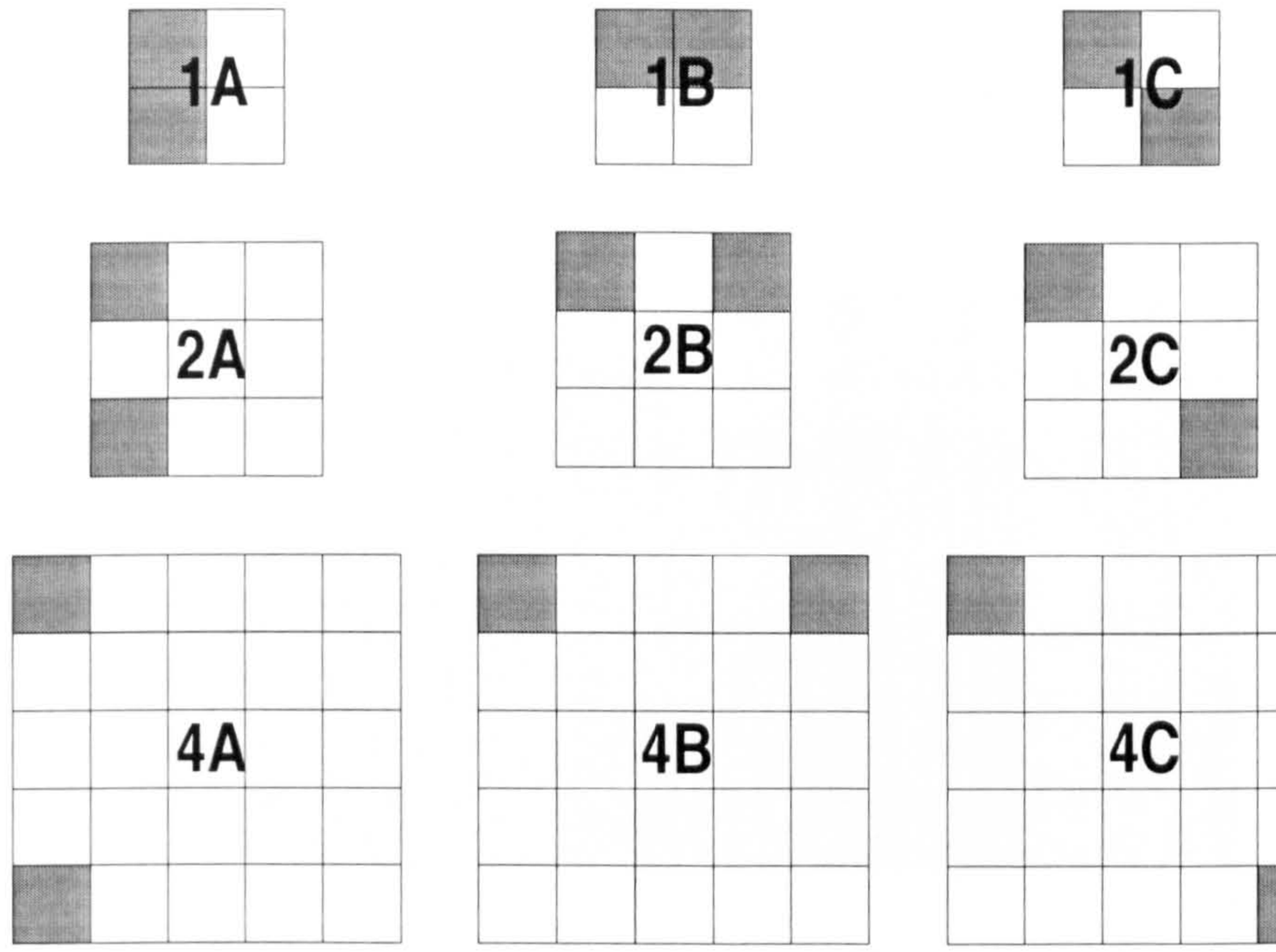


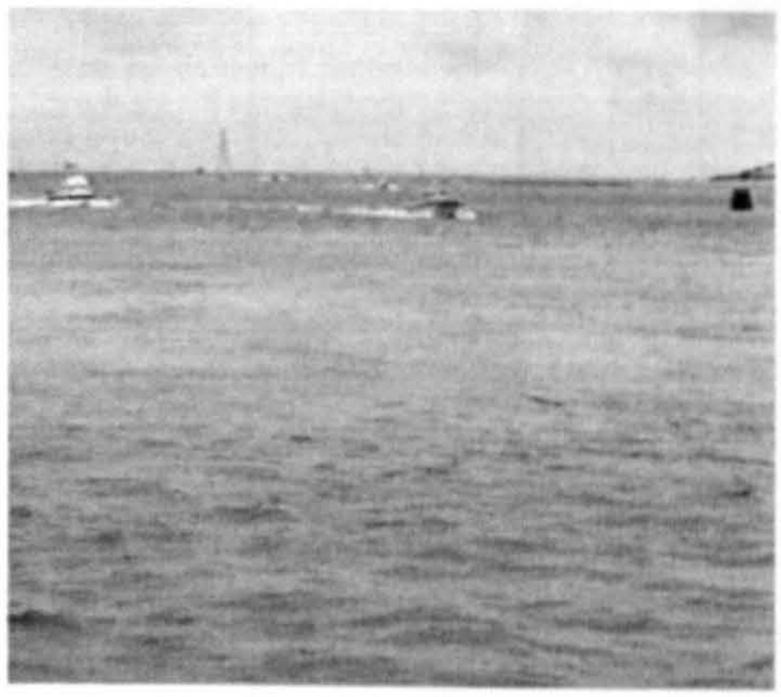
Figure 4.2: Spatial configurations of pixels used in the generation of co-occurrence matrices.

and visually checking the correctness of the segmentations. Any frames where the prior segmentation failed were excluded from the evaluation. Three criteria commonly used in evaluation of object segmentation techniques (Cohen and Medioni, 1998) were considered:

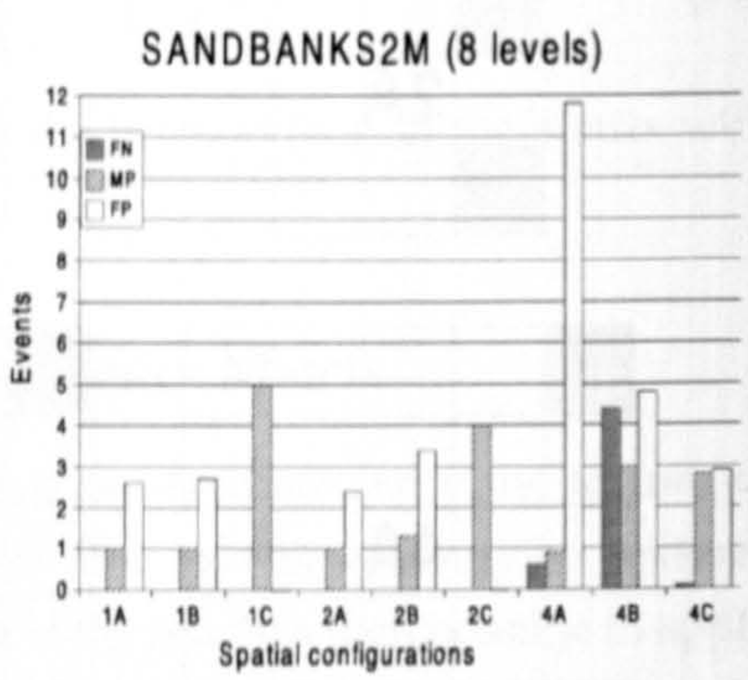
- false negatives (segmentation failed to identify an object),
- mismatched positives (an object was split into multiple parts or multiple objects were merged into a single one),
- false positives (segmentation identified non-existent objects in the scene).

The evaluation was applied to SANDBANKS2M (first 45 frames) and WEYMOUTH2E (first 100 frames) sequences. The graphs in Figures 4.3b-d and 4.4b-d show average numbers of detected false negatives (FN), mismatched positives (MP) and false positives (FP) per sequence.

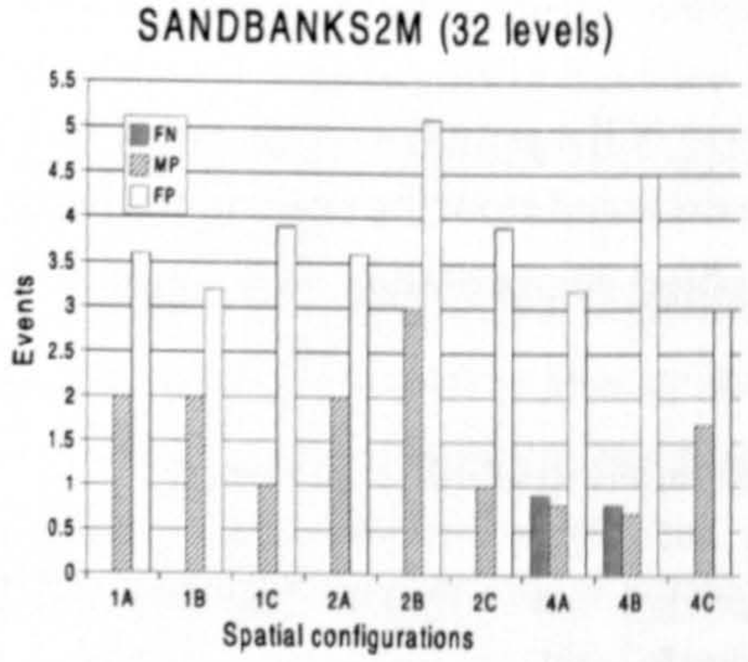
The results indicate that the features determined from co-occurrence matrices do not perform better than those determined straight from intensity data. Furthermore, there is an obvious dependency on the configurations of the pixel pairs as well as the intensity resolutions that must be set beforehand.



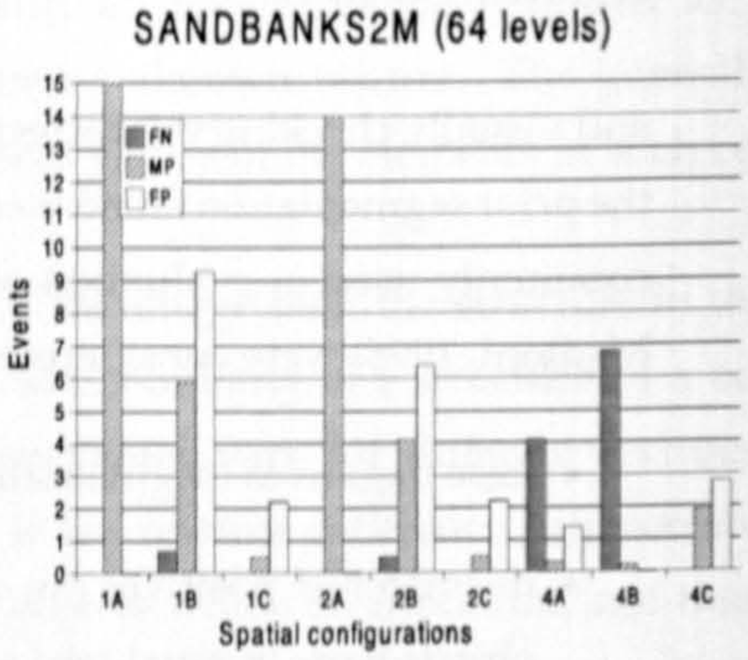
(a) sample frame



(b) 8 intensity levels



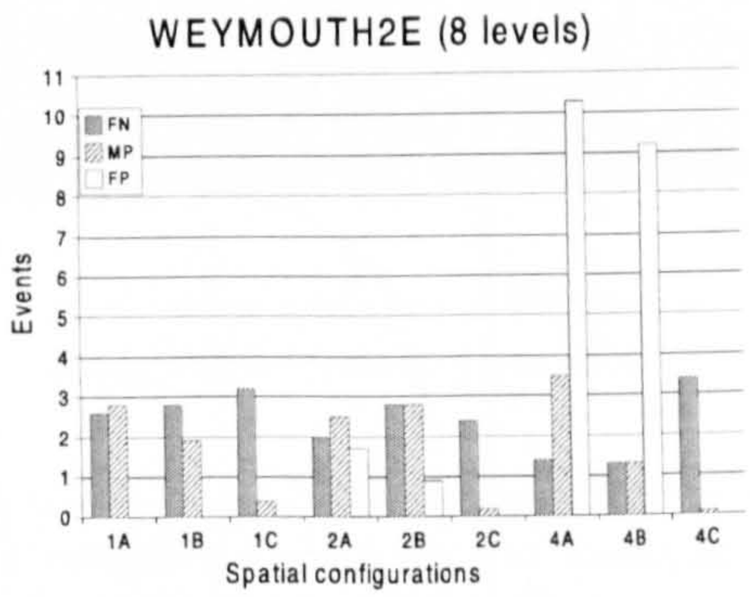
(c) 32 intensity levels



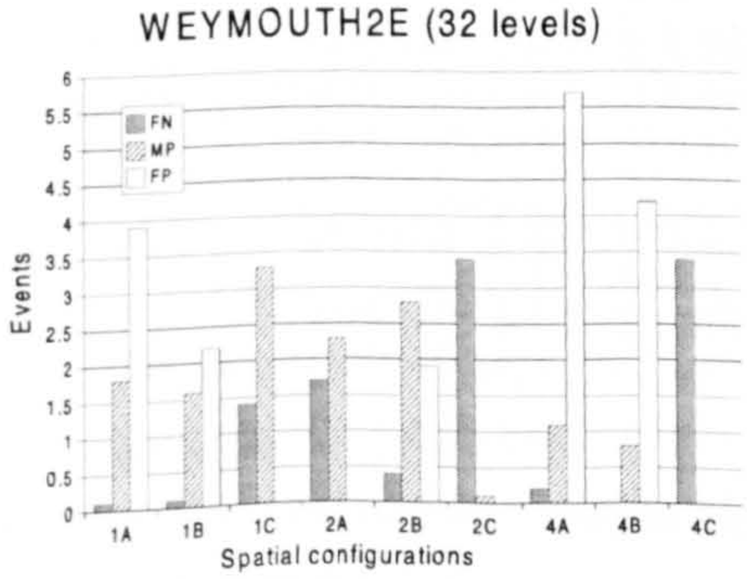
(d) 64 intensity levels

Figure 4.3: Evaluation of the co-occurrence matrix redundancy hypothesis using the segmentation described in Sections 4.3 - 4.7. Results shown for SANDBANKS2M sequence: (a) sample frame, (b)-(d) average false negatives (FN), misclassified positives (MP) and false positives (FP) per sequence for various spatial configurations of the pixel pairs. Three intensity resolutions of the co-occurrence matrices are considered - 8, 32 and 64 levels.

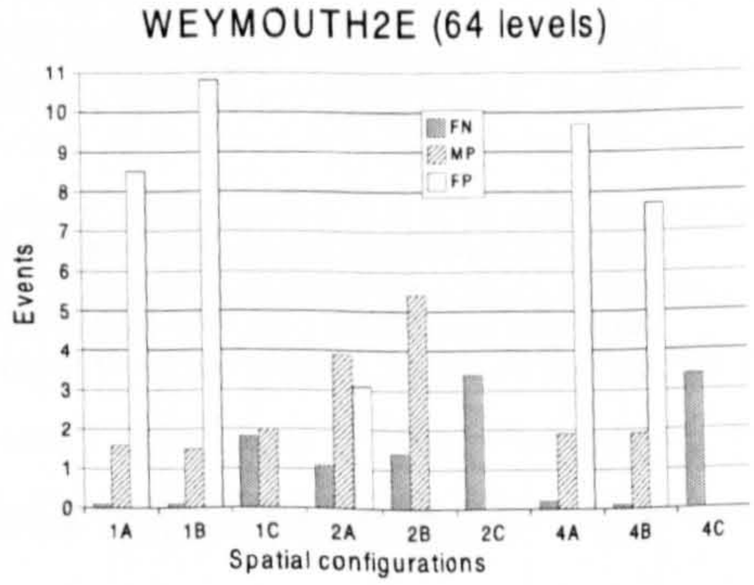
(a) sample frame



(b) 8 intensity levels



(c) 32 intensity levels



(d) 64 intensity levels

Figure 4.4: Evaluation of the co-occurrence matrix redundancy hypothesis using the segmentation described in Sections 4.3 - 4.7. Results shown for WEYMOUTH2E sequence: (a) sample frame, (b)-(d) average false negatives (FN), misclassified positives (MP) and false positives (FP) per sequence for various spatial configurations of the pixel pairs. Three intensity resolutions of the co-occurrence matrices are considered - 8, 32 and 64 levels.

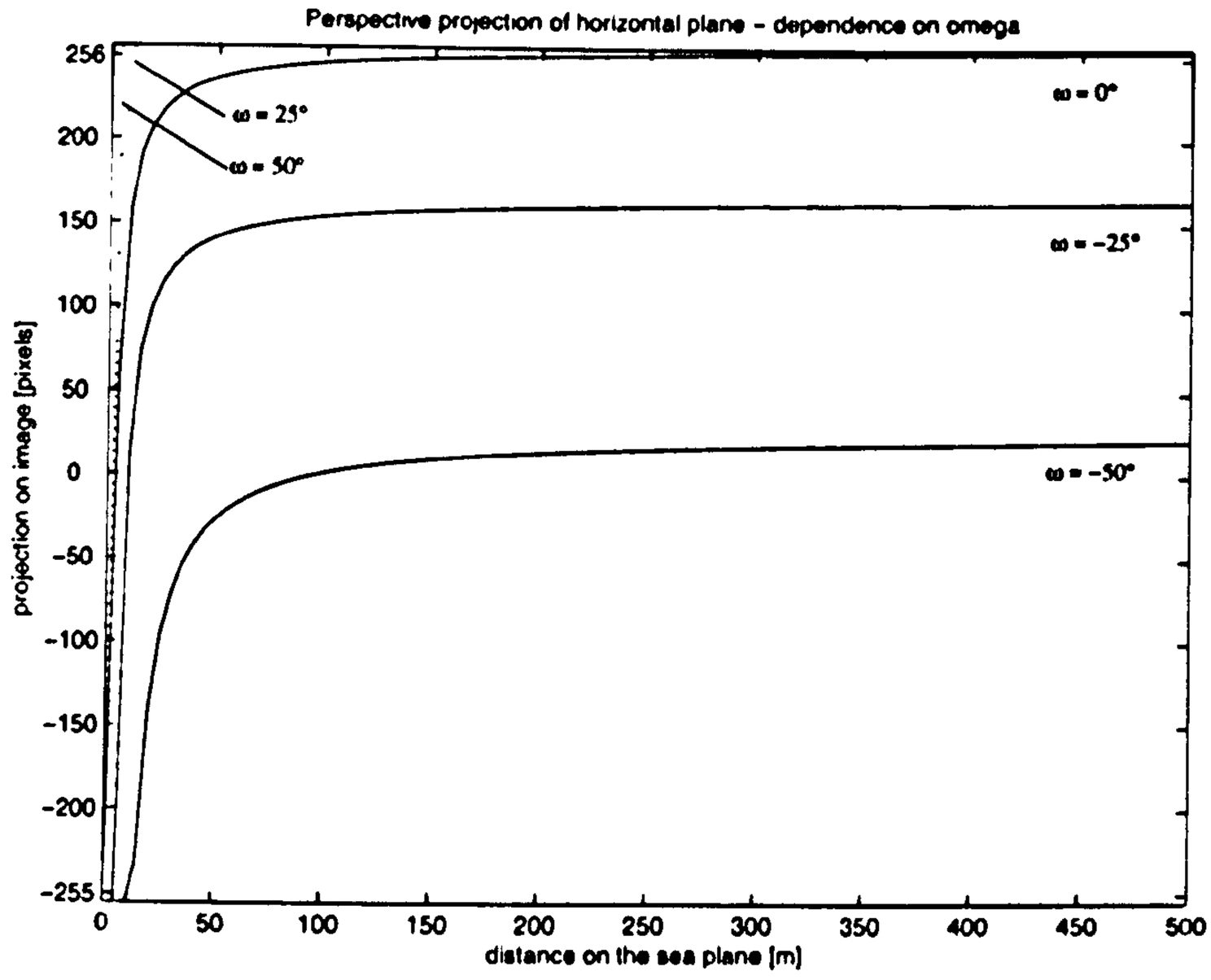
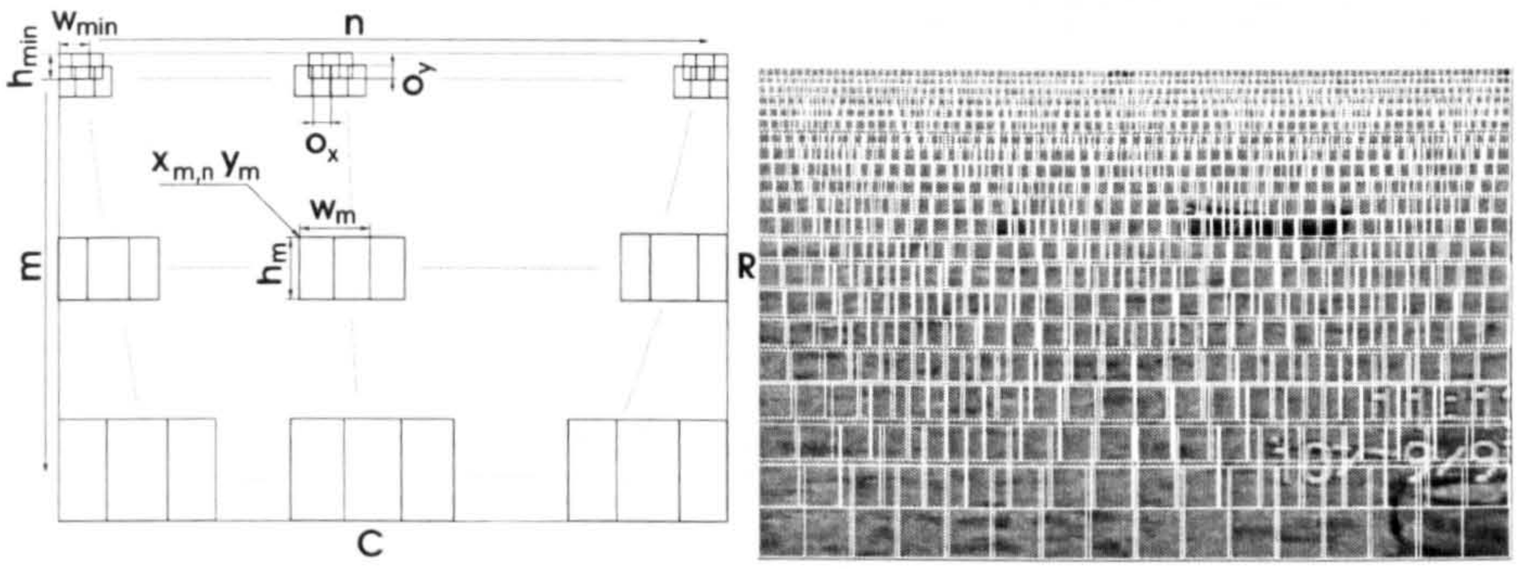


Figure 4.5: Perspective projection of the horizontal sea surface onto the image plane. Multiple projections are shown for varying camera pitch angle ω.

The conclusion from the experiment above is that the redundancy of a co-occurrence matrix in the segmentation process is justified.

4.3 Segmentation Geometry

Projection of maritime scenes is perspective which means that the scales of objects in the scene depend on their distances from the observation point. The objects are located on a planar surface representing the sea. Therefore, their vertical positions in the image change with their distances as well. Objects further from the camera project higher in the image. The scale of the scene gradually changes from bottom to top. The projection of the sea surface plane onto the image plane is shown in Figure 4.5. The setup is the same as in the Section 2.3. The horizontal axis shows the location of a point on the sea plane. The vertical axis shows the projected position of the same point in the image plane. The principal point is located at $(0, 0)$. Multiple curves correspond to various camera tilt angles ω.



(a) Structure of the segmentation grid.

(b) An example of optimised segmentation grid applied on a maritime scene.

Figure 4.6: The structure of the segmentation grid compensates for scale change due to perspective projection of maritime scenes (a). An example of the segmentation grid applied to a sample maritime scene (b).

The geometry of perspective projection places strong contextual constraint on the principles of the segmentation which must account for the fact that the scene being segmented is a perspective projection. Therefore the proposed segmentation does not operate on segments of the same scale across the whole image unlike other traditional segmentation algorithms that do not reflect the depth of the scene, (Pal, 1993). Here, the size of segment changes, depending on its position in the image. The change is monotonic reflecting the shape of the perspective projection profile (see Figure 4.5). The structure of the segmentation grid and an example of the grid applied to a maritime scene are shown in Figure 4.6.

The segments are smallest at the top edge of the image and their size increases towards the bottom of the image. Figure 4.7 shows a comparison of the perspective projection with the vertical position y_m of segments in the segmentation grid. The plot indicates the approximation of the change of scale in the segmentation grid structure. The segmentation is more gradual in the change of scale than the perspective projection. If the segmentation strictly adheres to the perspective profile, the segments near the top edge of the image would be too small to be analysed by texture features. The variable segmentation approaches the perspective projection while preserving the usability of the segments for textural analysis.

The position $(x_{m,n}, y_m)$ and size (w_m, h_m) of segments is given by the

following equations

$$x_{m,n} = n \cdot w_{min}(1 + \frac{\Delta_x}{100})^m(1 - \frac{o_x}{100}), m = 0, \ldots, m_{max} - 1; n = 0, \ldots, n_{m,max} - 1 \tag{4.5}$$

$$y_m = h_{min}(1 - \frac{o_y}{100})\frac{100}{\Delta_y}\left[1 - \left(1 + \frac{\Delta_y}{100}\right)^m\right], \Delta_y > 0; m = 0, \ldots, m_{max} - 1 \tag{4.6}$$

$$y_m = m \cdot h_{min}(1 - \frac{o_y}{100}), \Delta_y = 0; m = 0, \ldots, m_{max} - 1 \tag{4.7}$$

$$w_m = w_{min}(1 + \frac{\Delta_x}{100})^m \; m = 0, \ldots, m_{max} - 1 \tag{4.8}$$

$$h_m = h_{min}(1 + \frac{\Delta_y}{100})^m \; m = 0, \ldots, m_{max} - 1 \tag{4.9}$$

where w_{min}, h_{min} are the initial dimensions of the segments, Δ_x, Δ_y are the relative changes in segment sizes expressed in percent in both directions, o_x, o_y are the relative overlaps of segments expressed in percent in both directions, m, n are row and column indexes of the segments and $x_{m,n}$ indicates that the position depends on both indexes. The grid is grown from top (the smallest segments) to bottom. The following constraints apply for the index boundaries m_{max} and $n_{m,max}$

$$\exists m_{max} : \; y_{m_{max}-1} + h_{m_{max}-1} \leq R < y_{m_{max}} + h_{m_{max}} \tag{4.10}$$

$$\forall m = 0, \ldots, m_{max} - 1; \; \exists n_{m,max} : \; x_{m,n_{max}-1} + w_m \leq C < x_{m,n_{max}} + w_m \tag{4.11}$$

where $m_{max} - 1$ is the index of the last segment that fully fits into the image of height R, $n_{m,max} - 1$ is index of segment that fully fits into the image of width C. The extra index m indicates that the value of $n_{m,max}$ differs for each row of in the grid (see Figure 4.6a).

When applying the segmentation grid to an image, certain parts of the scene may be left uncovered. This occurs when the segments are overlaid horizontally or vertically and a gap is left uncovered at both ends. The size of this gap depends on the parameters of the segmentation. It is the same for all segments in a vertical direction, but varies for every row of segments in

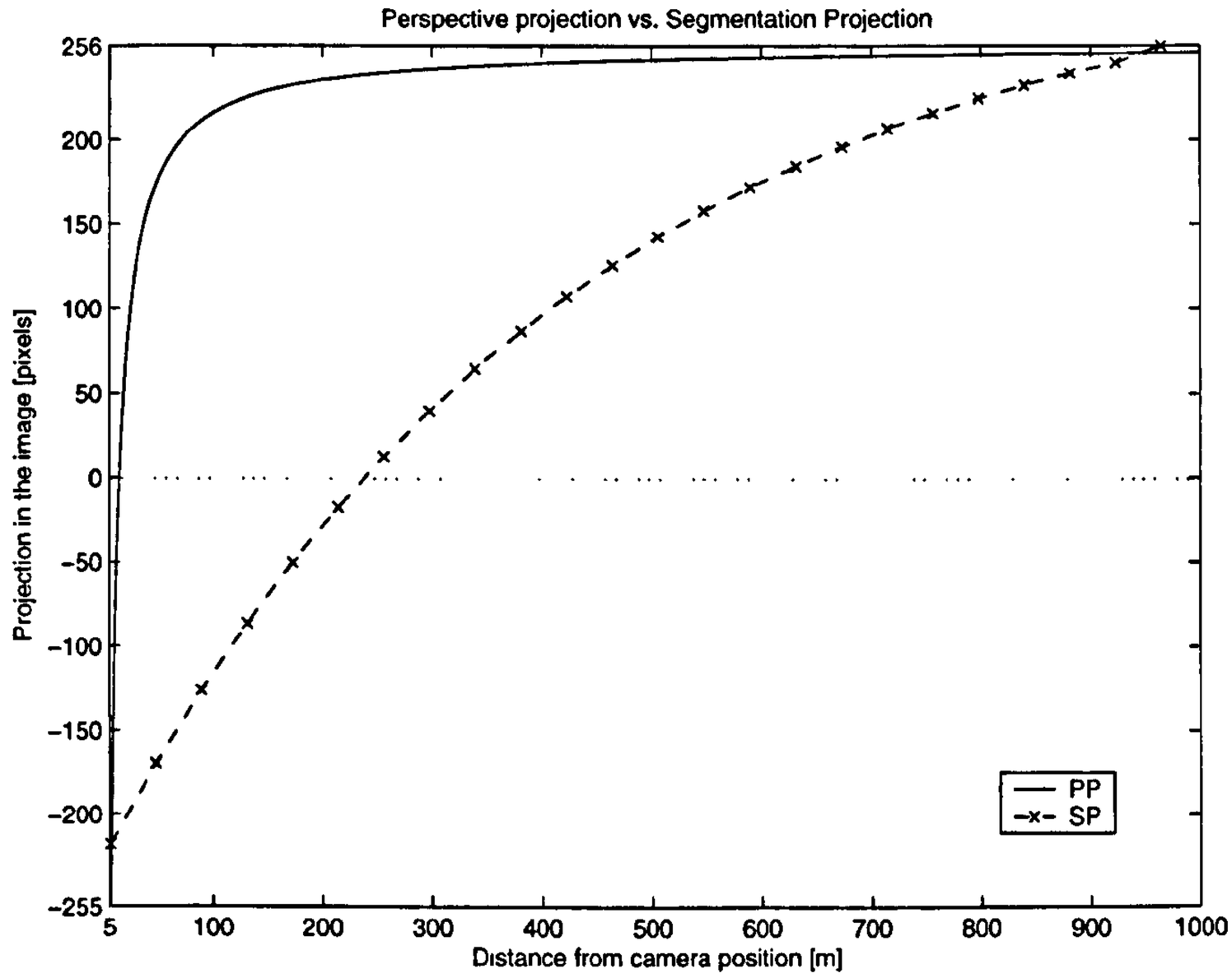


Figure 4.7: Comparison of perspective projection profile (PP) and vertical positions of segments in the segmentation grid (SP).

a horizontal direction. To fully cover the whole scene the layout is optimised by shifting every segment by a certain residual displacement in both directions. This corrected position $(\widetilde{x}(m,n), \widetilde{y}(m))$ of the upper-left corner of each segment can be expressed as

$$\widetilde{x}_{m,n} = x_{m,n} + \varepsilon_{m,x},\ n = 1, \ldots, n_{m,max} - 1; m = 0, \ldots, m_{max} - 1 \qquad (4.12)$$

$$\widetilde{y}_m = y_m + \varepsilon_y,\ m = 1, \ldots, m_{max} - 1 \qquad (4.13)$$

where $\varepsilon_{m,x}$ and ε_y are residual displacements in both directions that are determined from following relations

$$\varepsilon_{m,x} = \frac{W - (x_{m,n_{m,max}-1} + w_m)}{n_{m,max} - 1} \qquad (4.14)$$

$$\varepsilon_y = \frac{H - (y_{m_{max}-1} + h_{m_{max}-1})}{m_{max} - 1} \qquad (4.15)$$

where $\varepsilon_{m,x}$ indicates that the horizontal residual displacements vary for each row in the segmentation grid. Residual displacements are applied to all

segments in both directions except those in the first row where vertical displacement is inapplicable and the first column where horizontal displacement is inapplicable. Such a corrected segmentation grid covers the whole scene completely (see Figure 4.6).

4.4 Calculation of Features

Once the scene is subdivided into the segments by the variable windows grid the search for objects within the segments is initiated. Each segment is regarded as a sample of a texture that is characterised by features defined by Equations 4.1 - 4.4. Features are grouped to form a four-element vector that uniquely characterises each segment. The search for object is based on an assumption that there is a substantial difference between vectors characterising the segments with objects and vectors that characterise segments with sea only.

4.4.1 Intensity Unbiasing

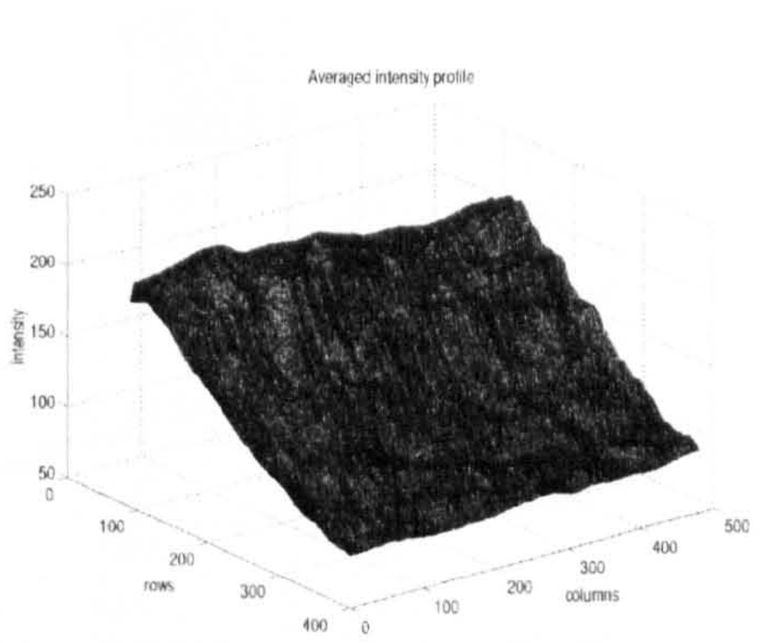
Illumination of the sea surface produces various effects under different environmental conditions. This is illustrated in Figure 4.8 that shows an average intensity profile of a typical maritime scene. The average intensity profile was obtained by convolving the original image with a 31×31 pixel averaging mask. The intensity values slope towards the bottom of the image. As intensity data are used to calculate features directly it is necessary to compensate for illumination offsets, otherwise features will become biased hindering the detection of objects.

The unbiased values of $f(r,c)$ in Equations 4.1, 4.2, 4.3 and 4.4 are defined by

$$f(r,c) = \left| f'(r,c) - \bar{f} \right| \tag{4.16}$$

where $f'(r,c)$ is the original intensity at position (r,c) and $\bar{f}$ is the offset value determined as a constant for each segment in the grid. The absolute value is necessary in order to determine the entropy feature (Equation 4.2). The value of $\bar{f}$ is either mean or median of all intensity values in a single grid segment. The difference between mean and median is analysed in Section 4.9.2.

(a) sample image



(b) averaged intensity profile

Figure 4.8: An example of intensity biasing in maritime scenes.

4.4.2 Segment Resizing

It is evident from each row of Table 4.1 that the values of features depend on the scale of the texture. The dependency is also apparent from Equations 4.1 - 4.4 with summations done across segments of different areas. This would introduce a systematic bias to the feature vectors which would complicate the search for vectors representing the objects.

One solution is to normalise the values by the area of the segment. This is possible for energy and entropy as the sum values are directly proportional to the area. The same, however, does not hold for homogeneity and contrast. The values of these features would remain significantly biased.

Another solution is to resize the segments to a single initial scale, $w_{min} \times h_{min}$. Details in larger segments are reduced and textures are smoothed. An alternative is to expand segments to their final scale, $w_{m_{max}} \times h_{m_{max}}$. This is inefficient for two reasons - the amount of data to process increases and the new intensities result from extrapolation of values already available thus no new information from the scene is obtained. Even though reduction of the scale reduces the details in the image by leaving out pixels in larger segments it reduces the amount of noise that is due to the appearance of the sea.

An optional method that compensates for the perspective projection of the scene is used in traffic applications (Kastrinaki et al., 2003), namely automated navigation of cars (Broggi and Berte, 1995). The method consists of an image transform that reverts the original perspective projection. The scale reduction is done in finer steps as each pixel is transformed instead of

Scales		100%	80%	60%	40%	20%
Orig.:	Eng	$2.65 \cdot 10^6$	$1.7 \cdot 10^6$	954275	431372	105622
	Ent	510421	328037	183212	82557.4	20196.5
	Hom	10063.5	7695.56	5319.93	3259.66	1288.03
	Con	$8.56 \cdot 10^8$	$3.5 \cdot 10^8$	$1.1 \cdot 10^8$	$2.2 \cdot 10^7$	$1.4 \cdot 10^6$
Norm.:	Eng	120.3	120.7	120.5	121.9	121.4
	Ent	23.1	23.2	23.1	23.3	23.2
	Hom	0.46	0.54	0.67	0.92	1.48
	Con	38860	24874	13972	6249	1555
Scaled:	Eng	105622	116749	99632.3	117741	105622
	Ent	20196.5	21743.9	19509.8	21773.5	20196.5
	Hom	1288.03	1385.17	1264.27	1383.04	1288.03
	Con	$1.4 \cdot 10^6$	$1.4 \cdot 10^6$	$1.3 \cdot 10^6$	$1.4 \cdot 10^6$	$1.4 \cdot 10^6$

Table 4.1: Linearisation of the features by scaling.

larger segments, so the resulting image is better compensated for perspective distortions. Because the rectilinear grid of the image is not projected to the same rectilinear grid, intensity extrapolation is needed and the resulting image takes the shape of an isosceles trapezoid, (Wolberg, 1990). Segmentation of such an image would have to take this into account by either cropping edges of the image to obtain a rectangular region of interest or applying a non-rectangular segmentation grid. Positions of objects detected in the transformed image would have to be transformed back in order to obtain the locations of objects in the original image. In contrast to that, the variable windows segmentation compensates implicitly for the perspective distortion without requiring to change the geometry of the segmented image.

Two rescaling methods are considered: re-sampling of the image and bilinear interpolation. Re-sampling reduces the scale of the segment by simply dropping out pixels at intermediate positions. No interpolation is necessary, re-sampling is simple and fast. The result, however, suffers from discontinuities at ramp edges that may introduce high frequency noise. Bilinear interpolation provides smoother results by linear interpolation of the intermediate values, it is, however, more complex.

Finally, the features determined for each unbiased and rescaled segment are arranged into vectors. Each vector represents a point in so-called feature space (Jain et al., 2000) (Figure 4.9). The segmentation of objects continues by partitioning and analysis of the feature space.

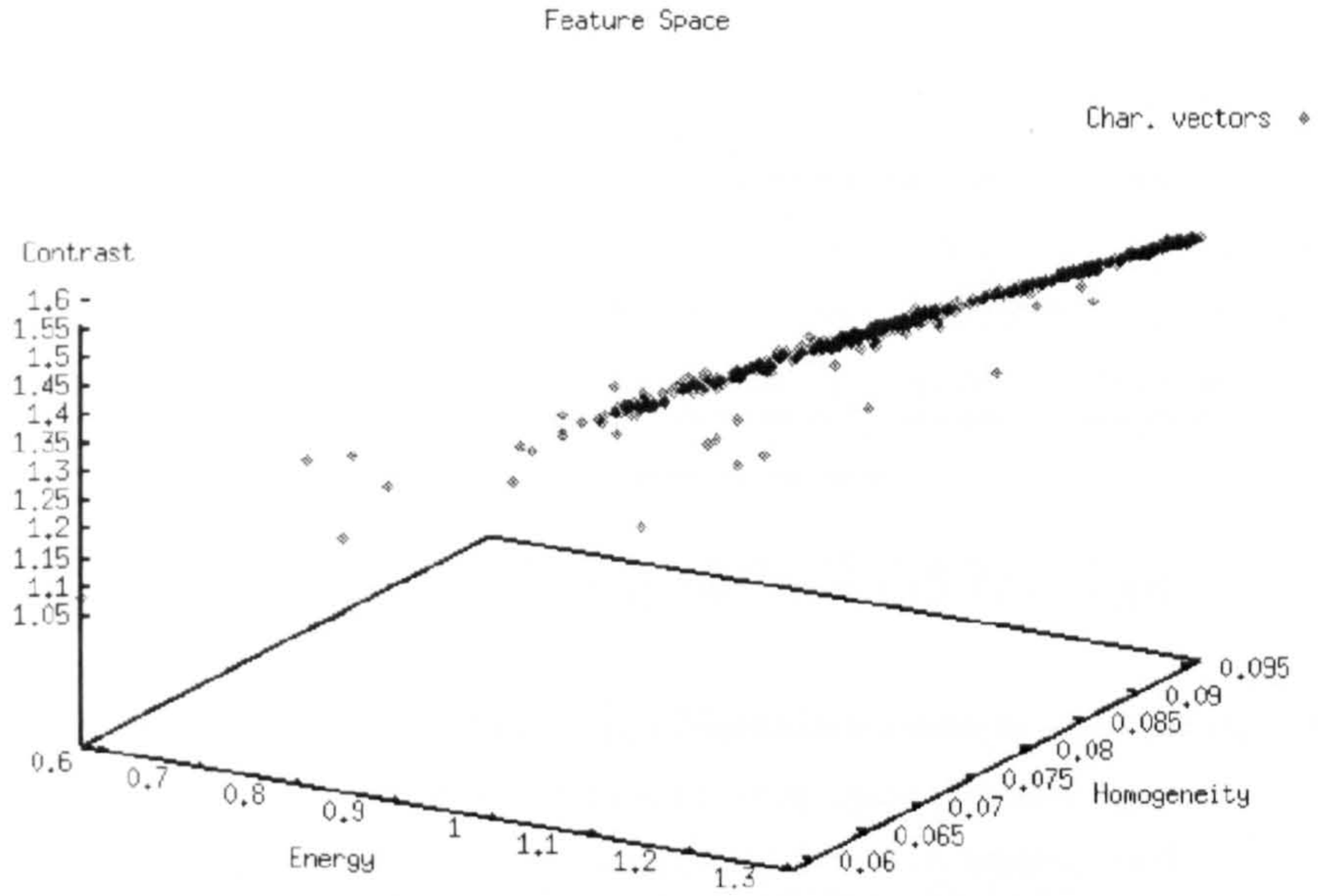


Figure 4.9: Feature space containing feature vectors describing the grid segments. The assumption is that outliers represent segments with probable objects in the scene. Only three coordinates are considered for illustrational purposes.

4.5 Partitioning of the Feature Space

The next step in the process of segmentation is the partitioning of the feature space into a main class cluster and outliers. The main class cluster contains vectors representing the most prominent structure in the image which is the sea while outliers correspond to segments with probable objects. This is in accordance with the assumption declared in Section 2.6 that states that objects occupy only minority of the scene. The partitioning produces a boundary maintained in the feature space that separates the main cluster from the outliers.

The partitioning of the feature space is a typical problem of pattern classification. The partitioning can be done in numerous ways (Jain et al., 2000) and the choice depends on the specific task. The categorisation of these methods is shown in Figure 4.10. The amount of information available about the data to be classified decreases from left to right in the tree. The choice of method depends primarily on the knowledge of class-conditional probability densities. Class-conditional probability states the probability of a feature having a certain value for a certain class. As this is not usually known

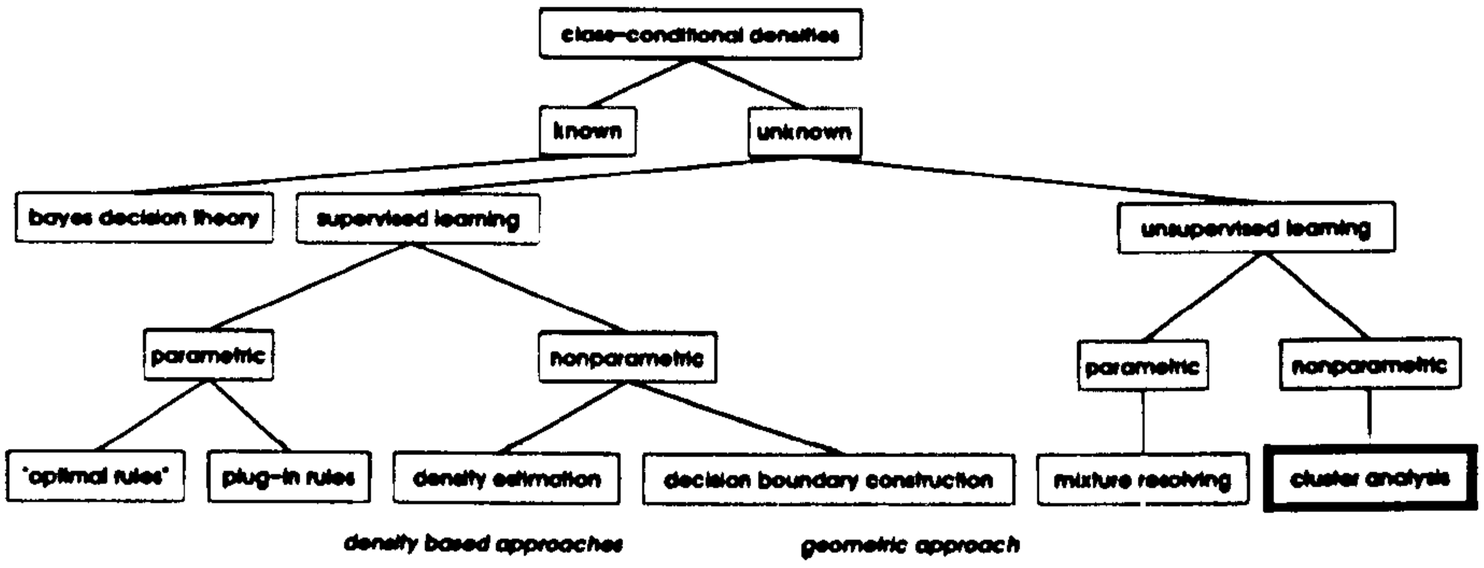


Figure 4.10: Commonly used classification methods, (Jain et al., 2000).

in advance in most real world applications, a majority of classifiers have to undergo learning stage prior to actual classification.

The method of learning depends on the availability of labelled data from which the classification rules can be inferred. If a representative set of labelled data for training is not available unsupervised classification methods such as cluster analysis should be considered. Cluster analysis is a valuable technique that groups and partitions a feature space into different class representatives. Widely used criterion in cluster analysis is based on iterative minimisation of square-error (Jain et al., 2000).

The objective of iterative square-error clustering is to obtain partitioning of the feature space that minimises the overall square error. If there are K classes then for every cluster C_k, $k = 1, \ldots, K$ representing a single class the centroid $\mu^{(k)}$ is defined as

$$\mu^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_i^{(k)} \tag{4.17}$$

where $\mathbf{x}_i^{(k)}$ is the i-th feature vector belonging to cluster C_k and n_k is the number of feature vectors belonging to cluster C_k. The square-error e_k^2 for cluster C_k is the sum of the squared Mahalanobis distances between each feature vector in C_k and its centroid $\mu^{(k)}$

$$e_k^2 = \sum_{i=1}^{n_k} (\mathbf{x}_i^{(k)} - \mu^{(k)}) \Lambda^{(k)} (\mathbf{x}_i^{(k)} - \mu^{(k)})^T \tag{4.18}$$

where $\Lambda^{(k)}$ is the inverse covariance matrix of the data in the cluster C_k. Equation 4.18 becomes a sum of Euclidean distances, in case the covariance matrix is an identity matrix, $\Lambda^{(k)} = \mathbf{I}$. The overall square error E_K^2 of the

clustering is the sum of all intra-class errors

$$E_K^2 = \Sigma_{k=1}^{K} e_k^2 \tag{4.19}$$

The minimisation of E_K^2 is generally achieved by an iterative process of repartitioning and re-evaluation of the criterion function. The iterative process also known as the K-means algorithm consists of the following steps (Jain et al., 2000):

1. Select an initial partition with K clusters. Repeat steps 2 to 4 until the the cluster membership stabilises.

2. Generate a new partition by assigning each pattern to its closest cluster centre.

3. Compute new cluster centres as the centroids of the clusters.

4. Repeat steps 2 and 3 until an optimum value of the criterion function is found.

Non-parametric clustering is a popular method of texture segmentation in natural scenes (Fauzi and Lewis, 2003; Pauwels and Frederix, 1999) for which labelled data are often difficult to obtain. Segmentation of maritime scenes is assumed to belong to the same category of problem. The clustering method introduced in the next section therefore follows the principle of the K-means algorithm.

4.5.1 Iterative Clustering

The standard K-means algorithm assumes that there are $K > 1$ clusters in the feature space. Considering the maritime scene with no objects in the scene there is only a single class representing the sea. When objects enter the scene their textures are characterised by features spanning additional classes. The number of these classes is unknown as there can be any number of objects in the scene constituting of any number of textures.

The unknown number of classes can be estimated from the density of points in feature space, (Pauwels and Frederix, 2000). These methods, however, require that the clusters do not contain less than a given minimum of points. The requirement often does not hold in maritime scenes. Maritime objects occupy a minority of the scene and, therefore, only a limited number of segments in

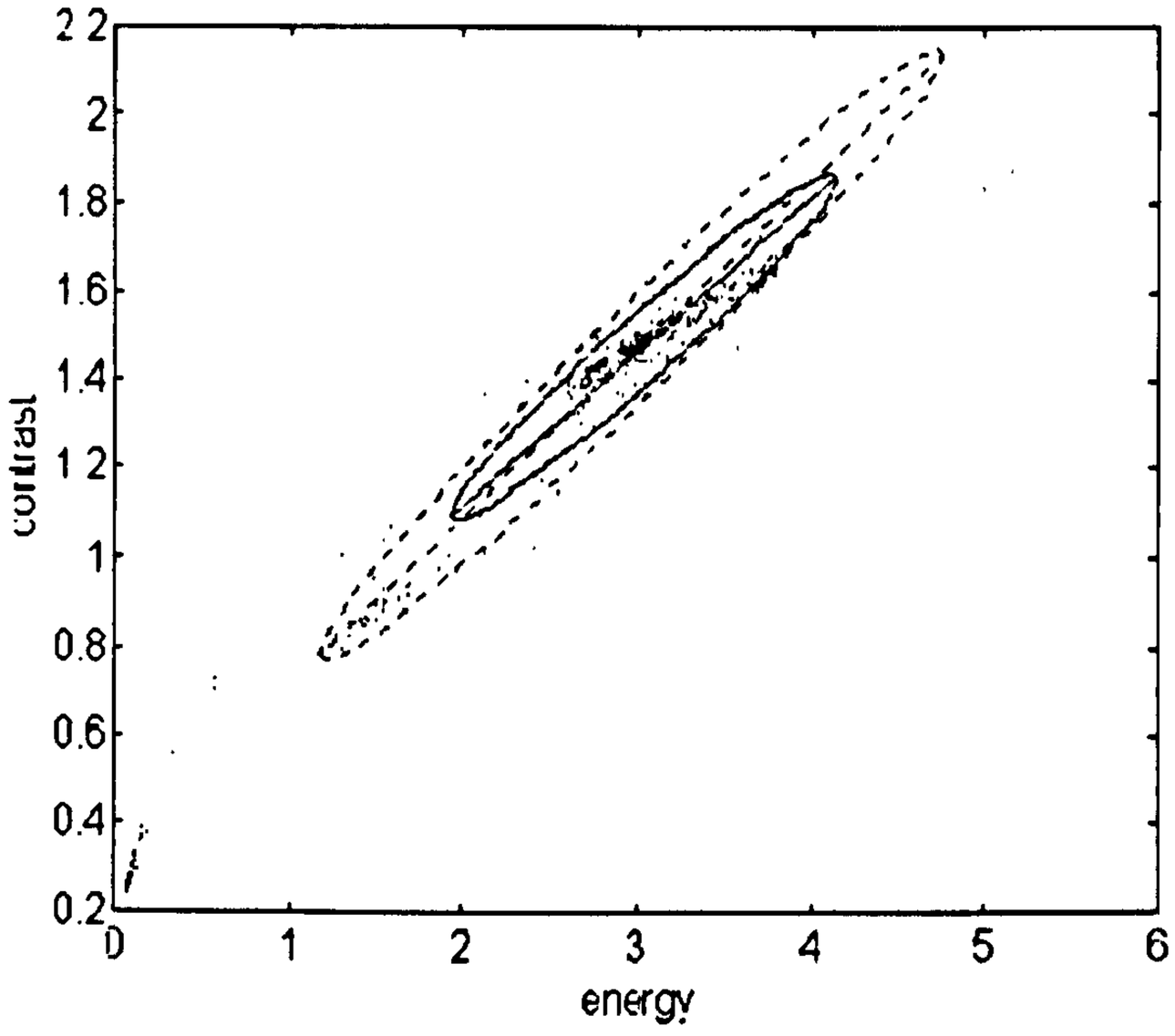


Figure 4.11: Two iterations of the centroid estimation procedure (two-dimensional projection). The first subset is inside a dashed ellipse, the second set is inside a solid ellipse.

the segmentation grid. This implies a limited number of points in the feature space that correspond to the objects. In addition, there is no guarantee that a compact cluster representing the objects in the scene will form as these objects can be composed of several parts with varying textures.

An alternative approach of the feature space partitioning is proposed. The centroid of the main cluster in the feature space is iteratively located and a decision boundary enclosing the main cluster is constructed. The points in the main cluster correspond to the sea and all points outside the boundary represent objects. The procedure is outlined as Algorithm 1. Two iterations are illustrated on the 2D projection of the data in Figure 4.11. The subsets are outlined as ellipses.

4.5.2 Optimal Number of Iterations

The population G_i of points in the subset used in centroid estimation reduces at each iteration step. The outliers are excluded from the subset, further

Algorithm 1 The iterative procedure of main cluster centroid estimation.

1. The position of the centroid $\mu^{(i)}$ in the i-th iteration is determined using points in a subset of the feature space

$$\mu^{(i)} = \frac{1}{G_i} \sum_{j=1}^{G_i} \mathbf{x}_j^{(i)} \tag{4.20}$$

where G_i is the number of points within the subset and $\mathbf{x}_j^{(i)}$ is the j-th feature point in the subset. Initially, $i = 0$ and $\mu^{(0)}$ is determined using a subset that contains all points G_0 in the feature space.

2. The average distance $D^{(i)}$ between points $\mathbf{x}_j^{(i)}$, $j = 1, \ldots, G_i$ in the subset and their centroid $\mu^{(i)}$ is determined as

$$D^{(i)} = \frac{1}{G_i} \sum_{j=1}^{G_i} d(\mathbf{x}_j^{(i)}, \mu^{(i)}) \tag{4.21}$$

where $d(\mathbf{x}_j^{(i)}, \mu^{(i)})$ is the Mahalanobis distance defined as

$$d(\mathbf{x}_j^{(i)}, \mu^{(i)}) = \sqrt{(\mathbf{x}_j^{(i)} - \mu^{(i)}) \cdot \mathbf{\Lambda}^{(i)} \cdot \left(\mathbf{x}_j^{(i)} - \mu^{(i)}\right)^T} \tag{4.22}$$

where $\mathbf{x}_j^{(i)}$ is the j-th feature point in the subset, $\mu^{(i)}$ is the centroid and $\mathbf{\Lambda}^{(i)}$ is the inverse of the covariance matrix estimated from the subset.

3. A new subset is generated containing points within the average distance $D^{(i)}$ from the centroid $\mu^{(i)}$.

4. $i = i + 1$ and steps 1 to 3 are repeated.

5. The procedure is stopped when i reaches predefined value.

improving centroid estimation. However, after a certain number of iterations the subset becomes too small and the information about the shape of the cluster is lost.

This can be illustrated by analysing the normalised eigenvectors of the covariance matrix for feature points within the subset used in centroid estimation. Eigenvectors of the covariance matrix are aligned with the major variance axes of the feature points (Dunteman, 1989; Iversen and Norpoth, 1987). As all points are involved at the initial iteration step the outliers will influence the covariances and consequently the orientation of eigenvectors. As the population of subset reduces, so does the influence of outliers on the covariance and the orientation of the eigenvectors will change. After a couple of steps the shape of the main cluster is well captured by the covariance matrix and the orientation of eigenvectors does not change substantially. However, if more steps follow, this orientation equilibrium is broken and eigenvectors start to shift again. This is undesirable as the resulting covariance matrix ceases to encode the shape of the main cluster causing the segmentation results to deteriorate.

Table 4.2 shows the evolution of orientations of normalised eigenvectors of the covariance matrix (two cases are illustrated - with and without the objects in the scene). The values in Table 4.2 are defined as $1 - \cos\alpha_j$, where α_j is the angle between j-th normalised eigenvectors at iteration steps i and $i+1$. This angle can be obtained from vector dot products:

$$\mathbf{v}_j(i) \cdot \mathbf{v}_j(i+1) = |\mathbf{v}_j(i)|\,|\mathbf{v}_j(i+1)|\cos\alpha_j \tag{4.23}$$

Both normalised eigenvectors have a length equal to one, therefore the dot product gives directly the values of $\cos\alpha_j$. It is clear from the definition of vector dot products that parallel vectors have α_i equal to zero. Change in orientation causes α_i to increase.

The values in Table 4.2 show that within the first two or three iteration steps there is an initial alignment of the covariance matrix eigenspace with the orientation of the main cluster. With more iterations the change of orientation is minimal but after five or six iterations the eigenspace stops following the orientation of the main cluster and starts to change the orientation again. Values greater than one in the Table 4.2 indicate an initial value of $\alpha_i > \frac{\pi}{2}$, ($\cos\alpha_i < 0$). Figure 4.12 shows the evolution of the position of the centroid over the iterations. The position changes linearly up to approximately five

Iterations	1	2	3	4	5
	No objects in the scene				
v_1	$0.48 \cdot 10^{-2}$	$0.12 \cdot 10^{-2}$	$0.15 \cdot 10^{-3}$	$0.27 \cdot 10^{-3}$	$0.83 \cdot 10^{-4}$
v_2	$0.15 \cdot 10^{-2}$	$0.16 \cdot 10^{-2}$	$0.18 \cdot 10^{-3}$	$0.23 \cdot 10^{-3}$	$0.24 \cdot 10^{-4}$
v_3	$0.56 \cdot 10^{-3}$	$0.59 \cdot 10^{-3}$	$0.55 \cdot 10^{-4}$	$0.59 \cdot 10^{-4}$	$0.16 \cdot 10^{-3}$
v_4	$0.43 \cdot 10^{-2}$	$0.37 \cdot 10^{-3}$	$0.37 \cdot 10^{-4}$	$0.13 \cdot 10^{-4}$	$0.89 \cdot 10^{-5}$
	Objects present				
v_1	0.68	$0.23 \cdot 10^{-2}$	$0.56 \cdot 10^{-3}$	$0.16 \cdot 10^{-3}$	$0.54 \cdot 10^{-4}$
v_2	$0.36 \cdot 10^{-1}$	$0.35 \cdot 10^{-2}$	$0.95 \cdot 10^{-3}$	$0.31 \cdot 10^{-3}$	$0.15 \cdot 10^{-3}$
v_3	$0.10 \cdot 10^{-1}$	$0.13 \cdot 10^{-2}$	$0.43 \cdot 10^{-3}$	$0.17 \cdot 10^{-3}$	$0.97 \cdot 10^{-4}$
v_4	1.30	$0.30 \cdot 10^{-3}$	$0.84 \cdot 10^{-4}$	$0.23 \cdot 10^{-4}$	$0.13 \cdot 10^{-4}$
Iterations	**6**	**7**	**8**	**9**	**10**
	No objects in the scene				
v_1	$0.20 \cdot 10^{-4}$	$0.98 \cdot 10^{-4}$	$0.14 \cdot 10^{-2}$	0.16	
v_2	$0.13 \cdot 10^{-3}$	$0.27 \cdot 10^{-4}$	$0.35 \cdot 10^{-2}$	1.3	
v_3	$0.13 \cdot 10^{-3}$	$0.18 \cdot 10^{-4}$	$0.40 \cdot 10^{-2}$	0.54	
v_4	$0.10 \cdot 10^{-4}$	$0.11 \cdot 10^{-4}$	$0.45 \cdot 10^{-3}$	$0.45 \cdot 10^{-3}$	
	Objects present				
v_1	$0.50 \cdot 10^{-4}$	$0.1 \cdot 10^{-3}$	$0.48 \cdot 10^{-3}$	$0.44 \cdot 10^{-3}$	
v_2	$0.82 \cdot 10^{-5}$	$0.20 \cdot 10^{-4}$	$0.36 \cdot 10^{-2}$	$0.21 \cdot 10^{-2}$	
v_3	$0.57 \cdot 10^{-5}$	$0.19 \cdot 10^{-4}$	$0.39 \cdot 10^{-2}$	$0.22 \cdot 10^{-2}$	
v_4	$0.51 \cdot 10^{-4}$	$0.10 \cdot 10^{-3}$	$0.12 \cdot 10^{-3}$	$0.20 \cdot 10^{-3}$	

Table 4.2: Evolution of covariance matrix eigenvector orientations.The values represent dot products between consecutive corresponding eigenvectors.

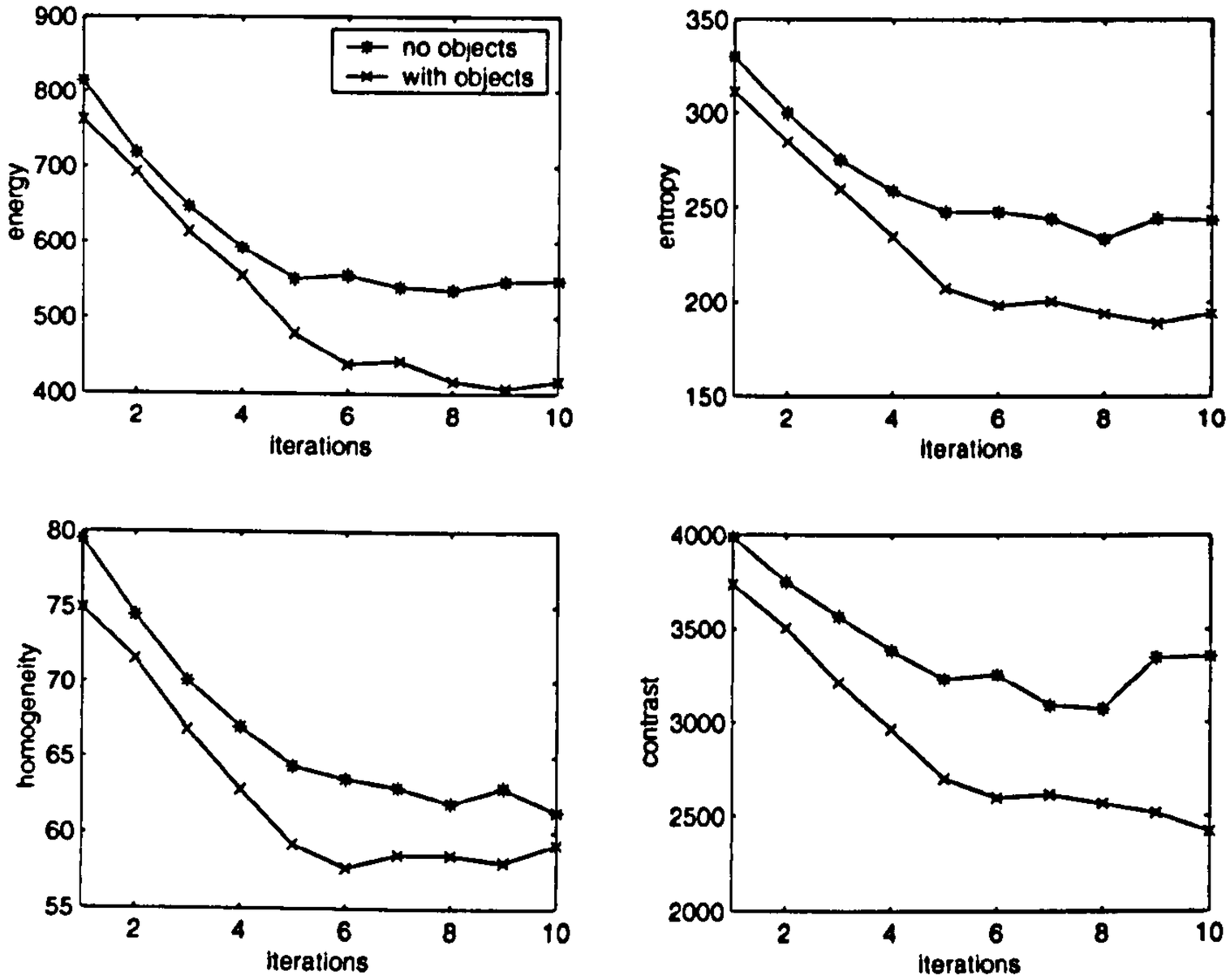


Figure 4.12: Evolution of centroid position with respect to iterative clustering steps. After approximately five iterations the position does not change significantly.

iterations but starts to oscillate after five iterations.

Taking into account the results in Table 4.2 and Figure 4.12 it can be concluded that the results start to deteriorate after four iterations as the data covariance matrix no longer characterises the structure of the main cluster and the centroid starts to fluctuate.

4.6 Adaptive Thresholding

The classification of outliers is based on a thresholding scheme that establishes and maintains a boundary surrounding the main class. The boundary is set up at a Mahalanobis distance D_b from the centroid μ of the main cluster estimated in the final iteration of the Algorithm 1.

Because the shape of the main cluster generally varies from scene to scene there is no single optimal value of D_b that would work for all the scenes. The threshold has to be established individually for every scene. Also, temporal

illumination changes in a single scene can influence the main cluster shape and thus the relative positions of outliers can change over time. All these factors should be considered when determining the threshold to provide time-consistent thresholding.

4.6.1 Distance Histograms

A modification of Mahalanobis distance between all feature points $\mathbf{x}_j$ and the centroid μ is introduced:

$$d_l(j) \equiv d_l(\mathbf{x}_j, \mu) = \frac{1}{2}\log((\mathbf{x}_j - \mu)\Lambda(\mathbf{x}_j - \mu)^T) \tag{4.24}$$

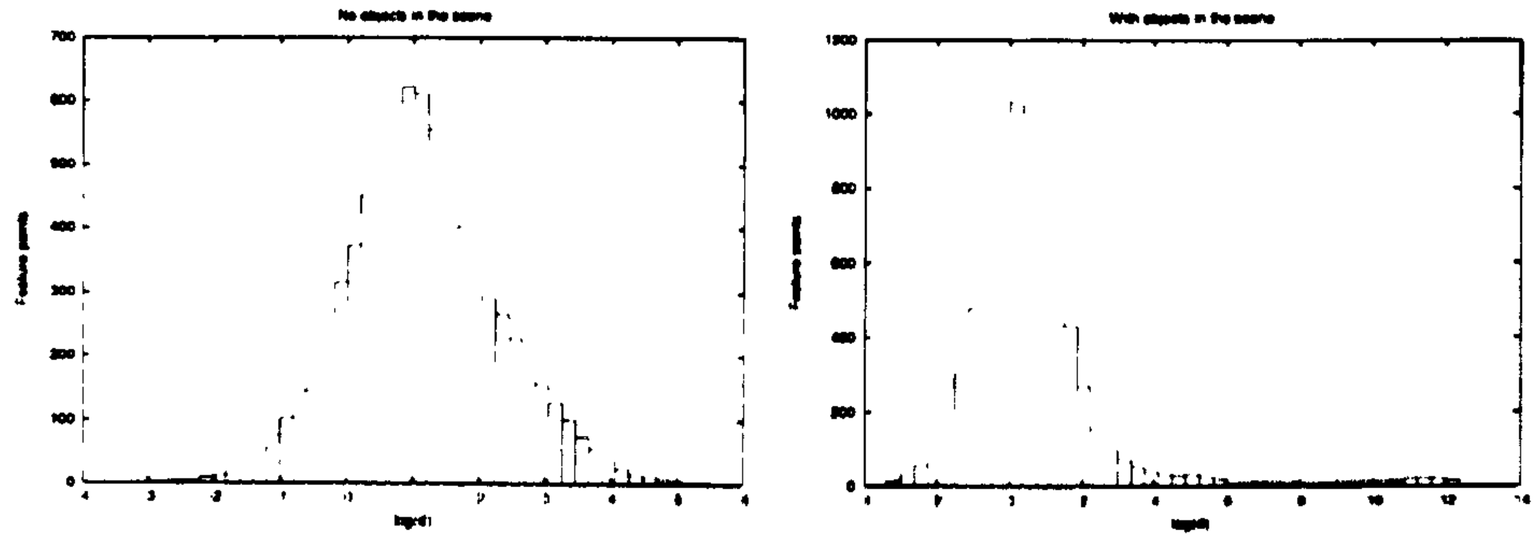
where Λ is the inverse covariance matrix obtained at the final iteration. The logarithm regularises extreme values of the distance that occur with outliers. The distances $d_l(j)$ can be regarded as samples drawn from two distinct statistical distributions. Figure 4.13 shows histograms of the distances for a typical maritime scene without and with objects present. The data for the histograms are collected over multiple frames and the histograms are smoothed by a moving average window filter 3 samples wide. Two distributions can be distinguished in Figure 4.13b - the major one on the left represents the background feature points while the minor distribution on the right corresponds to the feature points representing the objects. The distributions are separated by defining a threshold in the histogram. The search for the threshold is detailed in the following sections.

4.6.2 Distance Unbiasing

The image intensity unbiased values $f(r,c)$ obtained from Equation 4.16 have larger variances due to the prevalence of details in larger segments of the segmentation grid. The values of features obtained from Equations 4.1-4.4 and, consequently, the distances are larger for these segments. It is necessary to correct the bias prior to the construction of the histogram.

A partial correction of the bias is achieved by modifying each distance by subtracting a weighted mean

$$\tilde{d}_l(m,n) = d_l(m,n) - \frac{h_m}{h_{m_{max}}}\mu_1;\ n = 1,\ldots,n_{max}(m);\ m = 1,\ldots,m_{max} \tag{4.25}$$



(a) no objects

(b) with objects

Figure 4.13: Histograms of distances for scenes (a) without and (b) with objects. A secondary distribution in (b) on the right represents outliers in the feature space.

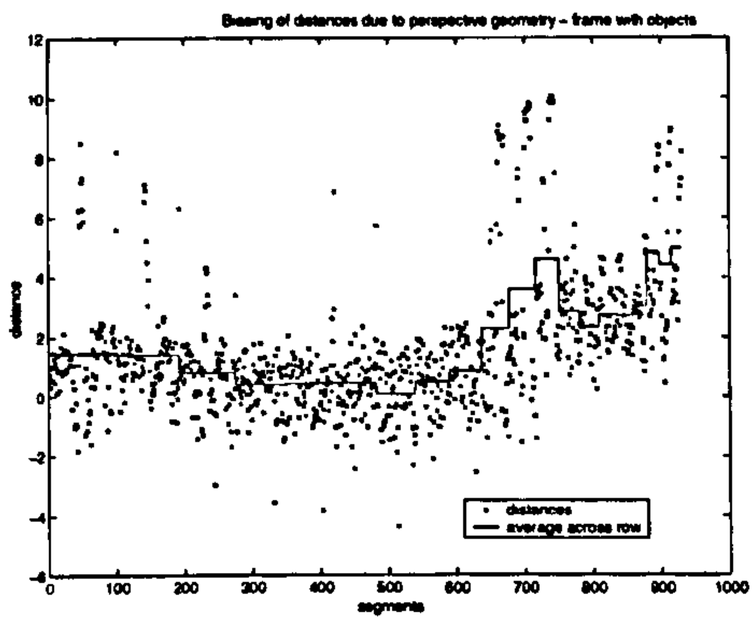
where $\tilde{d}_l(m,n)$ is the unbiased distance of feature point $\mathbf{x}_{m,n}$ for a segment at position (m,n) in the segmentation grid, $d_l(m,n)$ is the original distance for the same segment, μ_l is the average distance, h_m is the height of segments in a row m and $h_{m_{max}}$ is the height of segments in the last row of the segmentation grid. Indexing of the distances and feature points changes from a single index j to a pair of indexes m,n as the unbiasing value depends on the vertical index m of the segment in the grid.

Figures 4.14a,b show the original distances and their mean values across each row. Figures 4.14c,d show the original and the corrected mean values across each row together with the difference in these values.

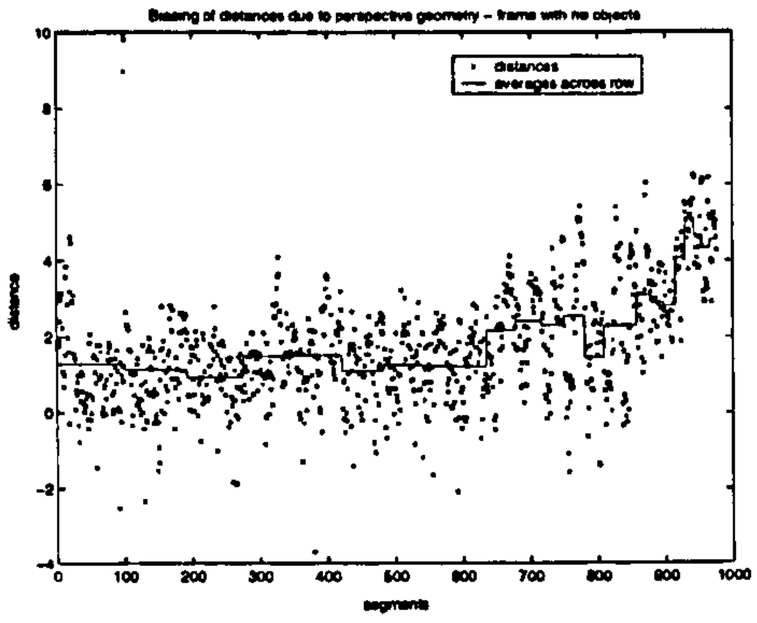
4.6.3 Threshold Selection

Histogram analysis methods are common in image processing as they provide straight-forward mechanism of separation of classes represented by modes in the histogram (Sezgin and Sankur, 2004). The analysis establishes thresholds between the modes usually by directly searching for the peaks and valleys in the histogram outline or by fitting parameterised curves such as Gaussians to the outline.

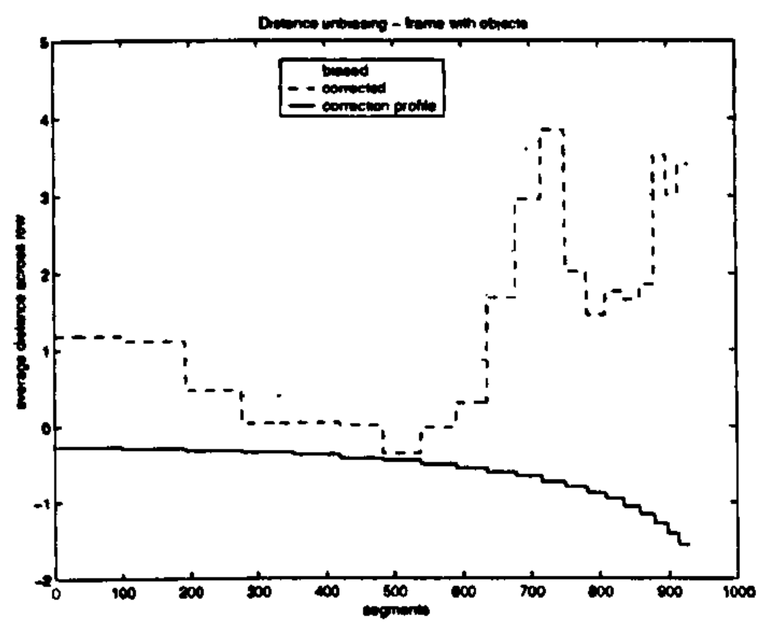
A minimum requirement for most methods is that the number of modes is known in advance. A popular thresholding method by Otsu (1979), for example, assumes a bimodal histogram. This is in contrast with the distance



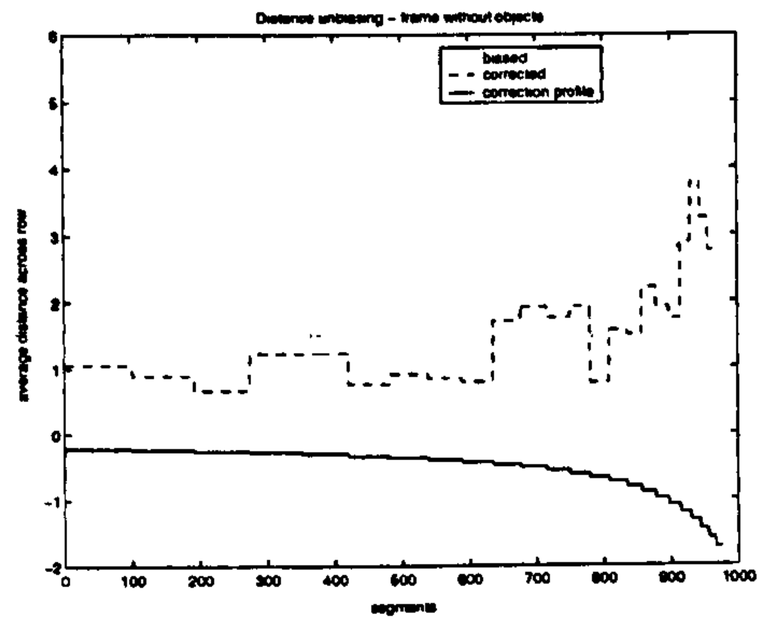
(a) effect of biasing - objects in the scene



(b) effects of biasing - no objects in the scene



(c) unbiasing - frame with objects



(d) unbiasing - frame without objects

Figure 4.14: Unbiasing of the distances. Distances are biased towards the bottom edge of the image due to suboptimal approximation of perspective projection in segmentation step.

histograms introduced above as their number of modes varies with the presence or absence of objects in the scene.

The histograms in Figures 4.13a,b show that the major mode is always present. By fitting a parametric curve to the outline of the main mode the presence of the secondary distribution can be established from a discrepancy between the fitted curve and the outline. The secondary distribution is always on the right side of the main mode as the feature points corresponding to objects lie further from the main cluster centroid.

McLaughlin (1999) provides an exhaustive list of probability distribution functions as well as definitions of their parameters. Figures 4.13a,b indicate that there is an apparent skew of the histogram with the right tail elongated. The Generalised Logistic Distribution (GLD) has been selected as an adequate approximation of the histogram outline. The GLD can be expressed as

$$PDF_{GL}(x) = \frac{C}{B} \frac{e^{\left(\frac{A-x}{B}\right)}}{\left[1 + e^{\left(\frac{A-x}{B}\right)}\right]^{C+1}} \tag{4.26}$$

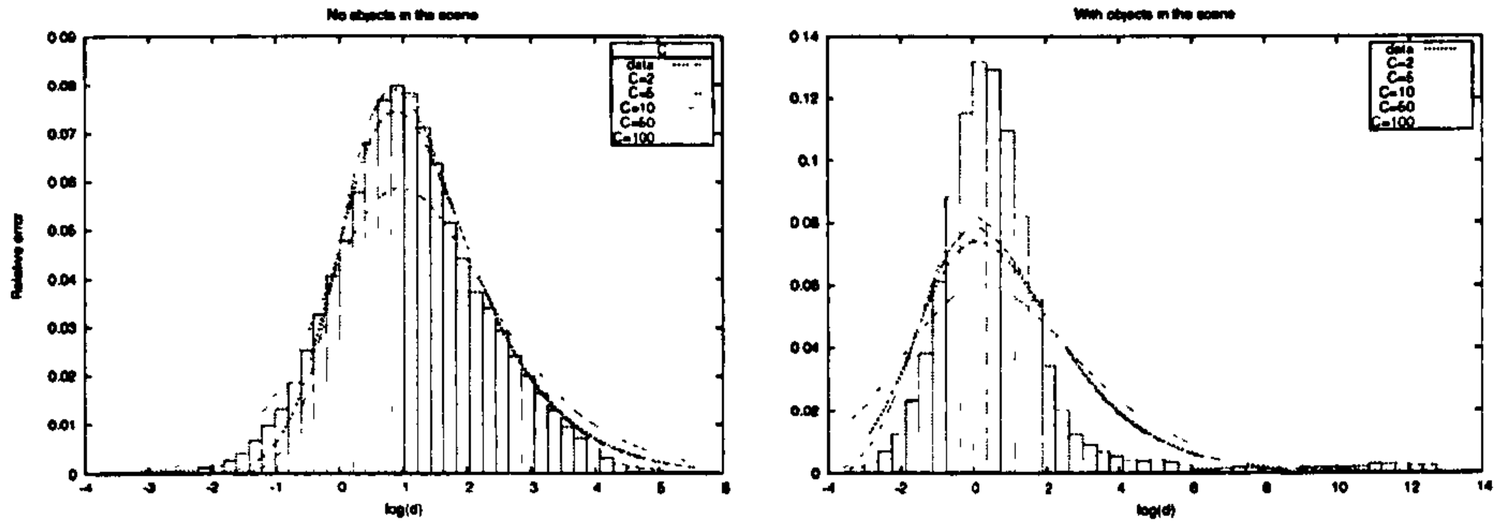
where parameters A, B are determined from following definitions relating to statistical moments calculated directly from the data

$$Var_{GL} = \left[\frac{\pi^2}{6} + \psi'(C)\right] B^2 \tag{4.27}$$

$$Med_{GL} = A - B\log(2^{\frac{1}{C}} - 1) \tag{4.28}$$

where $\gamma \doteq 0.57721566$ is Euler-Gamma constant and $\psi(x)$ is multi-Gamma function. A value of C is predefined. Figures 4.15a,b indicate that there is not a major difference in the shapes of the curves for $C > 5$. A value of $C = 10$ has been chosen. The variance and median of the distances $\tilde{d}_i(m,n)$ are used as estimates of the Med_{GL} and Var_{GL} model parameters.

The secondary mode in the histogram is detected as a positive deviation from the GLD to the right of the GLD's peak. By evaluating the relative deviation of the histogram from the GLD it can be established whether this deviation is significant enough to represent a secondary distribution of object distances. Figure 4.16 shows the relative deviation between fitted GLD curve and the histogram for two sample scenes. Secondary distribution is considered significant if the relative deviation exceeds 80%. The value is acceptable for the development scenes. A threshold is set up at the position between the two modes. The position is given as (Yusoff et al., 2000)



(a) no objects

(b) with objects

Figure 4.15: Distance histograms with General Logistic Distribution curves fitted. The secondary peak in the distance histogram (b) indicates the presence of objects in the scene.

$$t_d = m_o - k_\sigma \sigma_o \tag{4.29}$$

where m_o is the peak position, σ_o is the estimate of standard deviation of the secondary distribution and $k_\sigma = 2$ so that 95% values are included assuming that the secondary distribution is normal.

The distances determined from multiple frames are collected in order to smooth the histogram and, consequently, improve the estimation of GLD parameters. Results of the segmentation evaluation show that between 5 to 10 frames provide an ample amount of data for histogram generation.

A temporal long-term averaging is applied to avoid any short-term inconsistencies in the threshold values

$$T_i = \frac{((N_t - 1)T_{i-1} + t_d)}{N_t} \tag{4.30}$$

where T_i is the overall threshold in frame i, N_t is the length of the temporal filter and t_d is the current threshold obtained from Equation 4.29. If, for example, the relative deviation of the secondary distribution is close to the threshold value of 80% it occasionally drops below that value and the secondary distribution remains undetected. The histogram is then treated as unimodal pushing the threshold to higher values. That causes objects to flash on and off, making the segmentation inconsistent. The value of N_t is set to approximately 50 frames which corresponds to a couple of seconds.

Algorithm 2 Adaptive thresholding algorithm.

1. A histogram of $\tilde{d}_l(m,n)$ values is generated. An optimal bin size is determined by formula derived by Scott (1979)

$$b = 3.49 \cdot \sigma_d \frac{1}{\sqrt[3]{G}} \tag{4.31}$$

where σ_d is the standard deviation of the data for which the histogram is built and G is the number of values.

2. GLD is fitted. Parameters of the GLD are determined from the parameters (median, variance) of the real distribution.

3. The histogram is smoothed by moving average filter of length 3.

4. Both GLD and smoothed histogram are normalised with respect to their areas.

5. Relative deviation of the histogram from the GLD is determined for all values on the right from the GLD's peak.

6. If the relative deviation exceeds 80% then a presence of secondary distribution is indicated.

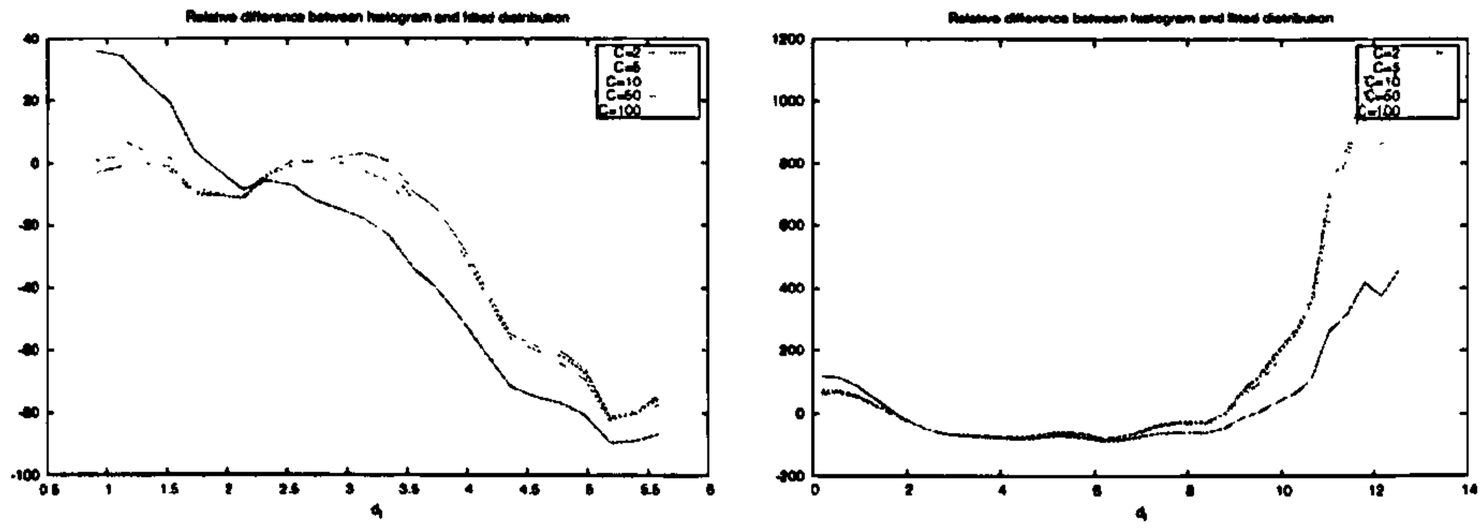
7. The secondary distribution parameters (modus, standard deviation) are estimated.

8. Current threshold is determined from Equation 4.29.

9. Overall threshold is updated by using temporal filtering in Equation 4.30.

10. The threshold T_i is applied to the distances of the current frame.

11. The distances $\tilde{d}_l(m,n)$ above the threshold are labelled 'object', distances below the threshold are labelled 'sea'.



(a) no objects

(b) with objects

Figure 4.16: Relative difference between the histogram outline and the fitted GLD curve for a scene without (a) and with (b) objects present.

The adaptive thresholding algorithm is outlined as Algorithm 2. Each labelled distance corresponds to a feature point in the feature space and in turn, each point corresponds to a segment in the segmentation grid. The threshold T_i represents the classification boundary that separates sea background from objects in the scene.

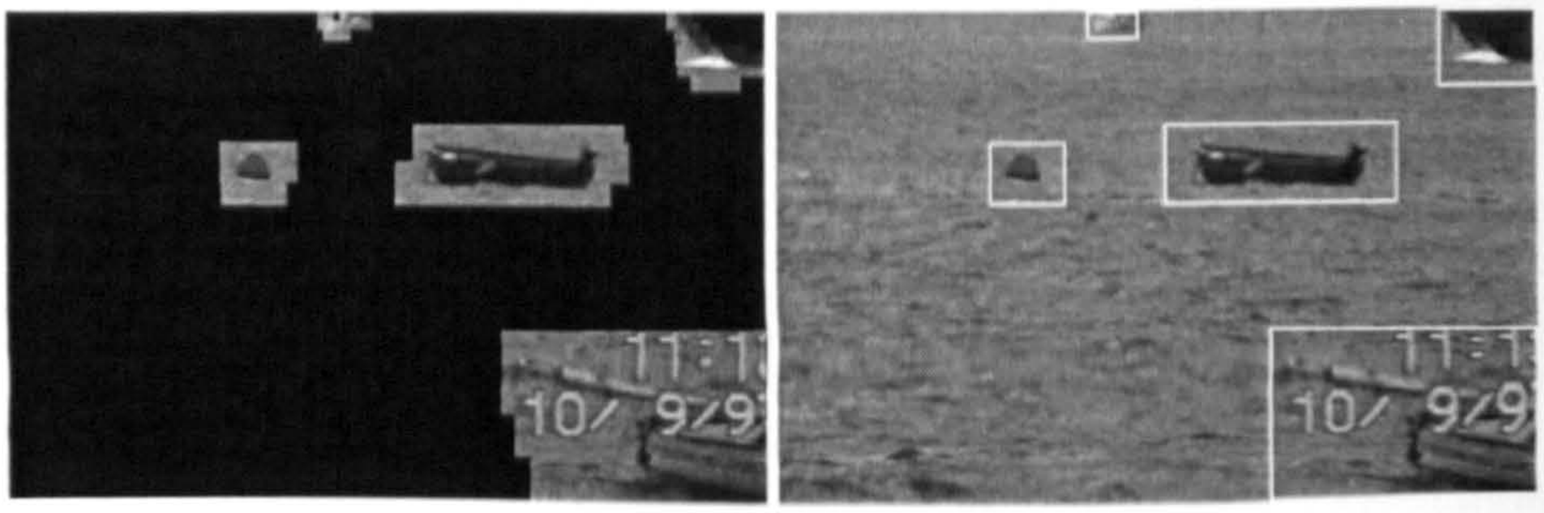
4.7 Remapping of Segmentation Results

Once the feature vectors are classified, segments corresponding to each class can be mapped back onto the original image as objects or background using the segmentation grid. There are three possible ways of the remapping. The first option is to output only those segments labelled as objects. This results in irregular segments that do not necessarily cover the complete objects (see Figure 4.17a). Such irregular regions would pose difficulties for consequent processing steps as they are not true representations of actual objects.

A second option is to enclose irregular regions into rectangles that outline the regions of interest in a more convenient way. This, however, leads to segmentation ambiguity when rectangles enclosing multiple segments of an object broken into several parts overlap.

The third option is to optimise the remapping by unifying the overlapping rectangles into a single one. This is the preferred approach as optimally each object is represented by a single enclosing rectangle.

The segments obtained by remapping provide coarse locations of likely



(a) remapped segments

(b) remapped segments marked by enclosing rectangles

Figure 4.17: Remapping of segments and resulting segmented image.

objects in the scene. The segments represent the input to the consequent processing stage that generates geometric characterisation of detected objects, namely the weak perspective model introduced in Section 2.3.

4.8 Structure of Segmentation Module

The segmentation steps described above are assembled into a processing path that represents a segmentation module in the maritime tracking framework. The structure of the module is outlined in Figure 4.18.

The segmentation module takes as an input the current frame in the sequence . A following set of operations is applied to the frame:

- *Segmentation*. Frame is split by the segmentation grid into individual segments.
- *Rescaling*. Each segment is rescaled to match the size of the smallest segment in the grid.
- *Calculation of features*. A vector of texture characterising features is calculated for each rescaled segment.
- *Partitioning*. Feature vectors span a feature space. A centroid of the space is located iteratively in several steps.
- *Mahalanobis distance*. A Mahalanobis distance between each vector and the centroid is determined.

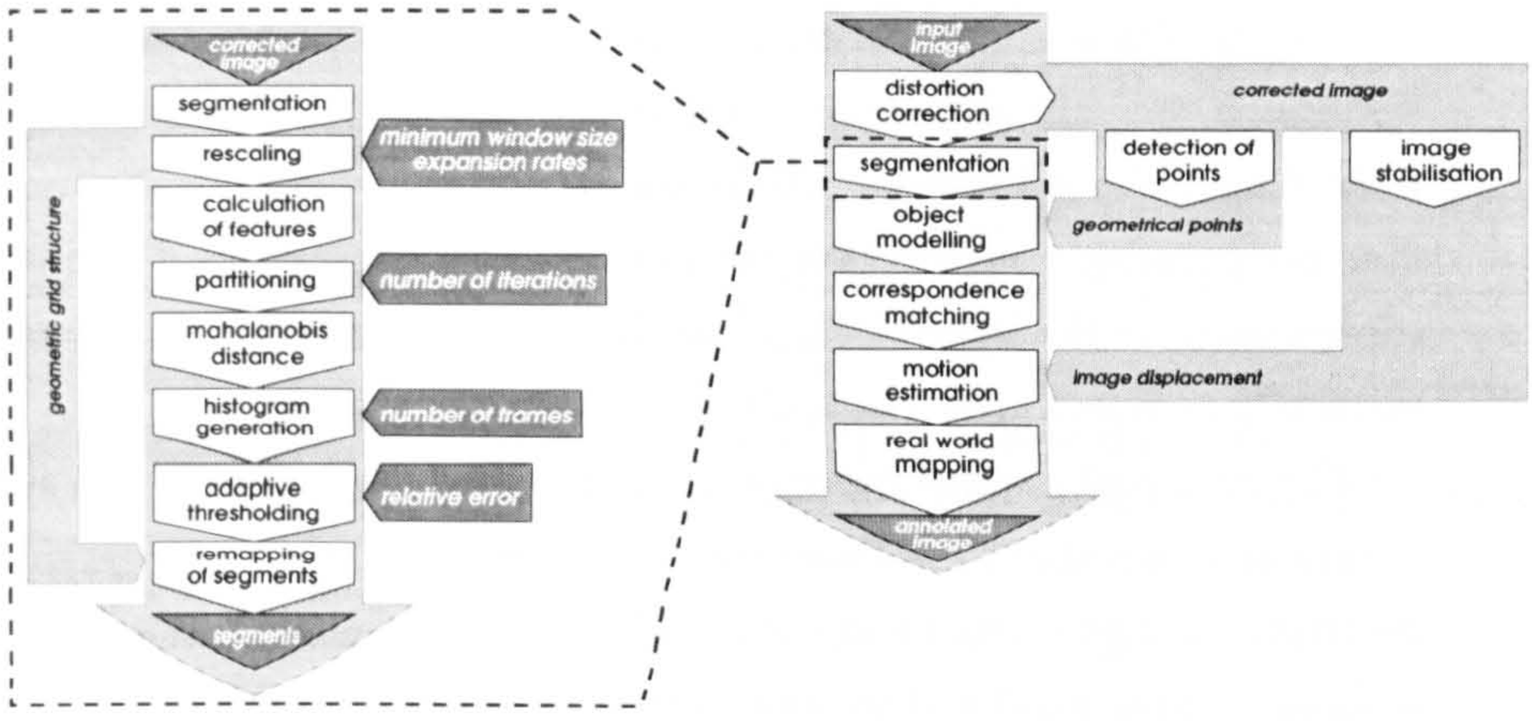


Figure 4.18: The structure of the segmentation module.

- *Histogram generation*. A histogram of the distances is updated. The histogram is initialised using data from multiple frames.

- *Adaptive thresholding*. The histogram is divided by an adaptive threshold into sections corresponding to the sea and to the objects. The value of the threshold is derived from the shape of the histogram and it is updated with each new frame.

- *Remapping of segments*. The thresholded histogram data are remapped back to the corresponding segments in the segmentation grid. Overlapping segments are grouped together.

The coordinates of segments containing possible objects form the output of the module. These coordinates are passed to the consequent module that builds a geometric model of each object in the scene. Some parts of the segmentation module require parameter values that are obtained from experimental evaluations described in the following sections.

4.9 Evaluation of Segmentation Performance

4.9.1 Introduction

An essential part of any classifier design is the evaluation of its performance. Success of classification is proportional to the ability of features to distinguish between different classes. Several tests have been designed in order to analyse the results produced at various stages of the classifier introduced above.

The initial test compares on a statistical basis the performance of the two options of intensity unbiasing described in Section 4.2. A standard t-test and f-test (Dowdy and Wearden, 1991) are employed to confirm the hypothesis that both, mean and median, unbiasing provides the same results. The confirmation of the hypothesis allows the selection of mean as the unbiasing method due to its lower computational demands.

The ability of feature descriptors to separate the outliers from the main cluster is evaluated in the next test. All possible combinations of feature descriptors are generated and each combination is evaluated on a set of sample sequences. Structural and temporal consistency of separability of the selected features is also evaluated. The performance of the classifier is optimised by selecting the combination of features that provide the best separation of features.

The final set of tests establishes adequate values of configuration parameters of the segmentation regarding the best achievable performance of the classifier. The tests involve evaluation of the segmentation for different scaling values Δ_x, Δ_y of segments in both directions, different overlaps o_x, o_y and varying numbers of frames involved in the distance histogram.

4.9.2 Evaluation of Intensity Unbiasing Methods

An essential step in the segmentation algorithm is the intensity unbiasing introduced in Section 4.2. Unbiasing is done by subtracting either mean or median of pixel values determined for each segment in the segmentation grid from each pixel value in the segment. The question is which method gives better results: mean or median ? One possible way to establish the answer is to perform an exhaustive evaluation test of the segmentation performance for a large number of images using both methods on ground truth data.

A more efficient solution is available, that does not involve exhaustive evaluation. Before evaluating the performance of segmentation for each frame in each sequence a hypothesis is made that the values of mean and median are similar. The values would be similar if the intensity values in each segment obeyed symmetrical, preferably normal distribution. This assumption is common in image segmentation algorithms such as (Elgammal et al., 2002).

If the hypothesis is rejected then the exhaustive performance evaluation is inevitable as the results would differ for each method. If the hypothesis is accepted, i.e. mean and median values are similar, the evaluation is

Frame	1	2	3	4	5	6	7	8
t test	A	R	A	A	A	A	A	A
f test	A	R	A	A	A	A	A	A
Frame	9	10	11	12	13	14	15	**A/R**
t test	A	A	A	A	A	A	A	**14/1**
f test	A	R	R	A	A	R	A	**11/4**

Table 4.3: Intensity unbiasing - results of hypotheses testing. The hypothesis is that mean and median of intensity values in each segment come from the same distribution and are therefore similar. The symbols indicate if the hypothesis is A - accepted or R - rejected.

unnecessary as it will provide the same results, no matter which method is chosen.

This test assumes that the values of mean and median calculated for each segment are considered as samples from two statistical distributions. One distribution represents the mean and the other median values. It is possible to find the similarity between these two distributions by using t-test and f-test.

T-test is designed to confirm or reject the hypothesis of two distributions having the same means (Iversen and Norpoth, 1987). In a similar manner, the F-test is designed to test the hypothesis of two distributions having the same variances. If both tests prove that the hypotheses are true then the conclusion is that means and medians determined for each segment and each frame are samples from the same distribution and are therefore similar.

A set of 15 frames randomly chosen from a sample of testing images was used for testing both hypotheses. The results summarised in Table 4.3 prove that both hypotheses are true for the majority of the scenes which means that there is no significant difference in the values of mean or median. As mean is generally easier to determine it is the preferred method of intensity unbiasing.

4.9.3 Evaluation of Separability of Outliers

The ability of features to distinguish between patterns corresponding to objects and those representing the sea is vital for classification. Efficient features should place an outlier in feature space away from the main cluster while preserving its compactness. The following section investigates the performance of various feature combinations listed in Table 4.4 for the purpose of selection of the most efficient combination.

Number	Combination
I.	energy-entropy
II.	energy-homogeneity
III.	energy-contrast
IV.	entropy-homogeneity
V.	entropy-contrast
VI.	homogeneity-contrast
VII.	energy-entropy-homogeneity
VIII.	energy-entropy-contrast
IX.	energy-homogeneity-contrast
X.	entropy-homogeneity-contrast
XI.	energy-entropy-homogeneity-contrast

Table 4.4: All possible combinations of features in segmentation.

4.9.3.1 F-test

Main objectives of the feature selection process are the efficiency of the classifier and possible reduction in complexity. The feature selection is important in cases where reduction of high number of features is necessary for manageable data representation. As the number of feature combinations grows approximately with the factorial of the number of features feature selection by an exhaustive evaluation of all possible combinations becomes impracticable. Dash and Liu (1997) provide an overview and evaluation of several alternative feature selection methods that provide faster but suboptimal feature selection.

As the highest possible dimensionality of data used in the segmentation of maritime scenes is limited to four, the exhaustive evaluation of all possible eleven feature combinations is feasible and there is no need to employ other, suboptimal feature selection techniques.

To evaluate the separability of outliers for different combinations of feature descriptors the following approach is proposed. The distance values $\tilde{d}_l(m,n)$ obtained from Equations 4.24 and 4.25 are considered to be samples of two distributions - one for background feature points and the other for object feature points. These two distributions will presumably have different means and variances.

The f-test (Dowdy and Wearden, 1991) is commonly used to evaluate the difference between two statistical distributions by checking the hypothesis that the two distributions have similar means and variances. The decision is based on the F-value that reflects the differences. If this value is close to one the hypothesis is true and the distributions are similar. If the F-value is larger than

one, the distributions are significantly different. The F-value is proportional to and therefore can also be considered as a measure of the difference between the two distributions. A higher F-value indicates that the distributions of feature vectors belonging to the main cluster and to outliers are less similar. This implies that the separation between outliers and the main cluster in the feature space is greater. The F-value is determined as

$$F = \frac{(\bar{d}_b - \bar{d}_o)^2}{\sigma^2(\frac{1}{G_b} + \frac{1}{G_o})} \tag{4.32}$$

where

$$\sigma^2 = \frac{G_b - 1}{G_b + G_o - 2}\sigma_b^2 + \frac{G_o - 1}{G_b + G_o - 2}\sigma_o^2 \tag{4.33}$$

and $\bar{d}_b$, $\bar{d}_o$ are the means of the distributions of background and objects feature points respectively, σ_b and σ_o are the standard deviations of the distributions and G_b and G_o are their magnitudes. F-values for all possible feature combinations listed in Table 4.4 are determined and the combination with the highest F-value is chosen.

4.9.3.2 Test of Separation Consistency

Another important criterion is the consistency of the separation of outliers with respect to time and scene appearance. The consistency of segmentation results depends directly on the consistency of separation of outliers. As far as possible, separation should remain independent of the scene illumination changes and the scale and appearance of the background and any objects. Adaptive thresholding partially compensates for possible short-term instabilities in data separation by accumulating data over multiple frames.

The algorithm assumes that the distance $\tilde{d}_l(m, n)$ for any objects in the scene remains similar. In terms of feature space, the constraint expresses a detectable and consistent 'gap' between the main cluster and the outliers. In addition, the outliers should be located at approximately the same distance from the cluster centroid to span a detectable and compact secondary distribution of distance values. These properties should preferably remain independent of the structure and appearance of the scene.

The evaluation of the separation consistency is done in the following steps:

1. Artificial sequences are constructed by placing a target at a known position onto the images of a varying background as described in Section

Bkg. - Tgt.	SEA1 - 01	MARINE01 - 02	PORTSM1 - 03	SEA4 - 04	SEA3 - 05	MARINE01 - 06	SEA2 -07	SEA2 - 08
I.	410	857	1108	1192	3208	934	787	3397
II.	414	965	1130	1222	3243	1076	811	3466
III.	681	1000	1162	1339	3668	1157	960	3662
IV.	975	1947	1661	1459	3057	2391	1311	4360
V.	1880	1926	1951	1624	2308	2440	2280	4541
VI.	3117	3587	4884	3650	1992	4391	2920	4091
VII.	423	887	1104	1202	3307	1016	788	3410
VIII.	422	907	1108	1202	3288	991	787	3409
IX.	457	1008	1118	1211	3233	1095	849	3548
X.	746	1878	1641	1330	3324	2331	1284	4160
XI.	416	833	1104	1189	3367	1040	767	3438

Table 4.5: F-values for combinations of different backgrounds and targets.

2.7.3.3[1].

2. The F-values for all the frames in the sequence are determined for a single sequence with a single target at a known position.

3. The previous step is repeated for multiple combinations of targets and backgrounds.

4. F-values are unbiased by their averages in each frame.

5. Average F-value for each sequence is determined.

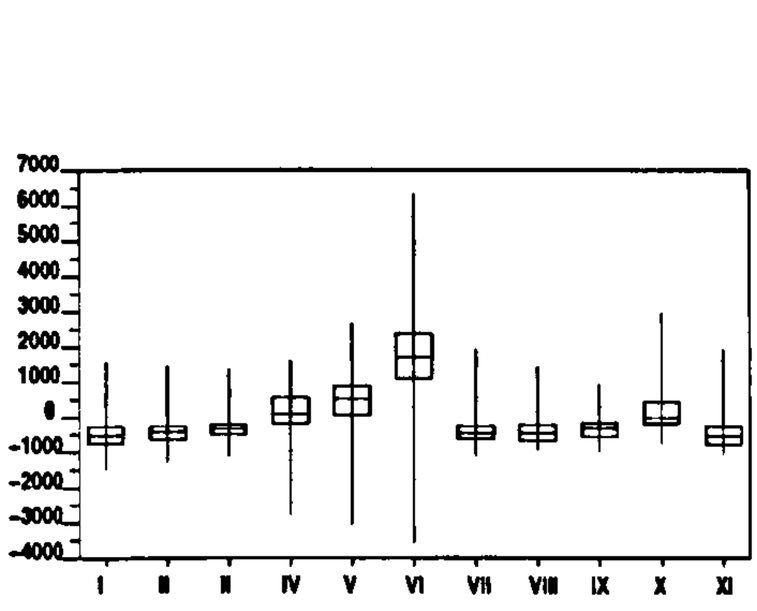
6. Box-plots of F-values are constructed for each combination of features.

Box-plots illustrate the distributions of F-values for each combination of features. Figure 4.19 shows the box-plot for real backgrounds with real targets superimposed. The box contains 95 % of all F-values, brackets delimit minima and maxima. Values are unbiased by their mean.

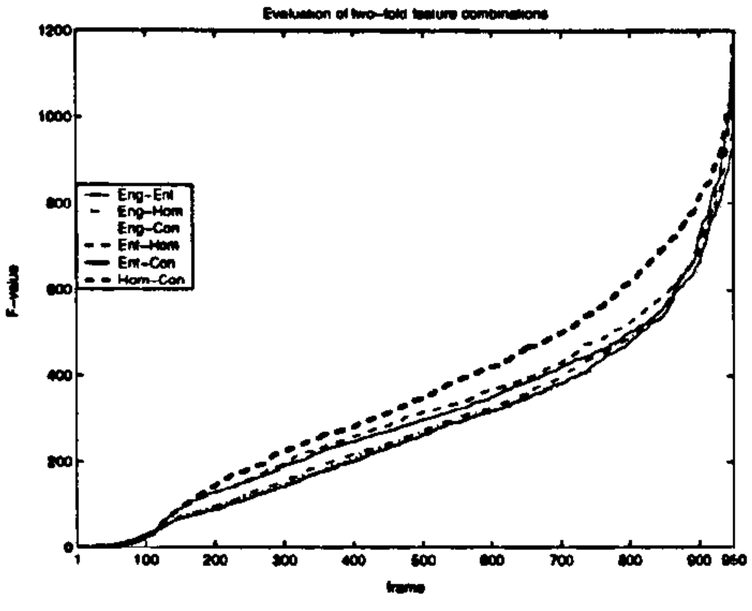
4.9.3.3 Evaluation Results

Test results are summarised in Table 4.5 and Figure 4.19. The results indicate certain combinations of descriptors separate the objects from background

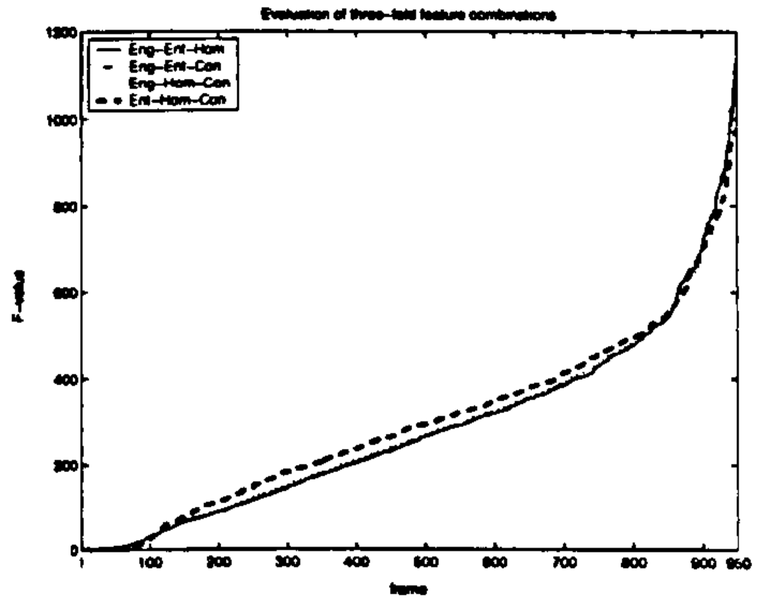
[1]F-test is not applicable if the population G_o of object distances is less than two. This would yield zero standard deviation. The targets are designed and placed such that they cover at least four segments in the segmentation grid.



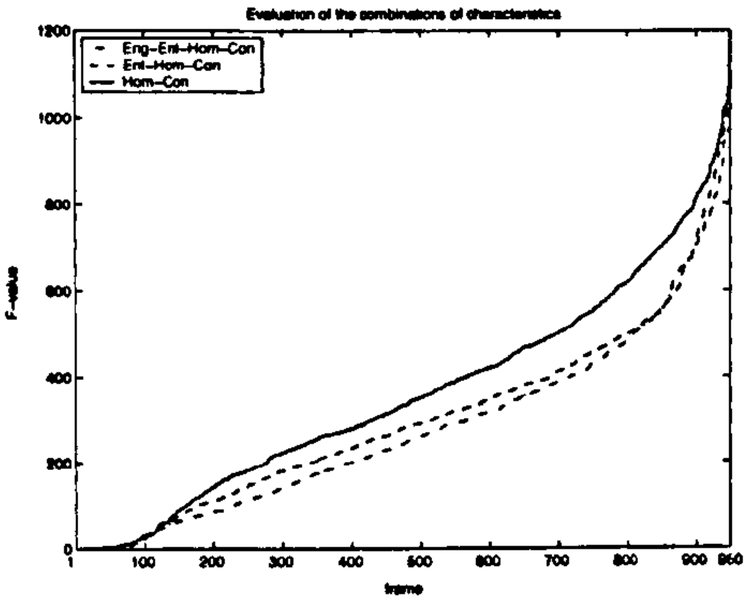
(a) Box-plot of unbiased F-values



(b) F-value of two-fold combinations



(c) F-value for three-fold combinations



(d) F-value for chosen feature combinations

Figure 4.19: Box-plot of unbiased F-values for different combinations of features (a). Comparison of F-values for different combinations of features (b)-(d).

better than others. Unfortunately, this is at the expense of spatial and temporal stability. For example, combination No. VI - homogeneity-contrast in Table 4.4 has an average unbiased F-value around 1900. However, the 95% interval is very broad, almost 1000, and the lowest F-value is -3500. This indicates poor spatio-temporal stability of separation and this combination is not favourable. Two-fold combinations No. I, II, III provide results similar to the combination No. XI - all features, and thus can be considered as viable alternatives.

Figure 4.19b-d shows the plots of F-values for different feature combinations determined from 950 frames randomly generated using real backgrounds and real objects. The values are sorted to emphasise the overall differences (McCane, 1997). The plots indicate that the best overall separation is obtained for a two-fold combination of homogeneity and contrast. As this combination is temporally less stable the second best threefold combination No. X - entropy, homogeneity and contrast is selected instead (see Figure 4.19d).

4.9.4 Evaluation of Structural and Temporal Parameters

The geometrical structure of the segmentation grid depends on a set of predefined parameters. In order to provide reliable functionality the scale of the segmentation grid should reflect the scale of objects in the scene. Segments that are not in proportion to the expected scale of objects encountered in maritime scenes would render the segmentation impossible. The scale conformance also applies to the amount of overlap. If the segments are not overlapping, objects may be only partially segmented or split into multiple regions leading to a misinterpretation of the scene structure. If the overlap of the segments reaches beyond a certain limit, segmentation granularity decreases and multiple objects are interpreted as a single one.

The task of the evaluation is to identify the scales and overlap limits of the segmentation grid (i.e.. structural parameters) to allow for the best segmentation outcome. The evaluation is done by changing the segmentation parameters and evaluating the segmentation on a sample sequence with a known ground truth.

Another important parameter of the segmentation is the number of frames involved in temporal filtering and the accumulation of data used in adaptive thresholding. If data accumulates over too many frames, short-term changes are filtered with a consequence of misinterpretation of activities in the scene. If not enough data is collected the histogram of distances becomes sparse making

the estimation of distribution parameters unreliable. To evaluate the influence of the number of frames on segmentation outcome the sample distance is processed with different numbers of accumulating frames and the results are compared against a known ground truth.

The criteria for evaluating segmentation performance take the following three indications into consideration:

- extra spurious segments that do not represent any rigid objects - false positives (FP)
- misinterpretations of the objects, i.e. object is segmented into multiple segments or one segment covers multiple objects - misinterpreted positives (MP)
- lost objects, i.e. segmentation failed to find an object present in the scene - false negatives (FN)

Tests are run for a sample sequence and the performance is evaluated over a range of values of tested parameters by using the indications mentioned above. Remaining parameters are kept constant through each test phase. The indications are counted each time they occur and the final score is established as a ratio of the number of occurrences to the length of sequence in frames. The score is expressed in percent.

The parameters tested are

- Δ_x - relative change of segment size in x-direction
- Δ_y - relative change of segment size in y-direction
- $o_{x,y}$ - horizontal and vertical overlays
- N_T - number of frames involved in histogram construction

4.9.4.1 Results

The results of the evaluation are presented in Tables 4.6-4.9 for POOLE sequence. The POOLE sequence contains multiple objects of different scales and appearances that provide a suitable sample of likely objects in a maritime scene.

The results indicate a relation between the geometry of the segmentation grid and the outcome of the segmentation. For small values of Δ_x,Δ_y (2%, 5%

Δ_x [%]	FP [%]	MP [%]	FN [%]
2	92	19	6
5	64	19	8
10	74	18	1
15	52	25	12
20	42	36	82
30	22	55	100

Table 4.6: Evaluation of the segmentation with regard to the relative change of the size of segments in vertical direction. Other parameters are set to the following values: $\Delta_y = 10\%$, $o_x = o_y = 50\%$, $h = 5$.

Δ_y [%]	FP [%]	MP [%]	FN [%]
2	100	100	0
5	69	44	0
15	42	14	28
20	52	36	6
30	39	44	74

Table 4.7: Evaluation of the segmentation with regard to the relative change of the size of segments in horizontal direction. Other parameters are set to the following values: $\Delta_x = 10\%$, $o_x = o_y = 50\%$, $h = 5$.

in both directions) objects that are close to the camera (large in scale and near the bottom of the image) are split into many parts and can be misinterpreted or lost completely. If the size of segments increases by more than 15%, objects that are small or far from the camera (small in scale and near the horizon) can be misinterpreted or lost completely. Best results are obtained for scale change of about 10% in both directions.

Overlapping influences the misinterpretation of the objects. For values of $o_{x,y}$ less than 20% the objects are either partitioned or missed completely. For values exceeding 50% objects are grouped together. If the overlapping drops below 10%, adaptive thresholding fails as the number of feature vectors for segments with objects drops below the 80% level of significance and the threshold is pushed towards higher values.

A number of accumulating images in the adaptive thresholding is evaluated in the final part of the test. Results show that the optimum number of frames for data accumulation is between 5 and 10 frames. Below this amount there are not enough feature vectors to make the secondary distribution significant enough to be separated by a threshold. On the other hand, more than 10 frames decrease temporal adaptivity of the thresholding.

The high values of false positives are due to a number of wakes that

$o_{x,y}$ [%]	FP[%]	MP[%]	FN[%]
2	29	78	77
5	59	19	15
10	16	86	47
20	70	19	6
33	64	17	0
66	86	50	0
75	61	62	0

Table 4.8: Evaluation of the segmentation with regard to the relative change of the overlap of segments in both directions. Other parameters are set to the following values: $\Delta_{x,y} = 10\%$, $h = 5$.

$h[frames]$	FP[%]	MP[%]	FN[%]
1	49	39	68
2	57	15	7
10	67	17	2
20	68	18	4
50	67	17	5

Table 4.9: Evaluation of the segmentation with regard to the number of frames used in estimation of distance distribution. Other parameters are set to the following values: $\Delta_{x,y} = 10\%$, $o_{x,y} = 50\%$.

appeared in the testing sequence. These are rapidly moving objects that leave traces in a form of bright spots or lines and are picked up by the segmentation algorithm but they are filtered out in consequent processing modules as they are present in the scene just for a couple of frames.

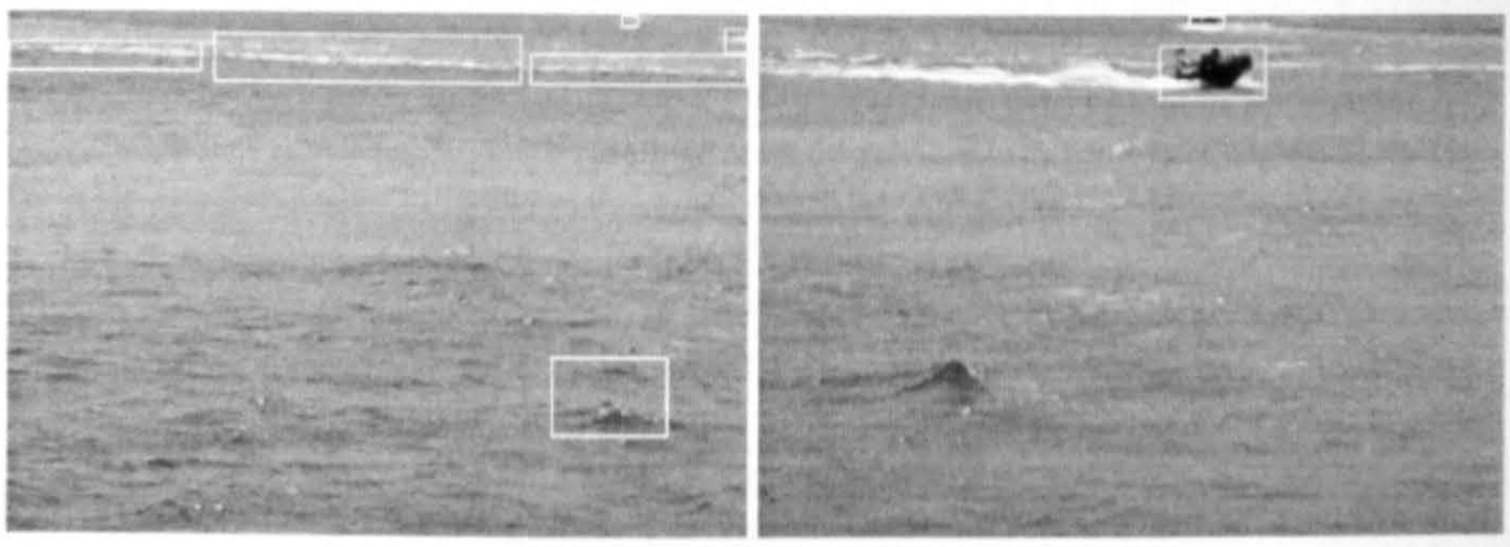
4.9.5 Cross-Validation

The segmentation is cross-validated on two other previously unused sequences in order to confirm validity of the values obtained by the evaluation described above. The segmentation is applied to the SANDBANKS2P and WEYMOUTH2B sequences with segmentation parameters set to values summarised in Table 4.10.

The SANDBANKS2P sequence is 330 frames long. It contains two RIBs

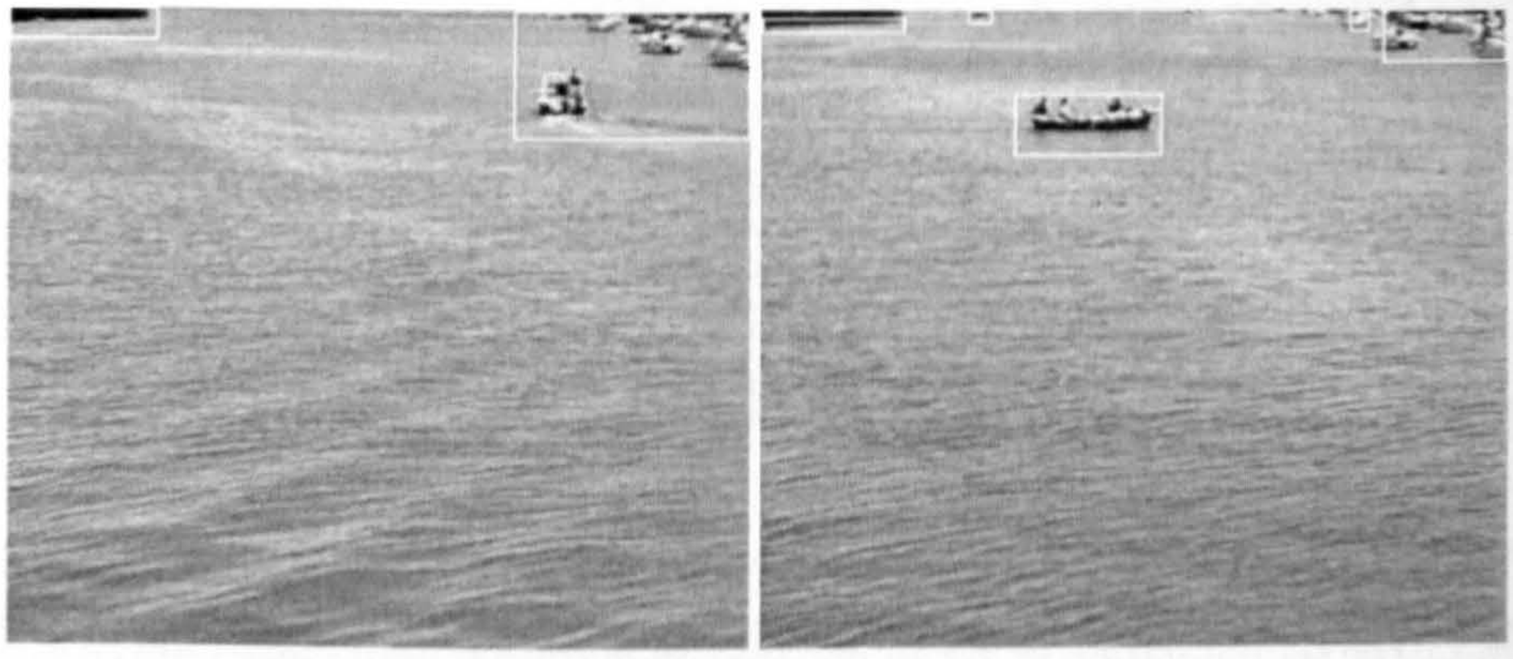
Δ_x	Δ_x	o_x	o_y	w_{min}	h_{min}	iter.	h
[%]				[pix]		[fr]	
10	10	50	50	10	10	2	7

Table 4.10: Segmentation parameters used in cross-validation.



(a) frame 21

(b) frame 142



(c) frame 21

(d) frame 61

Figure 4.20: Sample sequences used in segmentation cross-validation. SANDBANKS2P: (a) false positives caused by the wake, (b) correct segmentation. WEYMOUTH2B: (c) misinterpreted positive on right, (d) correct segmentation.

Sequence	FP	FN	MP
SANDBANKS2P	29.4	0.6	3.9
WEYMOUTH2B	0	0	2.3

Table 4.11: The results of the segmentation cross-validation. The values indicate the probability of occurrence of false positives (FP), false negatives (FN) or misinterpreted positives (MP).

moving fast across the scene, one at a distance, the other close to the camera. There is a large bright wake at the beginning of the sequence caused by a cargo ship passing. Two segmented sample frames are shown in Figure 4.20a,b.

The WEYMOUTH2B sequence is 422 frames long. It shows an entrance to Weymouth harbour on a calm day. There is a boat leaving the harbour on the right side of the scene. A small ferry passes across the scene in the middle. Another boat enters the harbour on the right side. Two segmented sample frames are shown in Figure 4.20c,d.

The segmentation results of the test are shown in Table 4.11. The values represent the probability of the false positives (FP), false negatives (FN) or misinterpreted positives (MP) to occur in each frame. The values are determined as average occurrences with respect to the length of the sequence.

There is a 29.4% probability of false positives in the SANDBANKS2P sequence. This is due to the presence of a large bright wake at the beginning of the sequence. Other values for both sequences indicate that the selected parameter values provide satisfactory results.

4.10 Summary

An essential task of the scene segmentation into background and objects is the initial stage of the processing. There are number of methods that try to achieve this. Due to a textural nature of maritime scenes, the selected segmentation method is based on statistical characterisation of the textures appearing in the scene. Statistical characterisation of textures often utilises a co-occurrence matrix. Statistical features, namely energy, entropy, homogeneity and contrast, are determined from the values in the co-occurrence matrix. The evaluation of these features conducted in Section 4.9.3 shows that combination of entropy, homogeneity and contrast provides the best segmentation results.

A hypothesis is tested that the co-occurrence matrix is redundant for successful segmentation of maritime scenes. The statistical features can be

determined directly from image intensity patches. The hypothesis is confirmed in multiple tests that compare the segmentation of sample scenes with and without co-occurrence matrix involved. The results in Figures 4.3 and 4.4 illustrate the deterioration of the segmentation performance when using the co-occurrence matrix.

The compensation for the perspective projection of the scene is inherent in the structure of the proposed segmentation grid. In order to regularise the feature values the segments are resized to the scale of the smallest ones. Calculated features for each grid segment form vectors in a feature space that is partitioned into main class cluster and outliers by an iterative search of main cluster centroid.

The boundary of the main cluster is obtained from a histogram analysis of distances between individual feature points and the main cluster centroid. The boundary serves as a decision rule that assigns outliers and their corresponding segments either to the sea background or to the objects.

The values of the segmentation algorithm parameters devised by a set experimental evaluations are cross-validated in Section 4.9.5. The results show that the devised values achieve less than 1% probability of objects being undetected (see Table 4.11).

The results of the texture-based segmentation provide primary localisation of areas with possible presence of objects in the scene. This information is utilised in the consequent processing steps.

Chapter 5

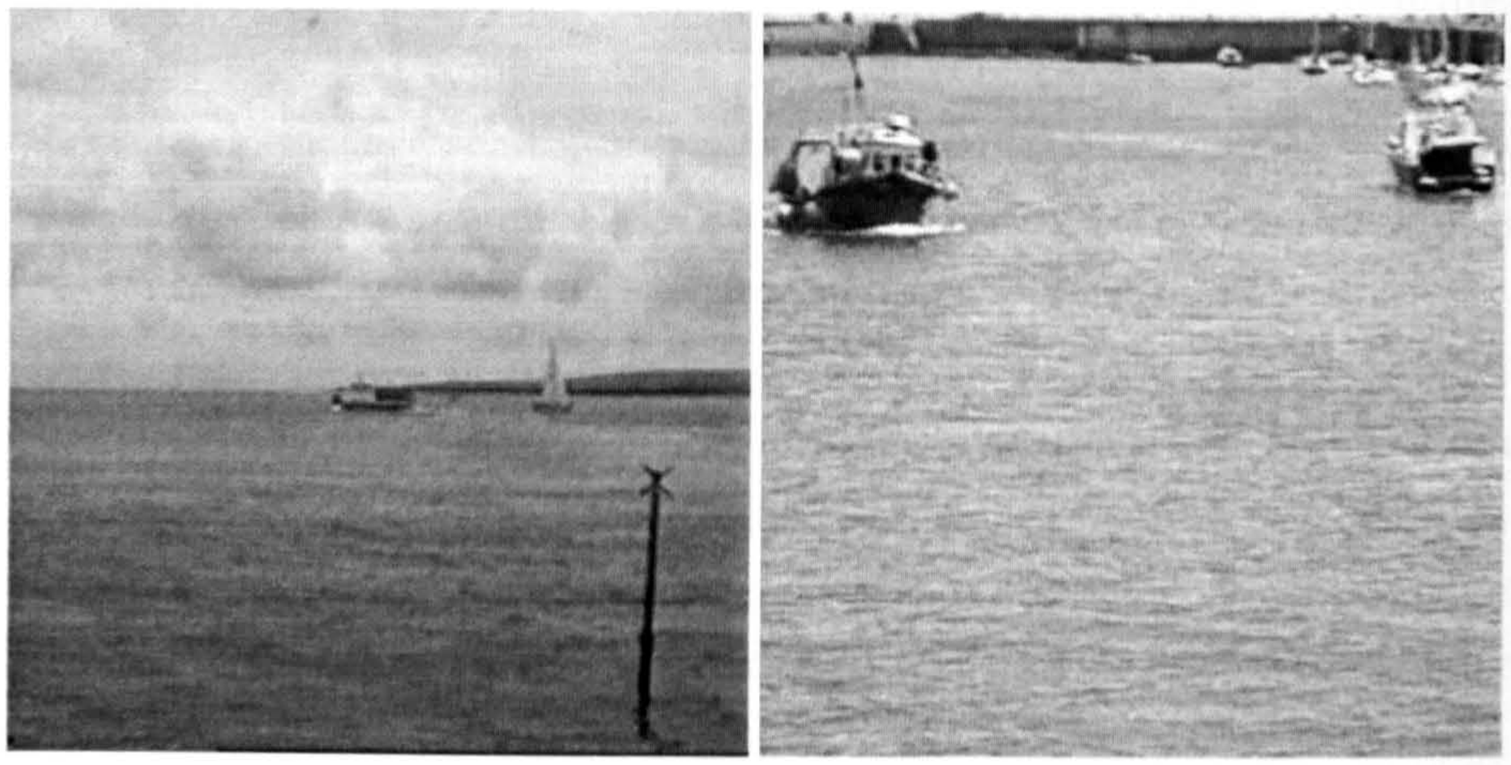
Detection of Geometric Features

5.1 Localisation of Objects

Once the segments containing likely objects are obtained, precise locating of objects to a pixel level within each segment can be determined. Many segmentation algorithms such as (Elgammal et al., 2002; Magee, 2004) directly work at a pixel level. However, such methods assume a temporally static background with only minor disturbances from scene noise such as shadows or tree branches moving in the wind that change the distribution of background intensities within a limited interval. An alternative to a pixel based segmentation is proposed which detects only geometric features useful for tracking of objects in the scene regardless of their structure and appearance.

5.1.1 Edge-based Segmentation

The pixel level segmentation algorithms are based on assumptions of spatial or temporal region consistency and/or edge continuity, (Pal, 1993). Unfortunately, the use of edges is not effective for objects in maritime scenes, as illustrated in Figure 5.2. A set of standard edge operators - Canny, Frei-Chen, isotropic, Marr-Hildreth, Prewitt and Sobel, was applied to a sample maritime scene in Figure 5.1a. The results show that the horizontal edges detected in the scene are buried in substantial noise originating from the presence of high



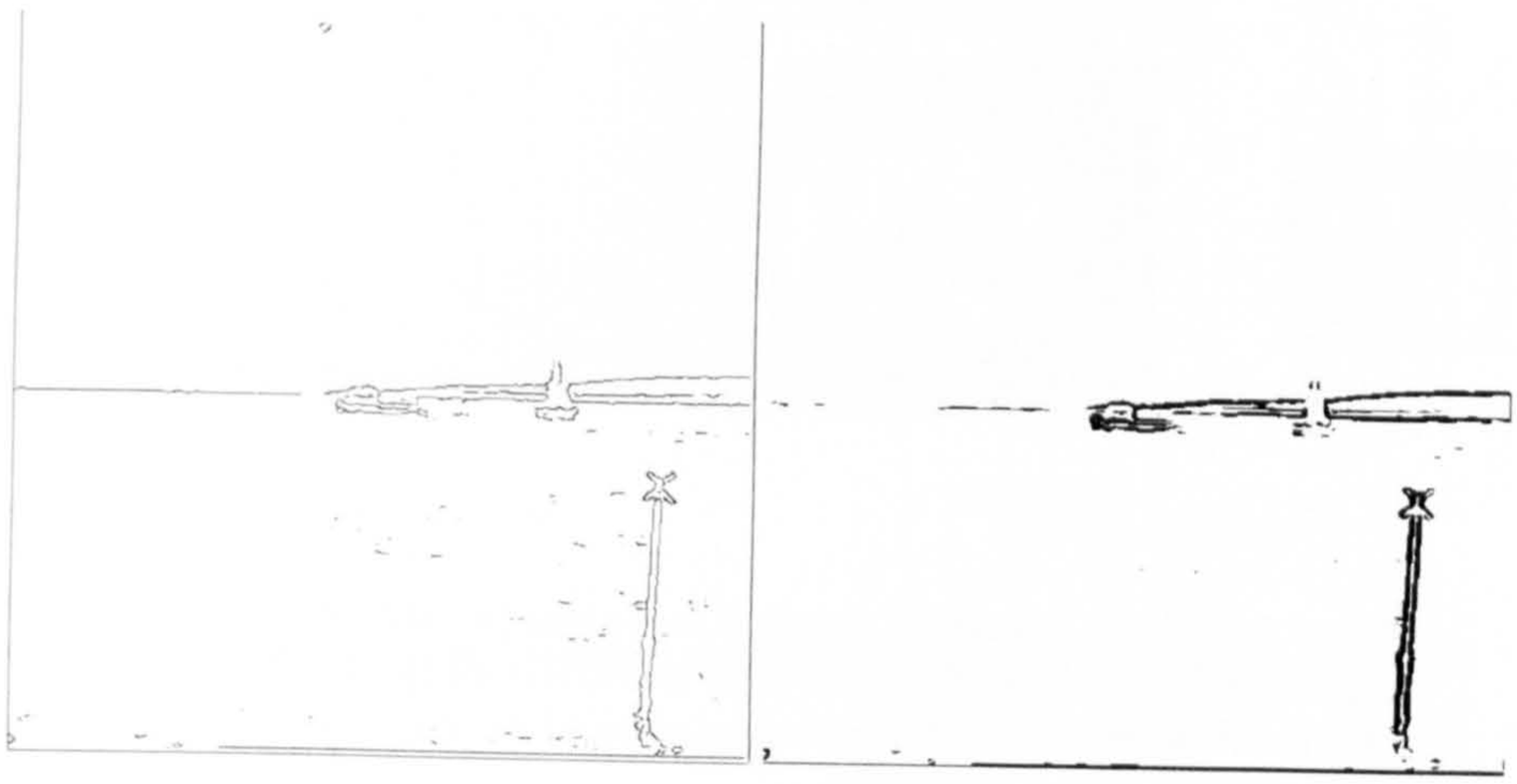
(a) SANDBANKS2D (b) WEYMOUTH2E

Figure 5.1: Sample maritime scenes used to evaluate standard methods of edge- and region-based segmentations.

contrast wakes on the sea surface.

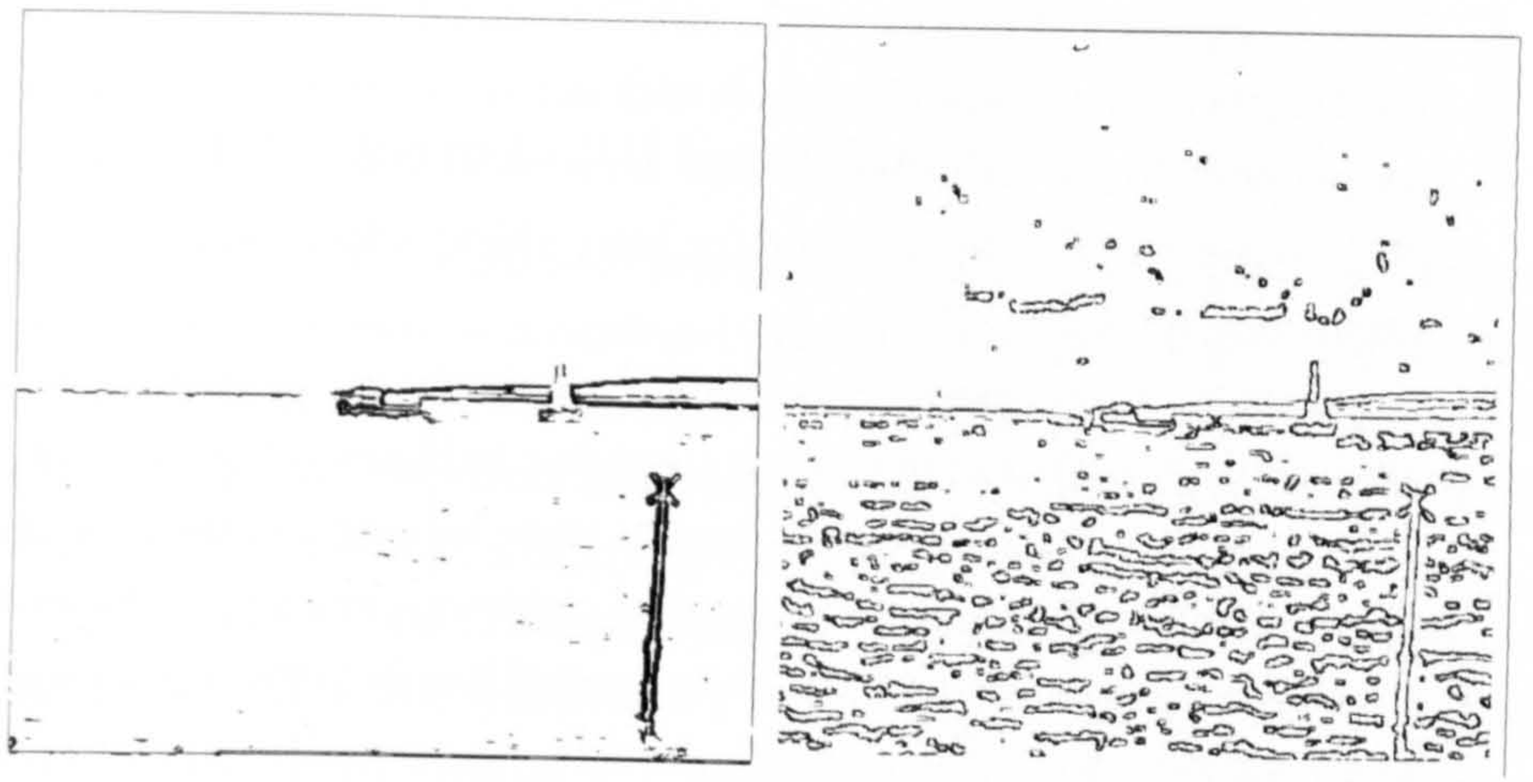
Maritime objects are often composed of parallel, mainly homogeneous regions with strong edges present along boundaries (see Figure 5.1). These edges trigger a strong detection response causing the objects to be split into multiple, disjointed, homogeneous regions. Such a situation is shown in Figures 5.2 where the vessel on the left is split into many disjointed parts, some of which have open boundaries that do not fully enclose the region.

Thresholding of edge responses is necessary for the majority of algorithms using edges for segmentation. The process is generally called non-maxima suppression. Edge responses are not uniform as illustrated in Figures 5.2 where the yacht on the right fails to be detected as a uniform region with a closed boundary even though it is clear from the original image that the hull of the yacht is a uniform region. Some form of adaptive thresholding or edge tracing is necessary to ensure boundaries are closed. However, the thickness of edges does not remain uniform even after non-maxima suppression, so an optional edge thinning would be needed to obtain an unambiguous edge representation. In general, because edge methods are mainly gradient-based, their noise sensitivity prohibits their employment in maritime domain where scenes are considered noisy in principle.



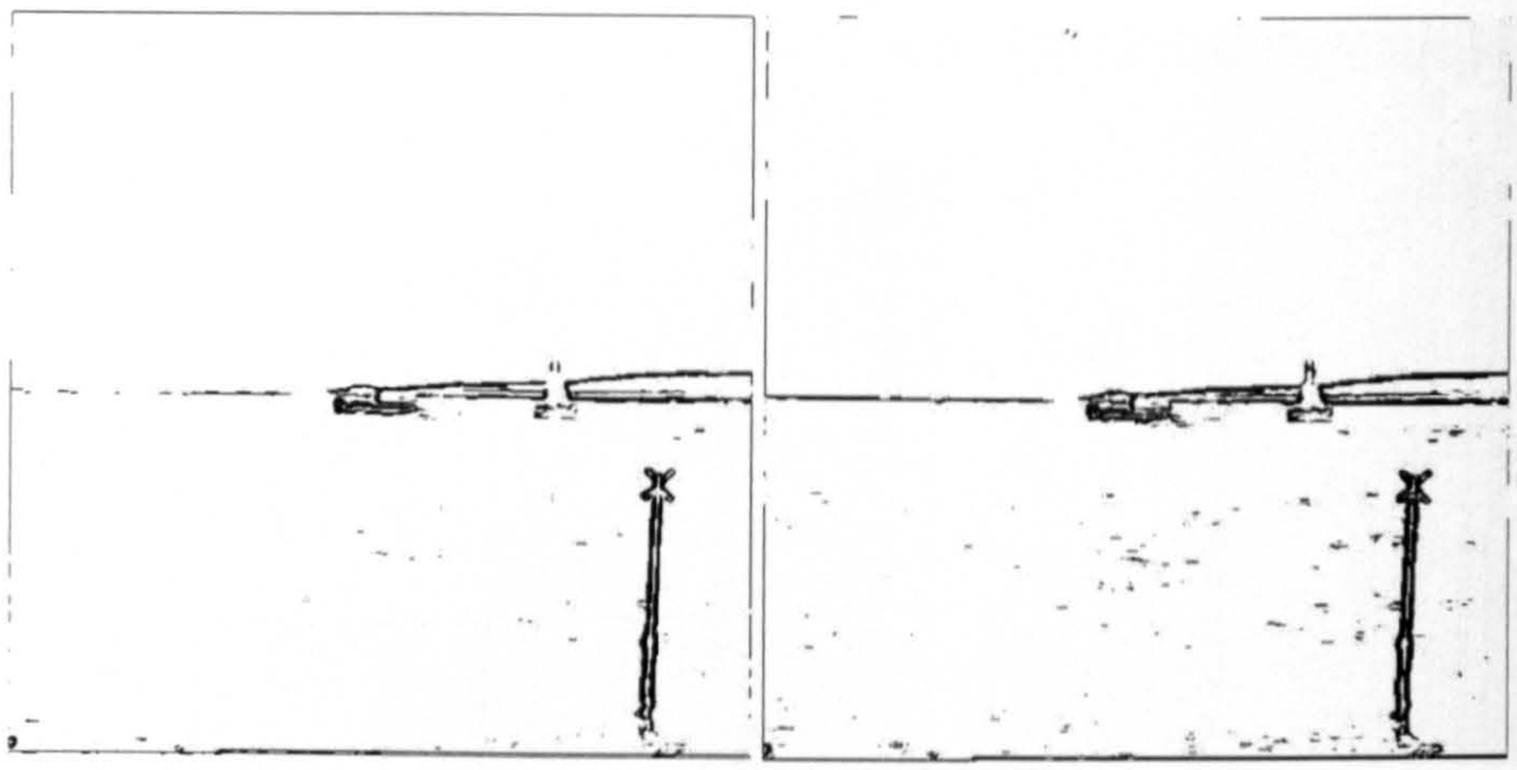
(a) Canny

(b) Frei-Chen



(c) Isotropic

(d) Marr-Hildreth



(e) Prewitt (f) Sobel

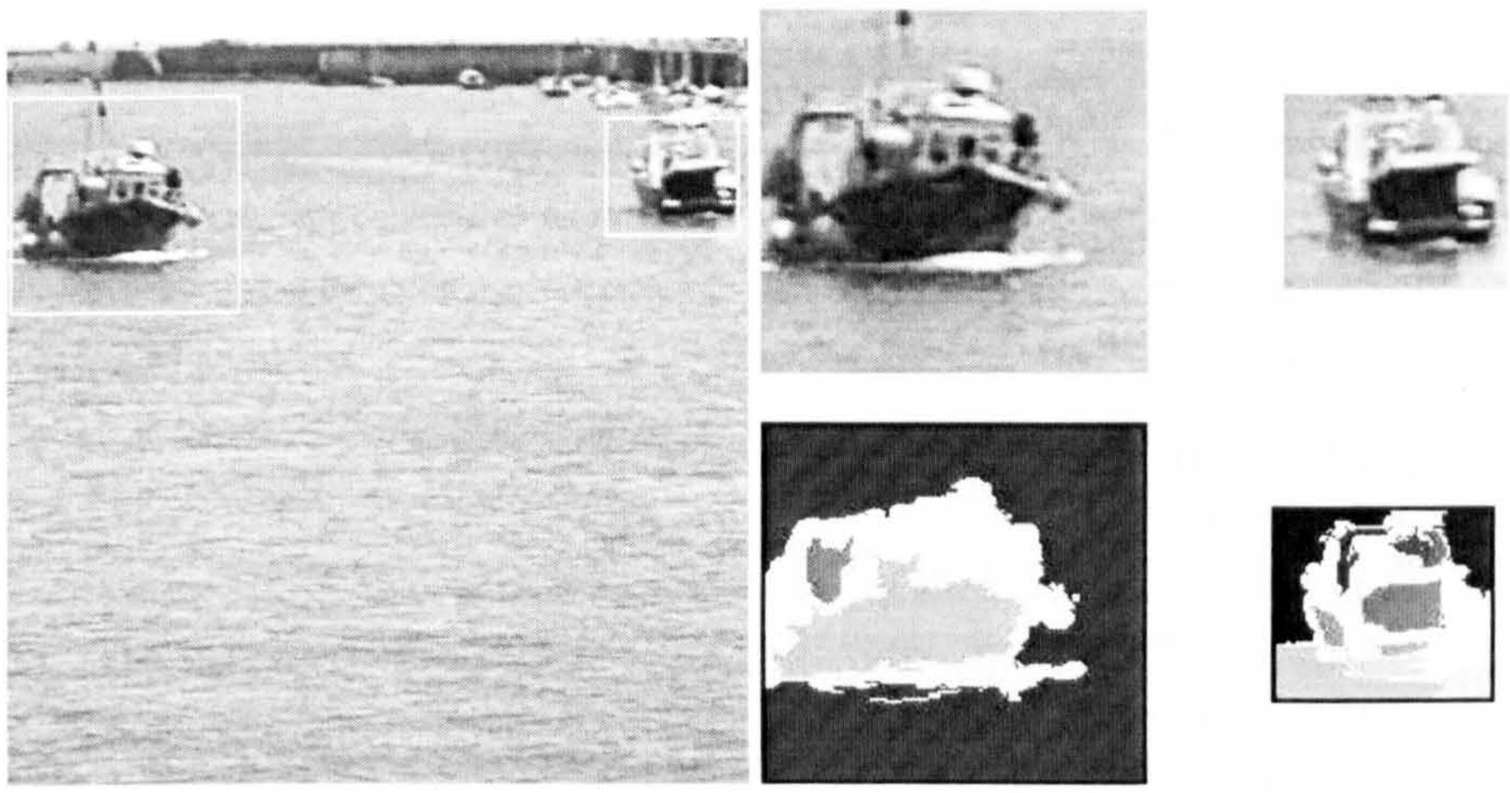
Figure 5.2: Edge operators applied to a sample maritime scene. Varying edge responses are indicated by varying thickness of lines. The thickness of lines remains different for different edges even after non-maxima suppression.

5.1.2 Region-based Segmentation

The homogeneity of regions can be also considered as a criterion for the pixel-level segmentation. Typically, pixels with similar intensities and spatially close are assigned to the same region. The segmentation based on classification of the pixels is usually iterative and/or hierarchical, (Pal, 1993), and it relies on predefined similarity criteria.

A set of pixel-based segmentation results based on similarity criteria applied to a sample maritime scene are shown in Figure 5.3. In addition, Figure 5.1b shows two different vessels, one of them leaving and the other entering port. A segmentation algorithm based on clustering, available in Khoros image processing suite by Khoral (2003), was applied locally to the segments obtained in the previous step. The algorithm assigns the same label to pixels within 8-connected neighbourhoods based on their similarity defined as the intensity difference below a given threshold. The number of labels is unknown and the merging factor is set to 7 % of maximum difference between intensities of neighbouring pixels which is the recommended value.

The results of the region segmentation are shown in Figure 5.3. Both objects are fragmented into multiple exclusive parts. The background in the segment with the second object (small boat on the right in the original frame)



(a) segmented objects (b) labelled objects

Figure 5.3: Region labelling applied to a sample maritime scene.

is misclassified and split into two separate regions. The results show that the homogeneity of regions is insufficient criterion for segmentation of maritime scenes due to their variable nature.

5.1.3 Effects due to Initial Segmentation

A partial failure of the primary segmentation described in Chapter 4 brings other issues that must be considered. Ideally, rectangular segments provided by the segmentation should contain whole objects plus a small part of the surrounding sea. Segmentation partially fails when the segments contain either only parts of objects or multiple objects in a single rectangle.

The solution in the first case would involve an expansion of the original segment to take in the whole object. The criteria of such an expansion are difficult to establish, especially if the object is composed of multiple varying parts.

The case of multiple objects in the segment is even more difficult to resolve, especially if the objects are close together. Unless the number of objects in the segment is known, it is difficult to decide whether the regions correspond to multiple separate objects or to a single object composed of multiple parts. The solution would probably involve temporal filtering for moving objects but there is no simple solution for static ones. For example, in (Lipton et al., 1998)

objects that cease to move for a certain period of time become parts of the background.

The conclusion is that the option of obtaining a representation at pixel-level of complete objects is unrealistic due to such factors as image and scene noise, complexity of the objects and the presence of both static and moving objects.

5.1.4 Motion-based Segmentation

For some surveillance applications in a traffic environment the motion estimation is their primary goal. Numerous segmentation methods provide the structure of the scene from information about the motion (so called 'structure from motion' methods). There is a vast number of algorithms for motion segmentation available in the machine vision area, (Zhang and Lu, 2001). They can be divided into two main categories - optical flow-based and feature-based. While the optical flow based algorithms are successfully used in many applications involving outdoor scenes, (Lipton, 1999), the temporally variable background of maritime scenes generates a lot of false motion cues. An essential overview of optical flow estimation algorithms is provided by Barron et al. (1994) and Beauchemin and Barron (1995). The authors also provide implementations of algorithms discussed.

To evaluate performance of optical flow estimation techniques on maritime scenes a group of algorithms was chosen and applied to a sample maritime scene. The algorithms that were chosen for evaluation are those implemented and presented by (Galvin et al., 1998a). These are namely:

- methods of Horn & Schunck and Lucas & Kanade (both standard and modified versions) based on the first order image derivatives
- methods of Nagel and Uras based on the second order image derivatives
- methods of Anandan and Singh based on region matching
- a method by Quenot (Quenot, 1996) based on dynamic programming

Due to the difficulty in obtaining the ground truth for the real scenes used in evaluation (Galvin et al., 1998b), the results presented in Appendix A provide mainly qualitative insight rather than objective quantitative evaluation. The results show that for many algorithms a substantial number of motion vectors are due to the temporal variability of the water surface. The methods of Lucas

& Canade (modified version) and Nagel provide the best results. Although other methods detect motion of rigid objects, this motion is not separated from the motion of the sea. The major factor in the failure of optical-flow based algorithms is the violation of the essential temporal intensity conservation constraint that stipulates that local temporal change of intensity is only due to the displacement. The violation occurs due to the changes in the appearance of the water surface that are not only due to the displacement but they are mostly caused by interaction of the water media with the incoming light. Despite an attempt to use optical flow for segmentation of natural scenes of water surfaces (Ablavsky, 2003), the principal requirement of intensity conservation is not met in maritime scenes.

5.1.5 Salient Features

An alternative to the pixel-based segmentation is proposed which is based on the following assumptions drawn from optical and geometric contexts of maritime scenes discussed in Sections 2.2 and 2.3:

- The depth of the scene is greater than the depth of the objects in the scene. A weak-perspective projection where objects are collapsed to the planes perpendicular to the water plane and parallel to the image plane can be used as object representations.

- For temporal tracking the structure of the object is not essential. Objects can be tracked as grouped sets of features[1] such as corners that obey consistency rules such as representation of underlying structure and localisation, a little change of appearance between two consequent frames, (Shi and Tomasi, 1994). These features are usually detected in the image as, for example, saliency points defined as image locations with high intensity gradients in more than one direction, (Harris and Stephens, 1988; Smith and Brady, 1995). All these points lie on the same plane regardless of the actual depth of the object when the planar representation of objects is considered.

- For threat assessment and collision detection the distance between the closest structural point on the object and the observation point is crucial,

[1]The word 'feature' is not used in connection with texture characterisation. In this and following chapters, the 'feature' corresponds to a structural attribute of an entity in the scene such as corner, line or other point of saliency.

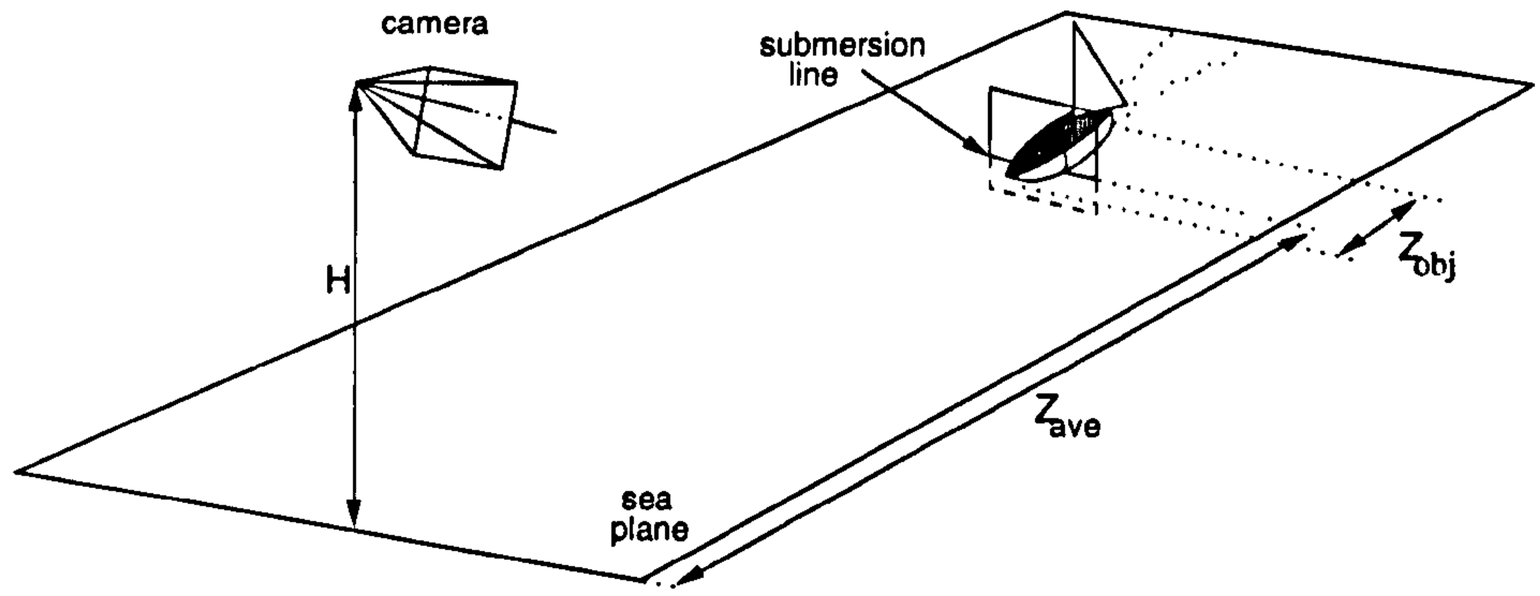


Figure 5.4: Projection of objects onto a single depth plane that represents the object in the scene (a weak perspective model).

not structure of the object. If the object is represented as a plane then its closest structural point lies on this plane.

This situation is illustrated in Figure 5.4. If $Z_{obj} \ll Z_{ave}$ then the object can be collapsed by a local orthographic projection to the plane with zero depth and parallel to the image plane. This plane is at position Z_{ave} in the scene. Note that the plane is not located in the middle of Z_{obj} as proposed by Shapiro (1995) but at the position of submersion closest to the camera so that it is bound to the location of the submersion line.

Based on the above assumptions a following approach of object localisation is outlined. Instead of detecting the objects in segments at the pixel level the alternative is to find salient points (corners) that presumably belong to objects and the submersion lines which locate the objects in the scene with respect to the camera. Each object is then presented by the segment that encloses it, the submersion line and a set of detected salient points that lie on the weak-perspective plane of that object. This set of features is used in consequent processing steps for location and motion estimation.

5.2 Detection of the Line of Submersion

The lines of submersion are chosen as features for locating the objects on the sea plane for the following reasons:

- Due to a low pitch angle of the camera the submersion lines of any objects project as horizontal, almost straight edges in the image no matter what

is the shape of the submerged object. This makes them simple to detect as their orientation is assumed to be known.

- The projections of the lines in the image are directly related to the positions of the lines in the scene through the perspective projection, i.e. if the submersion lines are detected correctly in the image, there are no depth ambiguities as the real submersion lines lie on the horizontal sea plane.

- The lines represent the closest points of impact for distances significantly larger than the depths of the objects. Even though for many objects an overhanging structure means that the actual closest point is closer to the observation point than the one on the submersion line (overhang of ship's bow, for example), the difference is negligible with respect to the depth of the scene.

Submersion lines are detected in all regions obtained in the initial segmentation. The position of the closest point of the object is important in collision prediction and avoidance as it is also the point of possible impact, discounting possible structural overhang. From the geometrical point of view, the submersion line represents the intersection of the weak-perspective plane onto which the object is orthographically projected, with the sea plane (see Figure 5.4).

5.2.1 Detection Algorithm

As the orientation of the submersion lines is assumed to be horizontal the detection task is reduced to a search of horizontal edges within each segment. The search is constrained by the width of the segment in which the object is likely to be. Even though standard edge detection methods seem to be well-suited for the task, Figure 5.2 shows that the edges are very often buried under the noise caused by wakes that, incidentally, project as horizontal edge fragments as well.

A method that is more robust to the noise and appearance variability is proposed. The method employs a vertically sliding mask divided into two parts of equal size. The pixels under both parts of the mask act in a χ^2-based test that determines the difference between the intensity distributions in the upper and lower parts of the mask. The χ^2 measure is suggested by Smith et al. (1998) as an alternative to a standard cross-correlation similarity

measure in matching spatial features. The advantage of the measure is that it scales inversely with the overall intensities due to the sum of intensities in denominator of (5.1). This allows line candidates to be detected even when the scene illumination conditions are not favourable (cloudy day, dark objects in shadow).

The difference value is defined as

$$d(k) = \sum_{r,c=0,0}^{R-1,C-1} \left[\frac{f(k-r,c) - f(k+r,c)}{(f(k-r,c) + f(k+r,c)} \right]^2 \tag{5.1}$$

where the $d(k)$ is determined for every vertical position k in the scanned segment, $f(r,c)$ is the intensity value in the mask at position (r,c) and R,C are dimensions of the mask. Figure 5.5a shows the structure of the detection mask together with the sample $d(k)$ profile obtained and the submersion line candidates being detected.

There are multiple peaks in the profile. Each peak represents significant difference between the intensity distributions in the top and bottom parts of the mask which indicates that there is a horizontal boundary between two different regions under the mask. Each horizontal boundary is a submersion line candidate. To localise the extremes in the profile and the corresponding line candidates Algorithm 3 is applied to every $d(k)$ profile.

The principle of the Algorithm 3 is similar to the principle of a compass operator proposed by Ruzon and Tomasi (1999). Ruzon uses a circular mask that is divided into two halves to detect boundaries between regions in colour images. His search is based on an Earth Mover's Distance (EMD) between colour signatures determined for each half of the mask. The mask is rotated by 180° and the EMD profile is calculated. The maximum of the profile indicates the orientation of the boundary at a given point.

The difference between compass operator and algorithm presented here is that the mask is rectangular and the distance profile is determined for displacements instead of rotations as the angle of the boundary is known. Also, the χ^2 measure is used instead of EMD as it is less complex while sufficiently discriminating.

5.2.2 Adaptive Parameters

The width of the sliding mask is governed by the width of the segment and the height of the mask is adjusted according to the absolute vertical position

Algorithm 3 Algorithm for detection of submersion line from χ^2 profile.

1. Profile is thresholded to avoid low, noisy peaks caused by waves or inhomogeneities in parts of the object. The threshold value t_d is established as

$$t_d = \mu + 0.1\sigma^2$$

where μ is the average and σ^2 is the variance of the whole profile (Figure 5.5c). The sum is chosen as it reflects the scaling of values in profiles for segments of different sizes.

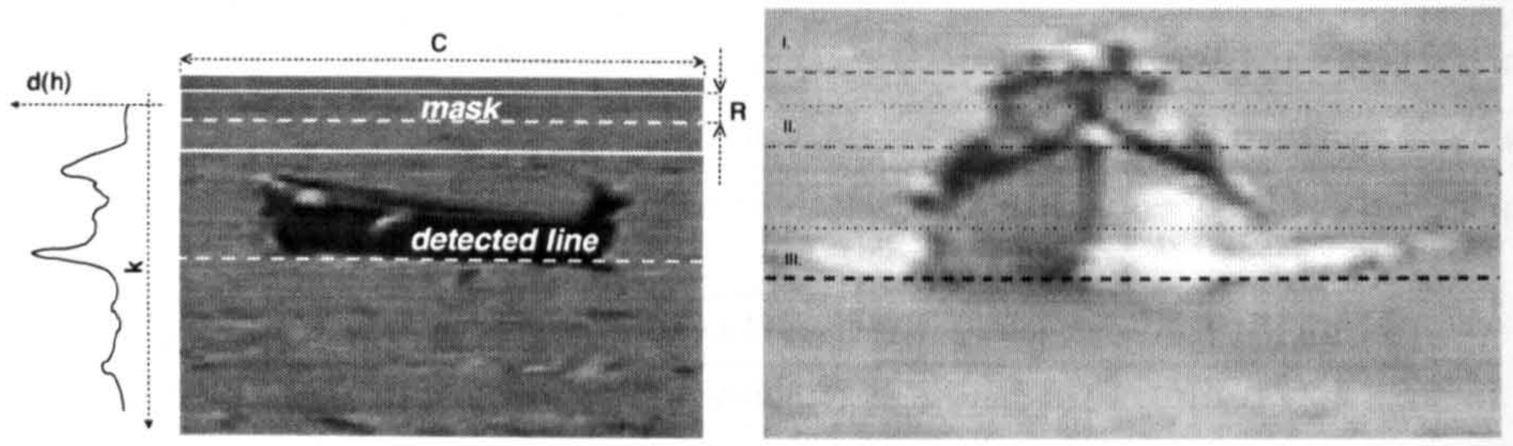
2. The thresholded profile is smoothed by a moving average filter three samples wide.

3. The 1st difference of the profile is determined so that extremes are mapped to zero crossings.

4. The differentiated profile is scanned for zero-crossings. Each crossing locates the line candidate (Figure 5.5d).

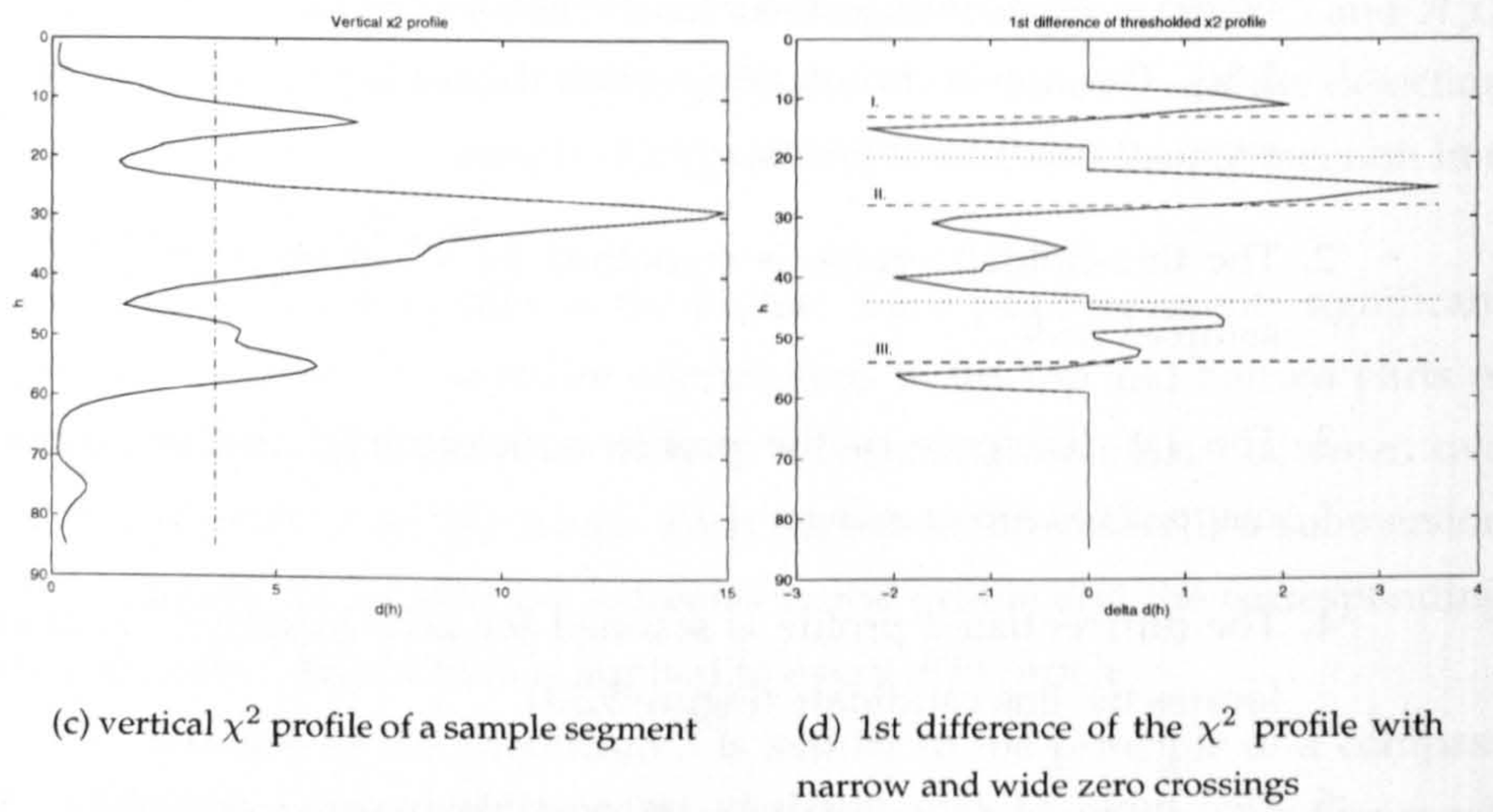
5. Two types of zero-crossings are possible: narrow and wide. Narrow crossing represents the extreme in the profile while wide crossing occurs due to thresholding. Only the narrow crossings are considered to be valid line candidates (Figure 5.5d, dashed lines represent the narrow crossings, dotted lines are the broad crossings).

6. The crossing at the lowest position in the segment is selected as a submersion line projection (Figure 5.5b, the thick dashed line corresponds to the chosen submersion line projection). A line at this position is most likely to represent the submersion line, any line above it might represent a boundary between various parts of the object.



(a) The structure of the scanning mask

(b) Submersion line candidates and the detected line (thick)



(c) vertical χ^2 profile of a sample segment

(d) 1st difference of the χ^2 profile with narrow and wide zero crossings

Figure 5.5: Detection of the submersion line.

of the top edge of the segment in the image. The adjustment accounts for the changes in scale in the scene due to perspective projection. The height of the sliding mask is larger for segments closest to the lower edge of the image and it decreases linearly in a series of discrete steps towards the horizon.

The height of the sliding mask would optimally follow the perspective projection profile (see Figure 4.5). The difference in the mask height at the bottom of the image to that at the top is, however, only a couple of pixels with typical values of 10 pixels for bottom position and 3 pixels for horizon position. Such a small range limits the capture of nonlinear perspective projection considering that the values of mask height have to be rounded to integer values. A linear change is therefore chosen as an approximation.

5.3 Feature-based Object Characterisation

Considering three advantages of low-level feature-based tracking mentioned by Shapiro (1995) - generality, opportunism and graceful degradation, feature based methods seem well suited for maritime scenes.

The generality is a necessary assumption considering the variability of objects in the scenes. Low-level features are local and therefore less constraining than a high-level object model. This makes them more general and feasible for applications where variable structure is encountered.

The objects in maritime scenes are mostly man-made and rigid. That means they have a fixed structure composed of straight or crossing lines with acute angles and homogeneous regions of differing intensities or textures. Such structures comprise points of local high intensity gradients that remain relatively unchanged when observed over multiple frames. These points can be detected and used in correspondence matching for motion estimation. Their appearance in the scene offers the opportunity for their tracking.

Another advantage of feature based methods is the reduction of data in the processing chain. Local features such as corners with a predefined neighbourhood occupy only a fraction of the original image, thus reducing the amount of data needed to be processed.

5.3.1 Corner Detectors

Despite an abundance of corner detectors (Smith and Brady, 1995; Sheng and Wang, 2000; Shen and Wang, 2001; Achard et al., 2000; Olague and Hernandez, 2002; Trajkovic and Hedley, 1998a; Ying and Lawrence, 1995; Cooper et al., 1993) in the image processing and machine vision areas, only a handful of them are considered competent for general applications. Three main groups of corner detection methods are recognised: contour based, intensity based and parametric model based. Both the contour and parametric model based methods rely on a substantial amount of prior knowledge, such as the location of thinned edges, which is not always available, particularly in a maritime scene. Corner detectors based on a parametric model are restricted by the number of corner types they can detect, typically L-type corners, (Olague and Hernandez, 2002; Wan-Ching and Rockett, 1997).

The main focus is, therefore, on the intensity based methods suitable for general machine vision applications. A thorough evaluation of intensity based

corner detectors is provided by Schmid et al. (1998). Based on the results presented in the evaluation study and by Smith and Brady (1995) two corner detectors were chosen as feasible for use in maritime scenes - SUSAN (Smith, 1992; Smith and Brady, 1995) and Harris (Harris and Stephens, 1988).

5.3.1.1 SUSAN Corner Detector

SUSAN corner detector, introduced by (Smith and Brady, 1995), is an intensity based edge and corner detector that employs a 'Univalue Segment Assimilating Nucleus' to detect one- and two-dimensional features in the image. The acronym 'SUSAN' stands for 'Smallest Univalue Segment Assimilating Nucleus'. The authors define the 'SUSAN' principle as:

> An image processed to give as output inverted USAN area has area edges and two dimensional features strongly enhanced, with the two dimensional features more strongly enhanced than the edges.

The functionality of the SUSAN detector consists of the following steps. A circular mask is applied to every pixel in the image. General diameter value of the mask is 3.4 approximated by a discrete mask with 37 pixels.

The pixels under the mask are weighted according to their intensity

$$c(\mathbf{r}, \mathbf{r_0}) = e^{-\left(\frac{f(\mathbf{r})-f(\mathbf{r_0})}{t}\right)^6} \tag{5.2}$$

where $\mathbf{r_0}$ is the position of the central point in the mask, $\mathbf{r}$ is the position of the actual point, $f(\mathbf{r})$ is the intensity at the position $\mathbf{r}$ and t is the similarity threshold. The weighted intensity differences $c(\mathbf{r}, \mathbf{r_0})$ are then added together over the entire mask

$$n(\mathbf{r}) = \sum_{\mathbf{r}} c(\mathbf{r}, \mathbf{r_0}) \tag{5.3}$$

The response of the filter is given by

$$R(\mathbf{r_0}) = g - n(\mathbf{r_0}) \quad if\ n(\mathbf{r_0}) < g \tag{5.4}$$

$$0 \quad otherwise \tag{5.5}$$

where g is a geometrical threshold given as $\frac{n_{max}}{2}$, where n_{max} is the maximum sum over the mask. A non-maximum suppression is performed by using a 5×5 pixels mask.

The results presented by Smith and Brady (1995) show very good localisation, noise robustness and a low number of false positives in detected features. The detection is also independent of the type of corner ('Y', 'T' and more complicated junctions are also detected). Another appealing fact is that no image derivatives are needed which explains the noise robustness of the detector. Also, the simplicity of the algorithm makes it a good candidate for real-time implementation as illustrated in (Smith, 1998).

Finally, the main advantage of the detector is that the detection sensitivity in terms of the number of detected corners is regulated by a single parameter t which has a clear, physical meaning. The parameter t is a threshold that represents the minimum intensity difference between USAN and surrounding pixels to indicate the presence of a corner.

When applied to maritime scenes (see Figure 5.6a,c) the SUSAN corner detector performs adequately, with precise localisation and sensitivity to low intensity salient features.

5.3.1.2 Harris Corner Detector

The corner detector by (Harris and Stephens, 1988) is an extension of Moravec's 'points of interest' detector. While Moravec used discrete displacements to determine the changes in image intensities that indicate the presence of an interest point, Harris employs autocorrelation matrix defined as

$$\mathbf{A}(r,c) = \begin{bmatrix} \sum_{(r_k,c_k)\in W}(\frac{\partial f(r_k,c_k)}{\partial c})^2 & \sum_{(r_k,c_k)\in W}\frac{\partial f(r_k,c_k)}{\partial c}\frac{\partial f(r_k,c_k)}{\partial r} \\ \sum_{(r_k,c_k)\in W}\frac{\partial f(r_k,c_k)}{\partial c}\frac{\partial f(r_k,c_k)}{\partial r} & \sum_{(r_k,c_k)\in W}(\frac{\partial f(r_k,c_k)}{\partial r})^2 \end{bmatrix} \tag{5.6}$$

where (r_k, c_k) are pixel positions in the window W centred at position (r, c) in the image and $\frac{\partial f}{\partial c}, \frac{\partial f}{\partial r}$ are directional image gradients. The window W is a two-dimensional Gaussian with a width parameter σ. The values in the sums are weighted by the window coefficients. If both eigenvalues of the autocorrelation matrix are large then the pixel is marked as a corner. To avoid eigenvalue decomposition, the corner response function is defined as

$$CRF(r,c) = \det(\mathbf{A}) - k \cdot trace(\mathbf{A})^2 \tag{5.7}$$

where $k = 0.04$, as suggested in the original paper.

After the $CRF(r, c)$ is determined for all pixels in the image a non-maxima suppression is performed. The size of the window used in non-

maxima suppression stage d is user defined. Finally, the remaining maxima in $CRF(r, c)$ are compared against another threshold T provided by the user.

The parameters, σ, d and T influence the performance of the detector. While σ and d have physical meanings, the values of threshold T are inferred by trial and error.

The major disadvantage of the detector lies in its computational overhead. For example, there are three multiplications by Gaussian coefficients over the window for every pixel. A number of modifications to the algorithm were presented. For example, Trajkovic and Hedley (1998b) suggest reducing the calculation of $CRF(x, y)$ to points with significantly high gradients.

The detector provides similar results to SUSAN when applied to maritime scenes as illustrated in Figure 5.6b,d.

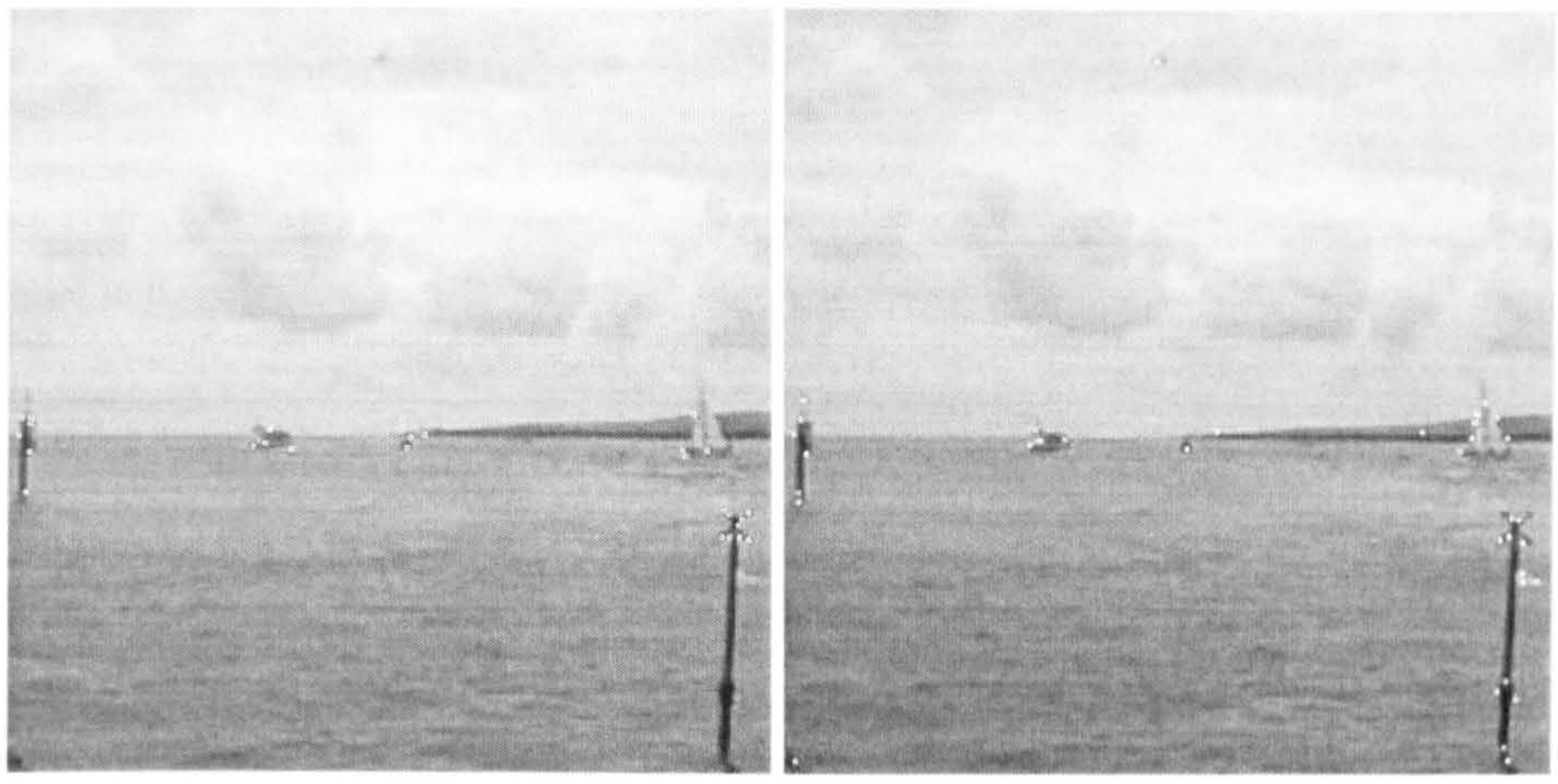
5.3.2 Comparison of The Detectors

A final choice of the detector suitable for maritime scenes is based on a following evaluation test. The test is designed to quantify the essential property of the detectors - the precision in corner localisation. The second quantity evaluated in the test is the total number of detected corners in each test image. When compared with the actual number of corners in the image, this measurement indicates the number of false positive detections that occur due to noise as well as any missed corners (detection dropouts). Similar tests are used by Trajkovic and Hedley (1998a). Nevertheless, they do not compare the results with a ground truth when evaluating the localisation precision.

5.3.2.1 Test Design

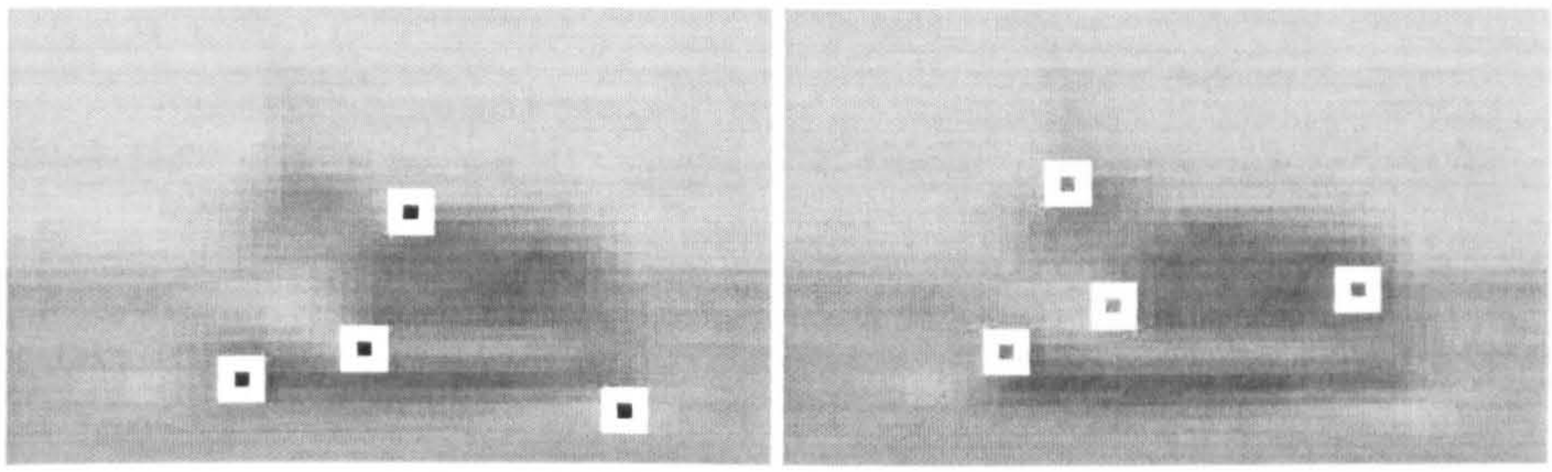
A testing sequence comprises of a square pattern of four homogeneous regions being rotated at a fixed angle around the pattern's centre. The rotation evaluates the geometrical invariance of the detectors. The background of the frames is constant. Each frame in the sequence is blended with Gaussian noise. The positions of the corners in each frame of the sequence are known precisely as the rotation of the square pattern is predefined. There are nine corners to be detected in the pattern, accounting for three typical corner junctions - L, T and X.

The sequence is subjected to both Harris and SUSAN detectors. Figure 5.7a,b shows a sample frame with corners found by both detectors. At first,



(a) SUSAN corner detector (t=20)

(b) HARRIS corner detector (sigma=2, r=3, T=10000)



(c) SUSAN - zoomed detail

(d) HARRIS - zoomed detail

Figure 5.6: Corner detection in sample maritime scenes using SUSAN (Smith and Brady, 1995) and Harris (Harris and Stephens, 1988) corner detectors.

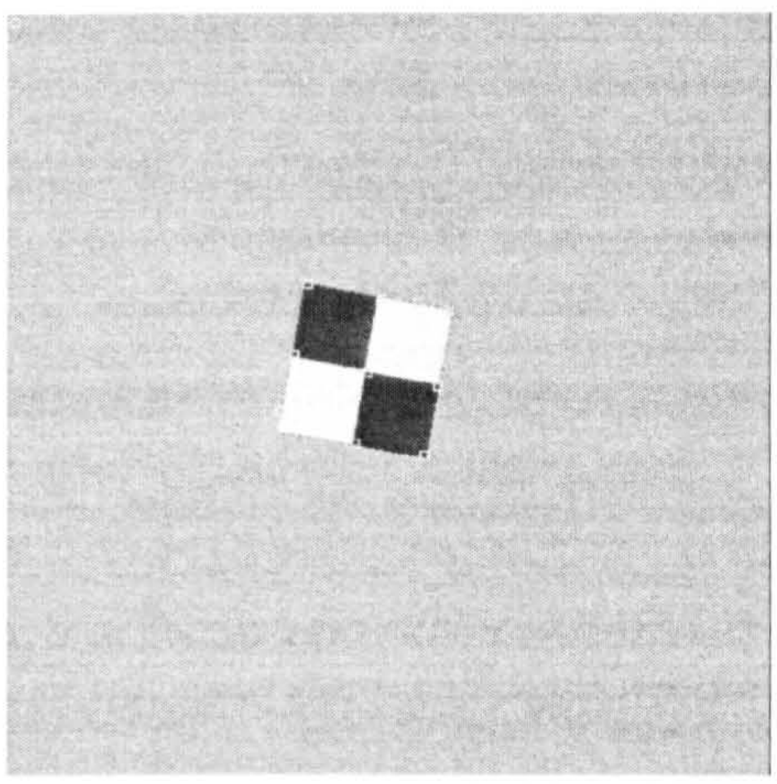
the total number of corners detected in each frame is stored. Then, the closest detected corner for each corner in the image is found and the offset is stored. If no corners are detected within the circular neighbourhood of a given radius the actual feature is marked as undetected. An average Euclidean distance of the closest detected corners from the actual ones is determined as the global displacement error for all frames in the sequence. This indicates the general localisation precision of each detector.

The parameters of the test are:

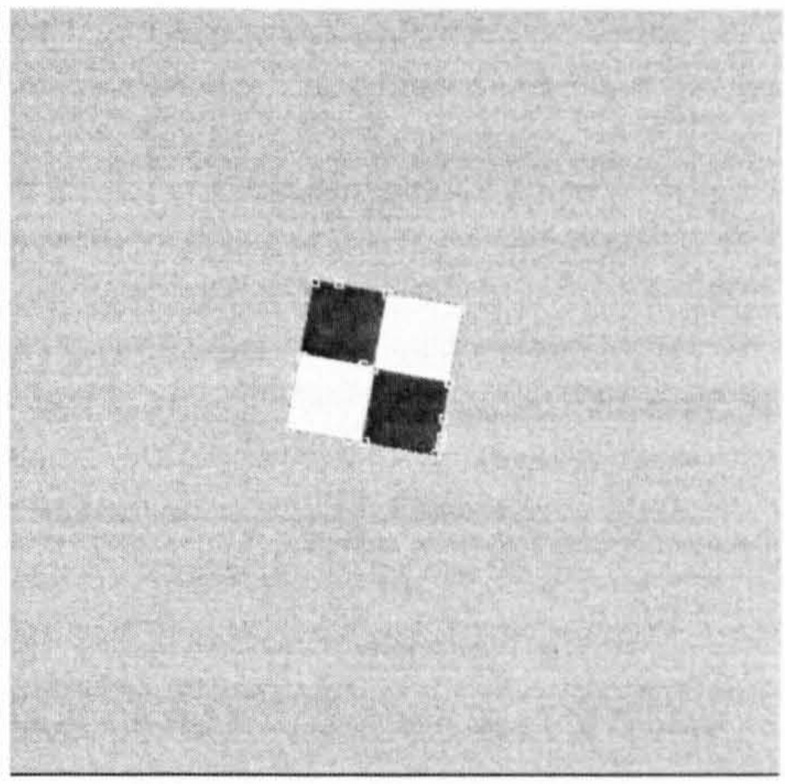
- test pattern: two pairs of squares (50×50 pixels) with alternating intensities of 20 and 195
- background (512×512 pixels): a constant 127 with added Gaussian noise ($\mu = 128,\ \sigma = 8$)
- sequence length: 20 frames
- rotation per frame: 5 degrees clockwise
- Gaussian noise parameters: $\mu = 128,\ \sigma = 8$ with unique instance for every frame; blending factor 0.5
- maximum corner matching radius: 7 pixels
- settings for the Harris corner detector: $\sigma = 2.5,\ d = 5,\ T = 15000$
- settings for the SUSAN corner detector: $t = 25$

5.3.2.2 Test Results

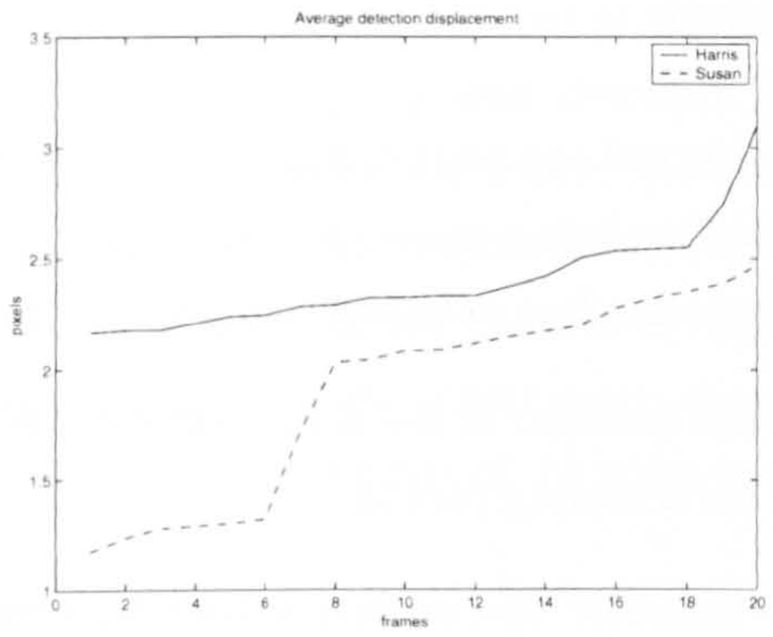
Both detectors found all the corners in the pattern in every frame of the sequence indicating that there were no detection dropouts. The results in Figure 5.7c show that the SUSAN detector localised the corners with a smaller global offset than the Harris detector. The higher localisation error of the Harris corner detector is well-known fact mentioned in a number of related works (Achard et al., 2000; Shen and Wang, 2001). The disadvantage of the SUSAN corner detector is the larger number of extra points detected (Figure 5.7d). This indicates that the SUSAN is less robust to noisy images, generating more false positive responses than the Harris detector.



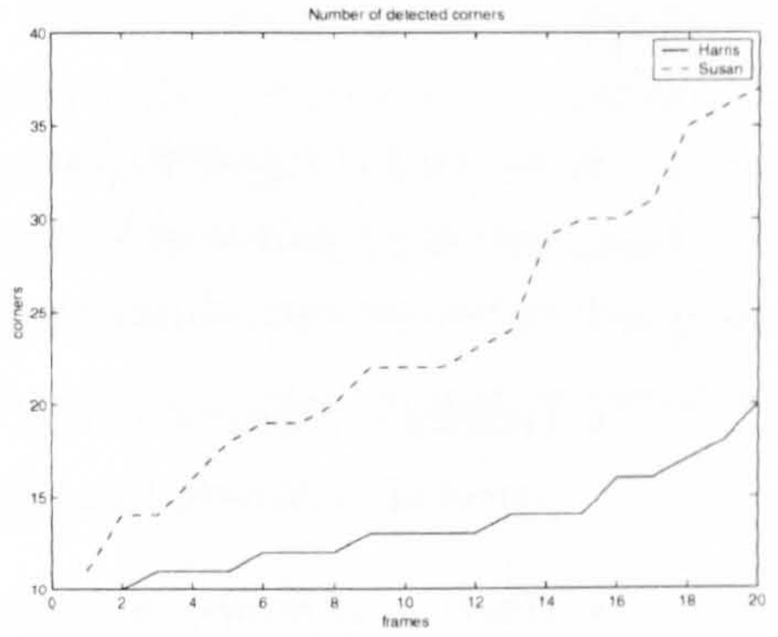
(a) Features detected by Harris corner detector



(b) Features detected by SUSAN detector



(c) Displacement error (values sorted increasingly)



(d) Number of features detected (values sorted increasingly)

Figure 5.7: Harris and SUSAN corner detectors - analysis of localisation precision. The values in graphs (c) and (d) are ordered according to the values for clearer illustration of the results. The order of the frames in the sequence has no importance to the test results.

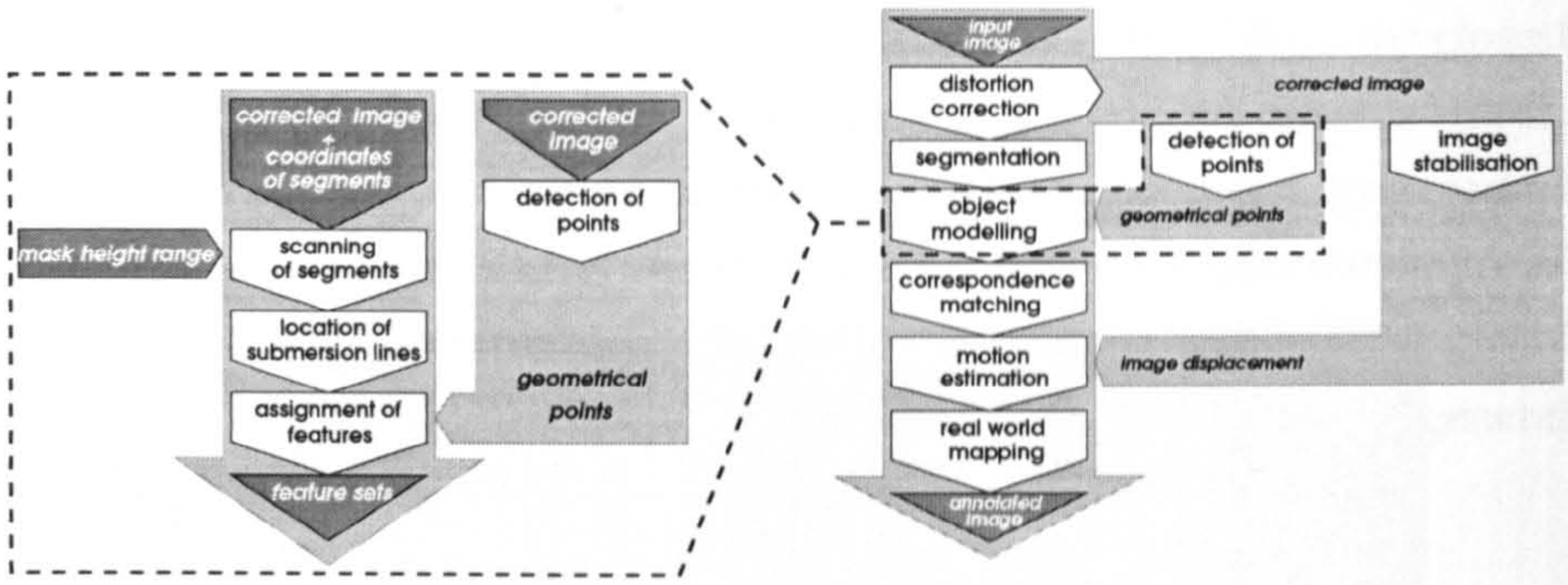


Figure 5.8: The structure of the object modelling module of the tracking system.

5.4 Structure of Object Modelling Module

The algorithms for detection of geometric features are assembled into an object modelling module that is a component of the framework. The structure of the module is shown in Figure 5.8.

The segmentation module takes as an input the current frame in the sequence and the coordinates of detected segments provided by the preceding segmentation module. The detection of features consists of the following four steps:

- *Scanning of segments.* Each detected segment is scanned for the submersion line candidates by vertically sliding mask. A scanning profile with peaks corresponding to submersion line candidates is generated.

- *Location of submersion lines.* The location of the actual submersion line is detected by thresholding of the scanning profile.

- *Detection of points.* A corner detector that detects salient features is applied to the input frame. Detected salient features represent geometrical points of interest within the scene.

- *Assignment of features.* Each detected segment is assigned corresponding vertical position of the submersion line and all geometrical points located within the segment.

The resulting feature sets represent the weak perspective models of objects within the scene. The sets are passed to the following processing module that establishes temporal correspondence between the sets.

5.5 Summary

A set of geometric features that represent an object in the scene are detected prior to motion estimation. An alternative approach to pixel-based segmentation of objects in the scene is employed that projects every object orthographically onto its average depth plane obtaining a weak perspective projection of the object. Geometric features are then bound to this plane.

The first feature that provides the location of the object in the scene is the line of submersion. A submersion line detection that uses a vertically sliding mask divided into two parts is proposed. The location of the submersion line is detected as a peak in a χ^2 profile.

Corner detection is applied to each segment in order to localise salient features characterising the object structure. Two candidates for corner detection are considered - SUSAN and Harris. An evaluation test is designed to compare their performances in terms of localisation precision. The results of the test indicate SUSAN to be superior in precision of corner localisation and it has the advantage of requiring a single user-defined parameter with clear physical meaning.

The planar representation of objects in the scene consists of the coordinates of the bounding segment detected during the primary segmentation, the vertical position of the submersion line and the geometric points - corners detected within the segment. All these geometric features are assumed to lie on a weak-perspective plane that represents the object structure. The planar representation of the object enters the process of motion estimation performed in the consequent parts of the framework.

Chapter 6

Correspondence Matching

6.1 Introduction

The motion of objects detected in the scene is estimated by temporal matching of geometric features detected at previous processing stages. Jain et al. (1995) states that the following three properties guide the matching process:

- discreteness - a measure of the distinctiveness of individual points,
- similarity - a measure of how closely two points resemble one another and
- consistency - a measure of how well a match conforms to adjacent matches.

Discreteness is assured by the use of corners as features for tracking. As discussed in the previous chapter, corners are suitable for tracking as they encode a high level of structural information. Another advantageous aspect is their spatial and temporal stability due to the fact that the corners are inherent in most man-made rigid objects which is the subject of the tracking.

Laws of physics, namely inertia and rigidity laws, and physical properties of rigid bodies constrain their possible motion. For example, the direction and velocity of motion do not change abruptly. The motion is assumed to be smooth with changes occurring only gradually. Furthermore, the rate of change of the velocity and direction of most rigid bodies in real world is usually much slower than the frame rate used in standard machine vision applications.

These assumptions imply that the projection of the object's structure between two consecutive frames in the sequence does not change significantly. As the structure is characterised by the geometric points, their resemblance will be preserved between the frames and will only degrade over a substantially longer period of time compared to the frame rate.

The inter-frame resemblance of geometric features is an essential principle in feature based applications such as (Smith, 1998; Shapiro, 1995). The transformation of the features between two consequent images is often modelled as affine, (Shi and Tomasi, 1994; Shapiro, 1995). Only displacement element of the affine transform is considered in maritime scenes, as rotation and scaling are limited and can be expressed in terms of displacement. This is due to a relatively high frame rate compared to speed of objects in the maritime environment.

The motion of the object in the scene projects into displacements of the detected geometric points. The task of finding these displacements involves a search for corresponding points between two consequent frames of the sequence. The correspondence search is typically based on evaluation of an affinity measure between two candidates for a correspondence match that quantifies their resemblance. An affinity measure based on correlation between image patches centered at the detected points is a typical example.

6.2 Affinity Measures for Corners

A traditional approach of correlating local intensity patches between frames is frequently used in machine vision applications, (Shapiro, 1995). The normalised cross-correlation measure is defined as

$$C(u,v) = \frac{\sum_{r,c=0,0}^{R-1,C-1} f(r,c)g(r+u,c+v)}{\sqrt{\left[\sum_{r,c=0,0}^{R-1,C-1} f(r,c)^2\right]\left[\sum_{r,c=0,0}^{R-1,C-1} g(r+u,c+v)^2\right]}} \tag{6.1}$$

where $f(r,c)$ and $g(r,c)$ are image patches, R,C are dimensions of the matching area and (u,v) is the offset at which the measure is determined.

Standard cross-correlation technique is used in many applications, either directly in a space domain or as a multiplication in a frequency domain, (Lewis, 1995). Because of the popularity, numerous methods for a fast calculation of cross-correlation are available. For example, a box filtering technique

presented by Changming (2002), together with subregioning of the matched images, significantly speeds up the cross-correlation process of the estimation of dense disparity maps commonly used in 'structure from depth' applications. Another optimisation approach exploits the fact that convolution becomes multiplication in the frequency domain, (Lewis, 1995).

Although these optimisations bring significant speed-ups when larger image regions are involved, for small patches the gain is minimal or contrary. These facts led to the development of other alternatives to cross-correlation. Smith et al. (1998) proposes the following affinity measure variants:

- zero mean cross-correlation (correlation coefficient)

$$C_{zm}(u,v) = \frac{\sum_{r,c=0,0}^{R-1,C-1} [f(r,c) - \bar{f}]\,[g(r+u,c+v) - \bar{g}(u,v)]}{\sqrt{\left[\sum_{r,c=0,0}^{R-1,C-1} [f(r,c) - \bar{f}]^2\right]\left[\sum_{r,c=0,0}^{R-1,C-1} [g(r+u,c+v) - \bar{g}(u,v)]^2\right]}} \tag{6.2}$$

- Sum of Squared Differences

$$C_{ssd}(u,v) = \sum_{r,c=0,0}^{R-1,C-1} [f(r,c) - g(r+u,c+v)]^2 \tag{6.3}$$

- χ^2 test

$$C_{\chi^2}(u,v) = \sum_{r,c=0,0}^{R-1,C-1} \frac{[f(r,c) - g(r+u,c+v)]^2}{[f(r,c) + g(r+u,c+v)]\,/2} \tag{6.4}$$

- Jeffrey divergence

$$C_{JD}(u,v) = \sum_{r,c=0,0}^{R-1,C-1} f(r,c) \log \frac{f(r,c)}{[f(r,c) + g(r+u,c+v)]\,/2} + g(r+u,c+v) \log \frac{g(r+u,c+v)}{[f(r,c) + g(r+u,c+v)]\,/2} \tag{6.5}$$

where the functions are the same as in Equation 6.1 and, in addition, the $\bar{f}$ and $\bar{g}$ are mean values of intensities in each patch. Even though the correlation coefficient does not provide any benefits in terms of simplification

of the measure, it is invariant to intensity offset and scaling. There are other two measures based on statistics also suggested by Smith et al. (1998) - Kolmogorov-Smirnov distance and Earth Mover distance. Smith et al. (1998) shows that Kolmogorov-Smirnov distance does not outperform the standard cross-correlation. Earth Mover distance that relies on a complex strategy of linear programming outperforms the standard cross-correlation. Nevertheless, it is outperformed by all other approaches except the zero-mean cross-correlation. These two measures are therefore excluded from the following evaluation for their poor performance and complexity.

6.2.1 Performance Evaluation of the Affinity Measures

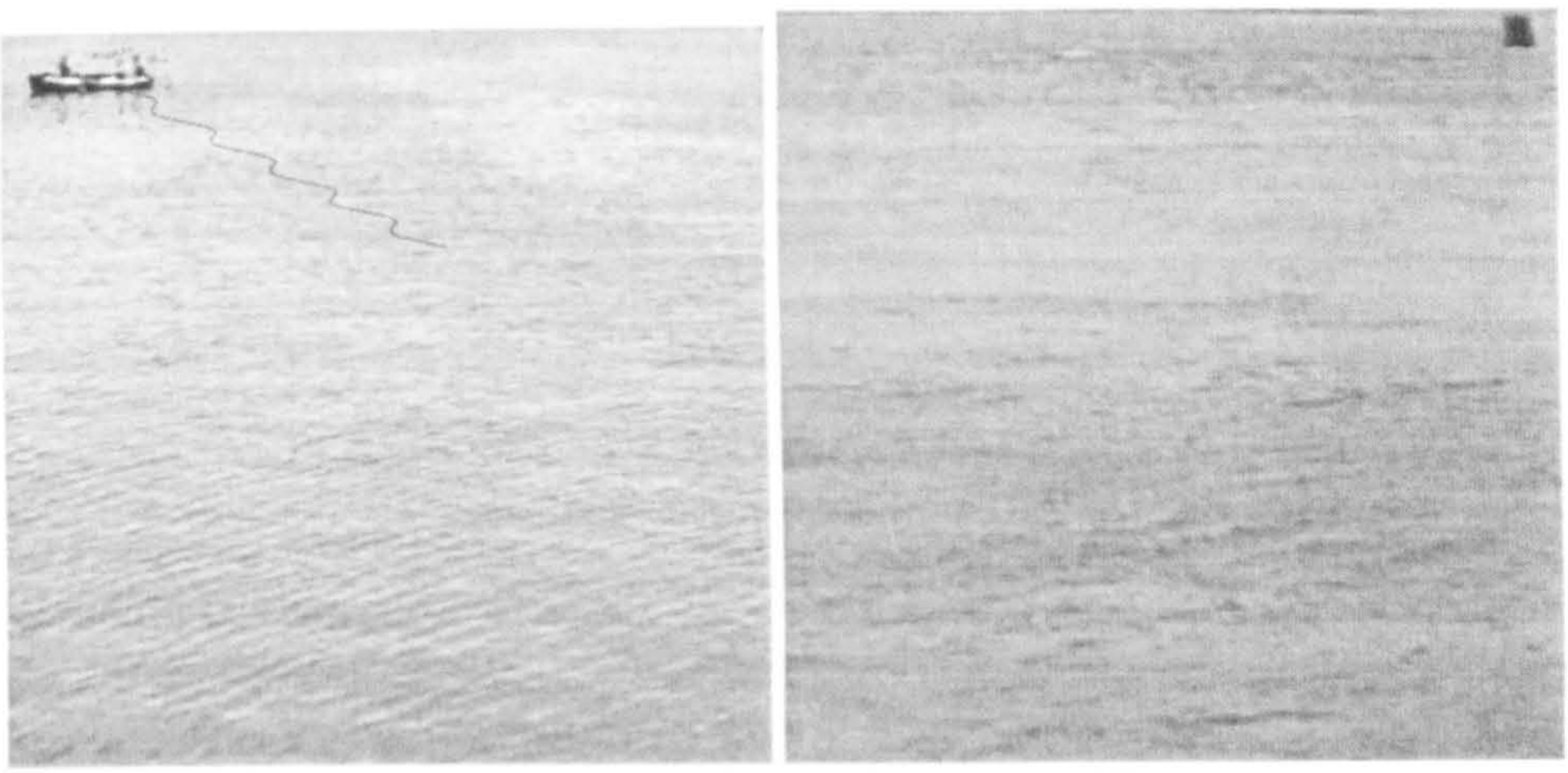
To assess performance of each measure in the intended maritime tracking task a set of evaluation tests is conducted. Multiple artificial sequences with moving targets are generated and then subjected to the correspondence matching using different measures defined above. The deviations between the actual and detected displacements are evaluated as indications of performance.

For such testing a knowledge of ground truth motion is a prerequisite. It would be rather challenging to obtain such information from real scenes, as the precise motion of the selected objects would have to be obtained. Instead, an artificial sequence is generated by overlaying an image of a target (vessel, buoy, etc.) over a background containing the sea as described in Section 2.7.3.3. The overlaid target is displaced by a known amount of pixels in either direction in each frame of a sequence. Gaussian noise with predefined parameters is added to each frame as well to simulate the effects of real noise generated during the image capture process. As Figure 6.1a shows, the resulting sequence is very close to a natural one.

Two artificial sequences were generated for the evaluation. The first sequence contains a large, highly structured vessel with approx. 70 detected corners in each frame. The second sequence contains a small rowing ferry with approx. 10 detected corners in each frame. Both objects undergo the same motion described by following recursive relations

$$r_{n+1} = r_n + 1\,[pix] \tag{6.6}$$

$$c_{n+1} = c_n + 2 + 3\sin(\frac{\pi n}{20})\,[pix] \tag{6.7}$$



(a) Artificial scene with target path superimposed

(b) Real scene

Figure 6.1: Sample frames from sequences used in evaluation of affinity measures.

where (r_n, c_n) and (r_{n+1}, c_{n+1}) are object locations in current and next frames and n is the frame index.

A real scene with a static object (a tied-up buoy) is also used in the evaluation testing (see Figure 6.1b). As the buoy is static it is possible to estimate the ground truth - the displacements of the buoy are assumed to be zero in both directions.

All three sequences have a similar length of about 100 frames. The size of the mask used in affinity measures is set to 9×9 pixels.

The summary of results is presented in Tables 6.1, 6.2 and 6.3 with dx and dy corresponding to localisation errors in horizontal and vertical directions. Localisation errors represent the discrepancy between the detected and the actual locations of the objects in the sequences. The different data fusion methods are also tested. Mean and median fusion methods simply delimit the resulting displacement for the object as either mean or median of all displacements of matched corners within a segment. The weighted mean method weighs each displacement by a coefficient proportional to the length of the trace to which the corner is assigned. Such a scheme preserves long coherent traces. The construction and maintenance of traces is discussed in detail in Section 6.5.

As the results show the errors are similar for all affinity measures tested.

	mean		median		weighted mean	
error [pix]	dx	dy	dx	dy	dx	dy
JEFFREY	0.91	0.24	0.37	0	0.78	0.2
SSD	0.7	0.22	0.37	0	0.63	0.2
X2	0.77	0.21	0.37	0	0.73	0.2
XCORR	0.68	0.21	0.37	0	0.64	0.2
ZM-XCORR	0.91	0.24	0.37	0	0.78	0.2

Table 6.1: Displacement errors of affinity measure variants for artificial motion - SCENE01

	mean		median		weighted mean	
error [pix]	dx	dy	dx	dy	dx	dy
JEFFREY	0.81	0.21	0.36	0.01	0.81	0.21
SSD	0.63	0.13	0.37	0	0.58	0.13
X2	0.69	0.15	0.36	0	0.7	0.16
XCORR	0.65	0.18	0.36	0	0.64	0.18
ZM-XCORR	0.81	0.21	0.36	0.01	0.81	0.21

Table 6.2: Displacement errors of affinity measure variants for artificial motion - SCENE02

All values are below one pixel. Noticeable differences in results are evident in the fusion methods. From all three methods the median fusion provides the best results. This is expected, as the median is insensitive to outliers in the data typically caused by false matches inconsistent with true displacements. Because of the feasibility for fast and efficient implementation (Nickels and Hutchinson, 2002), SSD in combination with median fusion is the preferred method used for correspondence matching in maritime scenes.

	mean		median		weighted mean	
error [pix]	dx	dy	dx	dy	dx	dy
JEFFREY	0.18	0.21	0.18	0.21	0.18	0.21
SSD	0.19	0.2	0.19	0.2	0.19	0.2
X2	0.18	0.21	0.18	0.21	0.18	0.21
XCORR	0.18	0.2	0.18	0.2	0.18	0.18
ZM-XCORR	0.18	0.21	0.18	0.21	0.18	0.21

Table 6.3: Displacement errors of affinity measure variants for real scene - SANDBANKS2R

6.3 Corner Correspondence Search

A detailed description of the feature matching and tracking algorithm is provided in a seminal work on affine motion analysis of image sequences by Shapiro (1995). Two essential parts of Shapiro's framework are corner matcher and correspondence tracker. The matching is based on an evaluation of zero mean cross-correlation similarity measure defined by Equation 6.2. To ensure a high level of generality throughout the framework great attention is paid to resolution of possible matching ambiguities. The tracking phase of the algorithm deals with dropouts by a simple prediction technique based on constant velocity or acceleration linear predictors. Because of the suitability of the feature-based technique for motion estimation of rigid objects in the sequence, the technique utilised in maritime tracking draws mainly from the aforementioned work of Shapiro.

6.3.1 Spatial Proximity Constraint

The worst-case scenario of matching search for two groups of detected corners in consequent images is to determine the affinity measures for all possible pairs. If all possible combinations of pairs are involved the total number of matchings would be $N_p N_c$ where N_p and N_c are the numbers of detected features in the previous and the current frames.

To avoid such an exhaustive evaluation a set of constraints that excludes unlikely matches is usually imposed onto the feature pairs. A frequently applied constraint is a spatial one: assuming that the motion of rigid objects in the scene implies gradual change of velocity and direction, the inter-frame displacements are relatively small (a couple of pixels in most cases). Therefore, the evaluation is reduced only to feature pairs that are no further apart than a certain predefined distance. Same constraint is utilised in the maritime tracking system - it is assumed that the inter-frame displacement of an object is less than 7 pixels in any direction. The value is adequate for maritime scenes used in the development of the framework. The value can be adjusted taking into an account the relation between the velocity of objects and their projected displacements as discussed in Section 2.4.3.

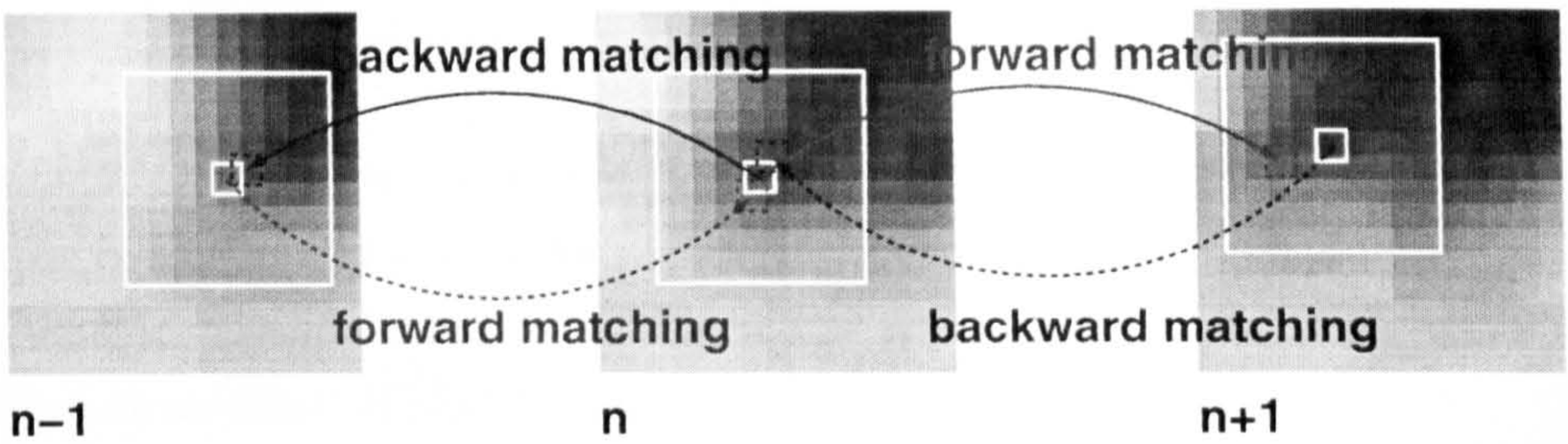


Figure 6.2: Mutual matching scheme. The corner detected in current frame n is matched against candidates in previous $(n-1)$ and next $(n+1)$ frames.

6.3.2 Mutual Matching

The principle of mutual matching based on a mutual affinity is shown in Figure 6.2. The mutual matching is necessary in order to resolve matching ambiguities (Shapiro calls them 'love triangles'). A greedy approach is used for matching, i.e. every corner is matched to all its proximate candidates with no constraints imposed on resulting smoothness or orientation of the resulting path, as any possible outliers are resolved by median filtering.

The mutual affinity measure is calculated in two steps. In the first step, so-called forward matching from previous to current frames finds the affinity between a corner in a previous frame and all valid candidates in the current frame. In the second step, a corner in the current frame is matched against its valid counterparts in the previous frame in so-called backward matching.

6.3.3 Stable Complete Matching

The matching results in a pair of sparse correspondence matrices for forward and backward matching with elements representing the affinity values between corners detected in both consecutive frames. The correspondence matrices will necessarily contain multiple ambiguous matches when one corner attracts more candidates and some of the corners in previous or current frames can remain without any candidate. To resolve such cases the information encoded in the matrices must be pruned leaving just unambiguous one-to-one correspondence between matched corners. All unmatched corners in previous and current frames must also be identified. The procedure can be reformulated as bi-partite matching.

Sara (1999) introduces a methodology of the bi-partite matching based on a concept of a stable complete matching for an application in stereo vision. The

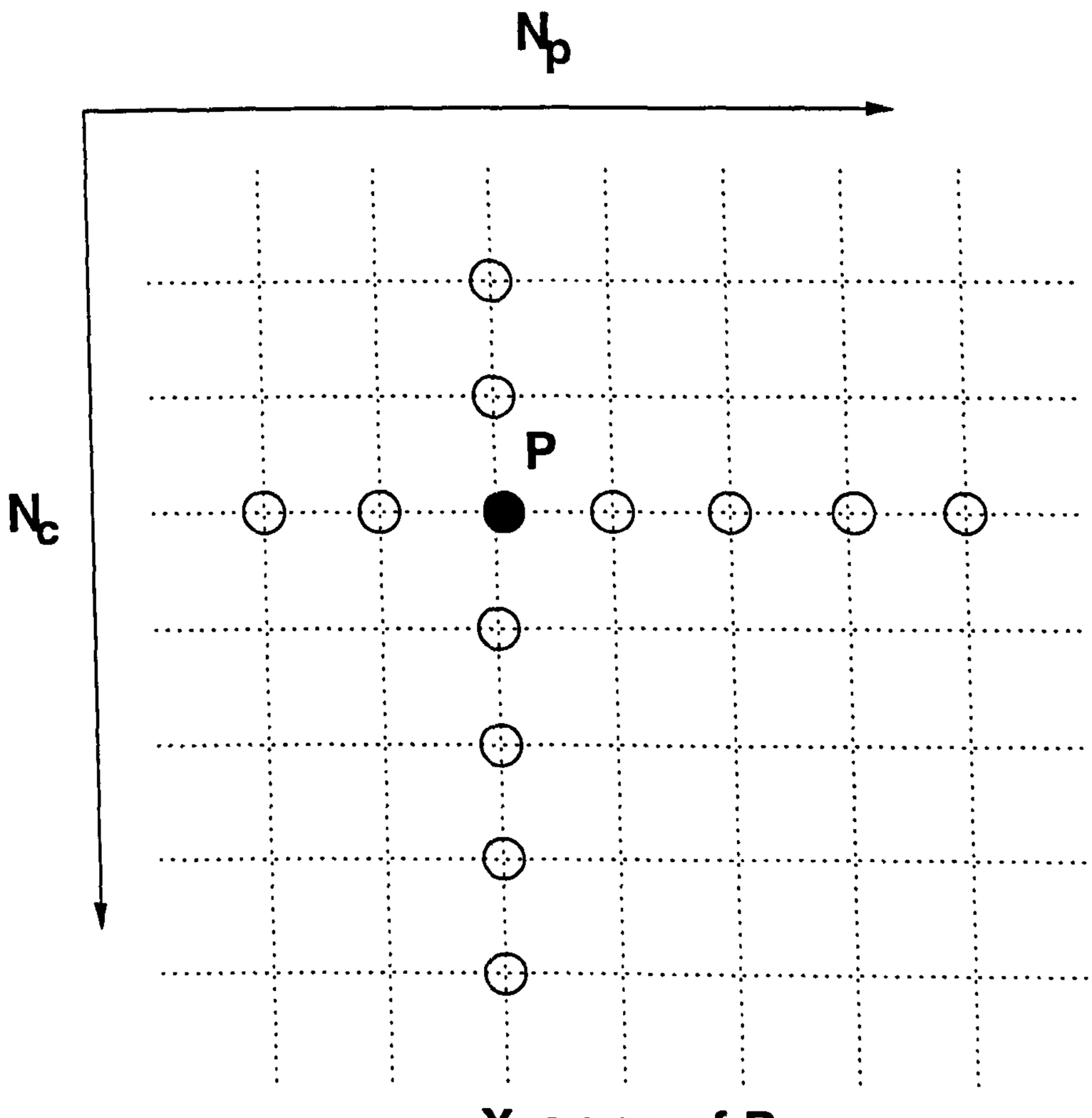


Figure 6.3: Definition of $X(P)$ zone (empty circles only) of an element in correspondence matrix (P).

concept, however, can be extended to any bi-partite matching problem.

Elements of the correspondence matrix are ranked with respect to their values. A zone $X(P)$ of an element P in the correspondence matrix is defined as shown in Figure 6.3.

X-dominant matching is defined such that the element P has the highest rank of all points in $X(P)$. Because there is no other element $Q \in X(P)$ such that it has a higher rank than element P then the X-dominant matching is also considered stable. If all the pairs are uniquely matched then the matching is pronounced stable and complete. Algorithm 4 to obtain stable complete matching is set out in (Sara, 1999).

Algorithm 4 Stable Complete Matching Algorithm as presented by Sara (1999).

1. Form a list L of all elements of correspondence matrix and sort them in descending order according to their rank. Initialise M (a set of elements representing pairs successfully matched) to an empty set.

2. If L is empty, terminate. The set M is a stable complete matching.

3. Let p be the first element in L. Add p to M and remove p together with all $q \in X(p)$ from L.

4. Go to step 2.

6.3.4 Modified Stable Complete Matching

Modifications to the matching algorithm provided by Sara (1999) are introduced in order to suit the problem of corner correspondence search. To change the asymmetry of the mutual matching due to directional (forward and backward) affinity measures C_F, C_B a mutual affinity C is proposed

$$C = \frac{C_F + C_B}{1 + |C_F - C_B|} \tag{6.8}$$

As Figure 6.4 shows the mutual affinity is symmetrical and it prioritises stronger and symmetric matches. The original matching method operates on rankings of the matches to ensure the independence of the measure used. The mutual affinity C that increases monotonically in directions of increasing affinities C_F and C_B enables these rankings to be established.

Another modification arises due to the fact that the algorithm in (Sara, 1999) assumes that the number of rows of the correspondence matrix equals to the number of columns, $N_p = N_C$. This assumes that there is the same number of points to be matched across the frames and the correspondence matrix is square and symmetrical (true bi-partite matching).

In corner matching the numbers of corners detected in each frame can vary and the correspondence matrix is not necessarily square. Furthermore, even for an equal number of corners some of them could remain unmatched in cases of proximity constraint violation, i.e. the corners are not close enough. A modified Algorithm 5 is presented that handles unmatched and new corners as well.

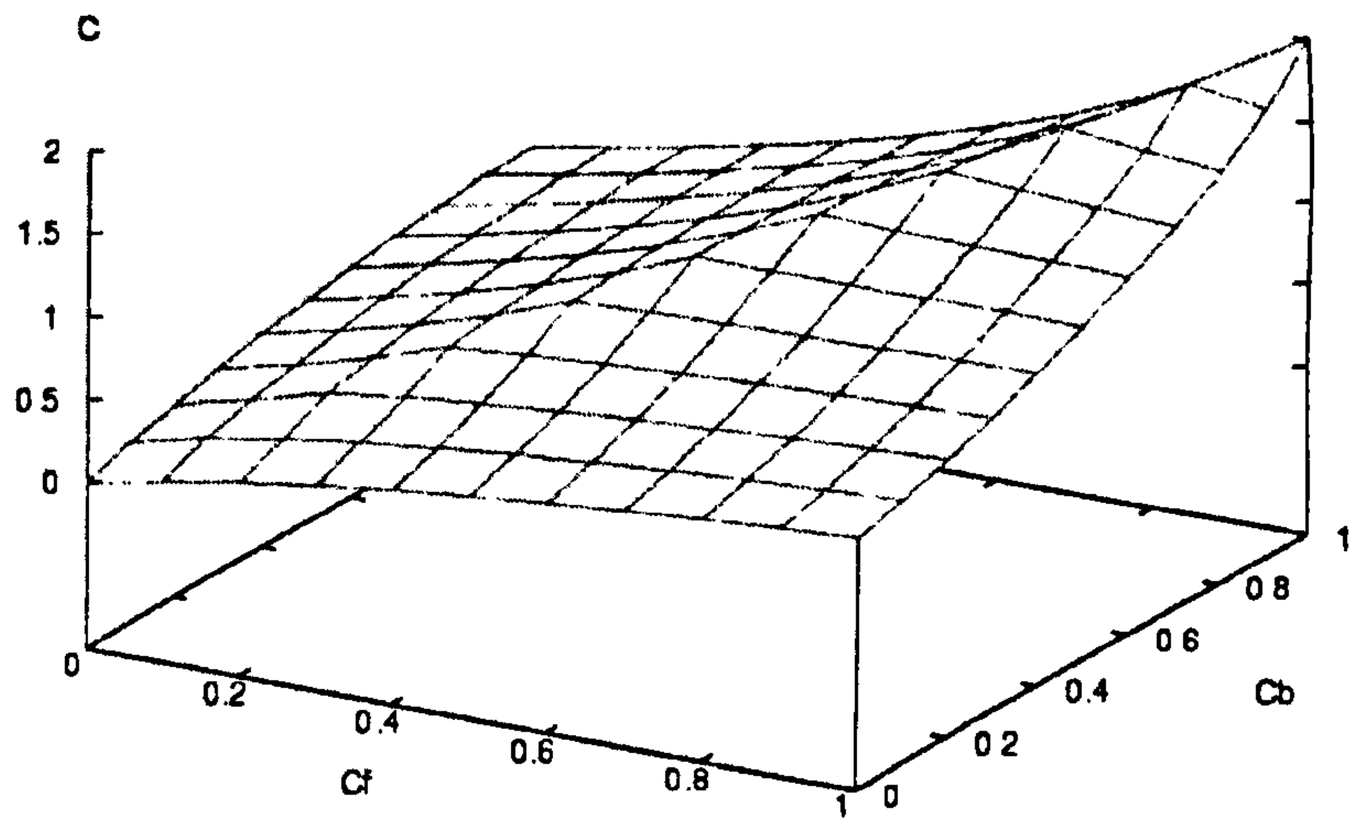


Figure 6.4: Mutual affinity as a function of normalised forward C_F and backward C_B affinities.

6.4 Segment Correspondence Search

The objects detected in maritime scenes are represented by segments in which they are likely to occur. The corners, together with the submersion line, are bound to these segments. To find a temporal correspondence between segments detected in each frame the segments have to be matched across frames in a way similar to the corners. In fact, the same Algorithm 5 is used with a modified affinity measure.

The same assumptions as in the case of corner matching are applied to segment matching. The velocity of rigid objects in the scene is limited by their physical properties. The displacements in the image are restricted to a couple of pixels, the image displacement is generally much smaller than the size of the object itself due to the sufficient frame rate.

When the object in the scene moves the segment does not change significantly in size and is displaced or resized by the amount corresponding to multiples of the overlap in the segmentation grid, either horizontal, vertical or both, depending on the direction of motion. In conclusion, corresponding segments will overlap either partially (rapid moving objects) or completely (slow moving objects, static objects) between frames.

Algorithm 5 Modified Stable Complete Matching Algorithm that identifies new and unmatched corners.

1. Scan the correspondence matrix for any empty rows and columns.
2. If an empty column is found, mark the corresponding corner in previous frame as unmatched.
3. If an empty row is found, mark the corresponding corner in the current frame as new.
4. Reduce the correspondence matrix by all empty rows and columns.
5. Apply Algorithm 4 on the reduced correspondence matrix.

6.4.1 Affinity Measure

The measure for matching the segments is based on the amount of overlap between the segments. If an object appears in the segment and it is either static or moving the detected segments in which it will appear in consequent frames will either completely or partially overlap. Forward and backward affinity measures are defined as relative overlaps between segments, namely

$$C'_F = \frac{(r_{right} - r_{left})(c_{bottom} - c_{top})}{(r^{prev}_{right} - r^{prev}_{left})(c^{prev}_{bottom} - c^{prev}_{top})} \tag{6.9}$$

$$C'_B = \frac{(r_{right} - r_{left})(c_{bottom} - c_{top})}{(r^{curr}_{right} - r^{curr}_{left})(c^{curr}_{bottom} - c^{curr}_{top})} \tag{6.10}$$

where

$$r_{left} = \max(r^{prev}_{left}, r^{curr}_{left}) \tag{6.11}$$

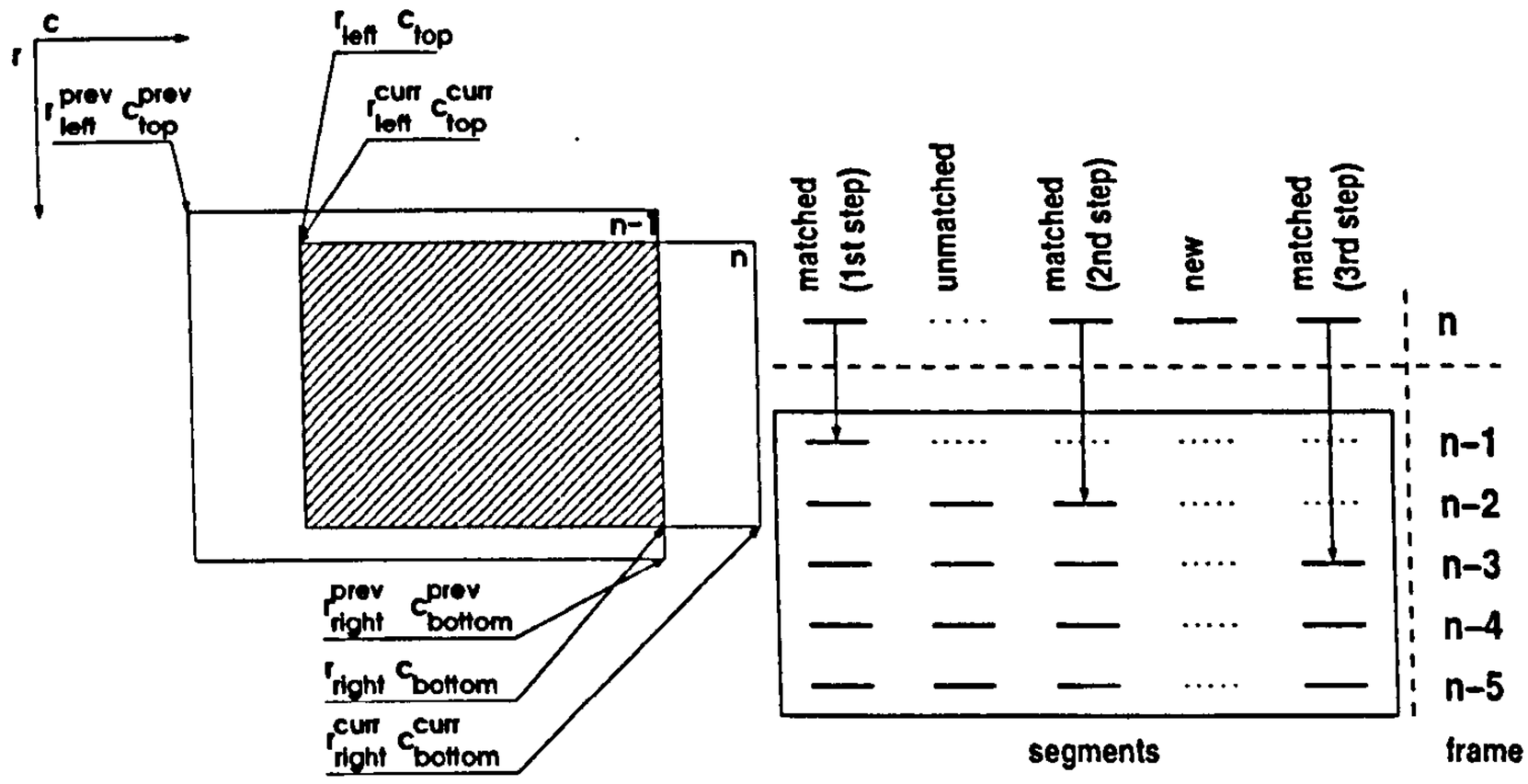
$$r_{right} = \min(r^{prev}_{right}, r^{curr}_{right}) \tag{6.12}$$

$$c_{top} = \max(c^{prev}_{top}, c^{curr}_{top}) \tag{6.13}$$

$$c_{bottom} = \min(c^{prev}_{bottom}, c^{curr}_{bottom}) \tag{6.14}$$

with $(r^{prev}_{left}, r^{prev}_{right}, c^{prev}_{top}, c^{prev}_{bottom})$ and $(r^{curr}_{left}, r^{curr}_{right}, c^{curr}_{top}, c^{curr}_{bottom})$ corresponding to the coordinates of the segments in previous and current frames respectively (see Figure 6.5a). The mutual symmetric affinity is then given as an average overlap measure

$$C' = \frac{C'_F + C'_B}{2} \tag{6.15}$$



(a) segments coordinates used in affinity measures

(b) segments matching against history stack

Figure 6.5: Matching of segments using their overlap (a). The temporal stack contains a list of all segments detected in last 5 frames (b).

6.4.2 Temporal Stack

Temporal dropouts are prevented by keeping a stack of segments detected in the scene for a given number of frames. The segments in frame n are matched against all segments in frame $n-1$. If there remain some unmatched segments in frame n the matching is repeated against unmatched segments in frame $n-2$. The matching is then repeated for any unmatched segments in frame $n-3$ and so on until the last frame in the stack is reached. If there are still some segments unmatched they are labelled as new and stored (see Figure 6.5b).

6.5 Feature Traces

For every newly-detected and unmatched corner a new trace is spanned. Traces describe the motion of each corner across the sequence by storing the information about all previously matched candidates for that particular corner. A single trace corresponds to a single corner. Even though the traces are not essential for determining the motion of the objects (this is done in the following module using Kalman tracking and fusion of inter-frame displacements), they are convenient for maintenance of the coherence of the object path and also

allow to predict the position of the corners and their displacements in case a dropout in corner detection occurs.

6.5.1 Position Prediction

When the corner in the previous image remains unmatched due to a dropout in corner detection, or when all possible candidates for a match are further apart than the proximity constraint permits, the trace would have to be terminated which would cause a loss of valuable information about the motion of the object. Should the corner re-appear in the following frame a new trace would be spanned starting where the previous one terminated.

To avoid losses due to short-term detection dropouts a simple prediction scheme is used that tries to establish the position of the undetected corner from previous displacements. While some approaches use a Kalman filter, (Galvin et al., 1999a), for the majority of tracking applications where rigid objects are involved such an approach is often unnecessary. For example, at least one four-dimensional matrix and one four-element vector would have to be updated and stored for every trace and every frame in the sequence if both position and velocity are taken as elements of the motion state of the corner. Instead, Shapiro (1995) suggests predictors based either on constant velocity ($\frac{\partial^2 \mathbf{x}}{\partial t^2} = 0$) or constant acceleration ($\frac{\partial^3 \mathbf{x}}{\partial t^3} = 0$) assumptions.

The first one, called linear predictor, can be expressed after discretisation of the differential as

$$\mathbf{x}(n+1) = 2\mathbf{x}(n) - \mathbf{x}(n-1) \tag{6.16}$$

and the second one, called quadratic predictor, as

$$\mathbf{x}(n+1) = 3\mathbf{x}(n) - 3\mathbf{x}(n-1) + \mathbf{x}(n-2) \tag{6.17}$$

where $\mathbf{x}(n)$ is the location of the corner in the frame n. The linear predictor needs two previous successful matches to establish a new position of the feature while the quadratic predictor needs three such matches. In maritime scenes the rigid objects mainly undergo motion with constant velocity as acceleration changes only gradually due to the nature of the environment. The linear predictor therefore produces results that are generally in accordance with the actual motion of objects in maritime scenes.

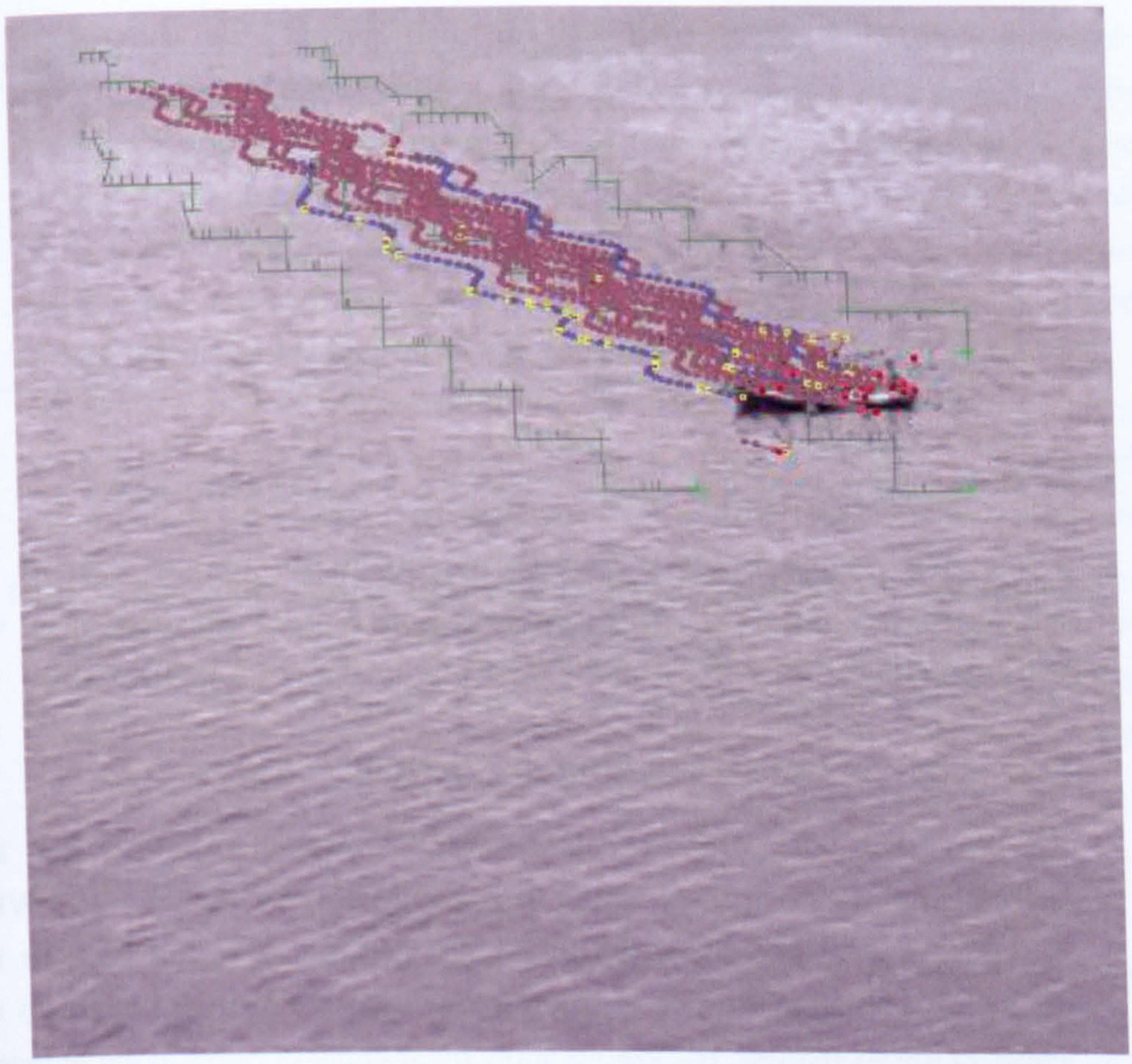
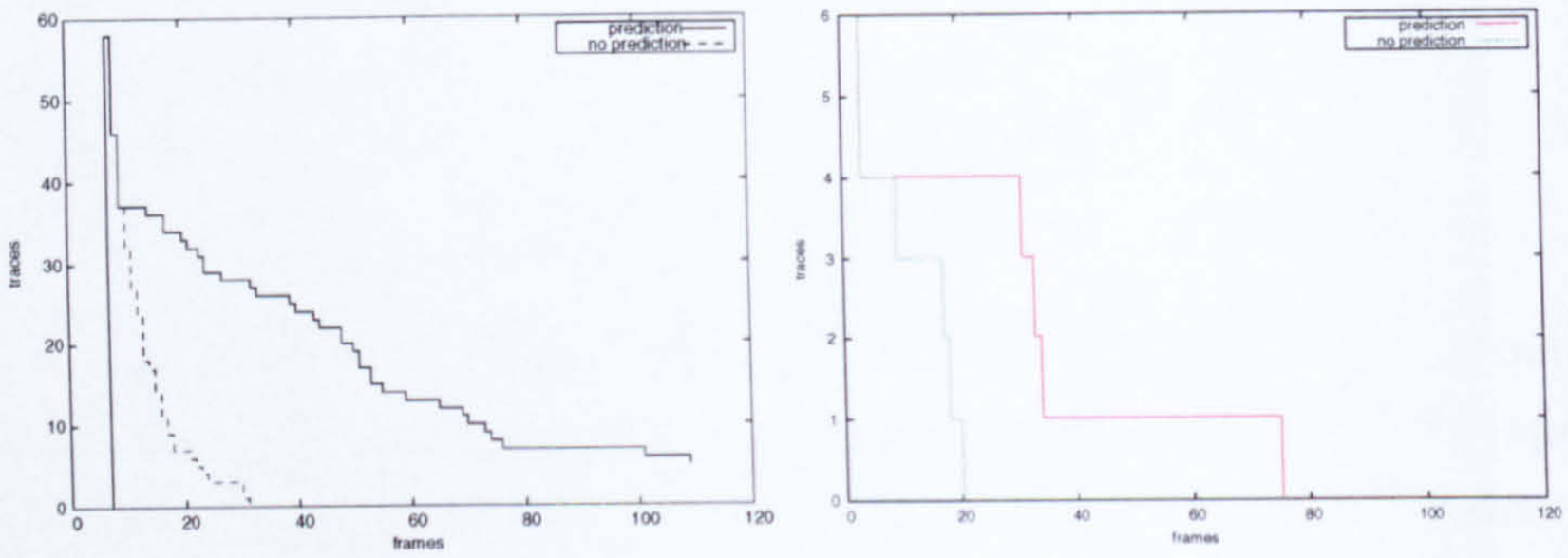


Figure 6.6: Detected traces in the artificial sequence. Red markers indicate traces with corners matched in the current frame, yellow markers indicate the predicted corners and blue markers are without a successful match in the current frame.

The trace is terminated when five consequent predictions occur without any successful match with a detected corner. This enables traces that tend to "stray" or those that ceased due to occlusion in the scene to be terminated.

An example of traces detected in the artificial scene (see Figure 6.1a) is shown in Figure 6.6. The yellow markers indicate corners obtained by prediction. Red markers correspond to matched corners. Blue markers are traces without a match in the current frame. Green markers outline locations of the matched segments.



(a) Artificial (b) WEYMOUTH2A

Figure 6.7: Evaluation of position prediction. The graphs indicate that the number of the longest traces decreases rapidly without the prediction. The prediction maintains long, coherent traces.

6.5.2 Evaluation of Position Prediction

The benefits of the position prediction are illustrated in Figures 6.7a,b. The artificial sequence described as the first sequence in Section 6.2.1 and WEYMOUTH2A sequence are subjected to the tracking with and without the prediction. Relative lengths of the traces are measured with respect to the length of the sequences. The graphs show the number of traces in the sequence with the same length as that of the sequence, i.e. the longest possible traces. The graphs indicate that without the prediction the number of coherent long traces decreases rapidly with the position in the sequence. The relative average lengths of the traces in the artificial sequence are 57% with prediction and 45% without prediction and 37% with prediction and 32% without prediction in the WEYMOUTH2A sequence. This indicates that prediction maintains long traces by filling the gaps caused by dropouts in corner detection.

6.5.3 Sub-pixel Localisation

To improve the precision of tracking and to allow detection of displacements less than a pixel per frame, sub-pixel techniques are commonly used in tracking applications.

A comprehensive study of sub-pixel precision in motion estimation is provided by Borman et al. (1999). The authors infer a mathematical model of a transfer function of an optical system with a CCD imaging device as a

combination of the characteristic function of the optical device and the point-spread function of the CCD chip. The resulting model shows significant anti-aliasing occurring due to the limited resolution of the imaging device. The authors employ the inferred model in tests that evaluate three commonly used block-matching motion estimators, namely Sum of Absolute Differences, Mean Square Error and Normalised Correlation. In the test, a step edge is displaced and the displacement is determined using the above-mentioned estimators.

Borman et al. (1999) conclude that an achievable sub-pixel resolution is firmly limited and cannot be further improved beyond a certain minimum. The minimum is based on statistical distribution of the residual errors that should be uniformly distributed on the interval given as $\pm\frac{1}{2p_{res}}$ where p_{res} is a reciprocal value of the desired precision. The results show that for all three methods the value for which the residual errors are uniformly distributed inside the interval is around 5, which leads to achievable precision of about ± 0.1 pixel for a standard CCD imaging device.

6.5.3.1 2D Interpolation

The most common technique of sub-pixel localisation is based on interpolation of the cross-correlation surface at maximum peak and surrounding values by an analytical function of two variables. Correlation surface is composed of correlation values obtained and located at different offsets (u, v) in Equations 6.1-6.5. The analytical function typically fitted is a paraboloid (Gleason et al., 1991)

$$f(u,v) = au^2 + bv^2 + cuv + du + ev + f \tag{6.18}$$

The parameters of the paraboloid are obtained by solving a set of linear equations

$$\begin{bmatrix} C(u_{-1},v_{-1}) \\ C(u_0,v_{-1}) \\ \vdots \\ C(u_{-1},v_0) \end{bmatrix} = \begin{bmatrix} u_{-1}^2 & v_{-1}^2 & u_{-1}v_{-1} & u_{-1} & v_{-1} & 1 \\ u_0^2 & v_{-1}^2 & u_0v_{-1} & u_0 & v_{-1} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_{-1}^2 & v_0^2 & u_{-1}v_0 & u_{-1} & v_0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} \tag{6.19}$$

or

$$\mathbf{C} = \mathbf{A}\mathbf{b} \tag{6.20}$$

where $C(u, v)$ is the correlation value at position (u, v) and (u_m, v_n); $m, n = -1, 0, 1$ are coordinates of points surrounding the minimum at position u_0, v_0. A pseudo-inverse method provides a closed solution

$$\mathbf{b} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{C} \tag{6.21}$$

From there, the sub-pixel location of the interpolated maximum of the correlation surface is obtained as

$$u_{max} = \frac{(2db - ce)}{(c^2 - 4ab)} \tag{6.22}$$

$$v_{max} = \frac{(2ae - dc)}{(c^2 - 4ab)} \tag{6.23}$$

Even though the method provides the results in a simple, closed form the results are often unstable and sensitive to noise. Such a case is illustrated in Figure 6.8. There is a clear maximum present, surrounded by values along the diagonal that are close to that maximum. The fitted paraboloid has the maximum at position $(-4, -6.5)$ which is clearly the wrong location of a true maximum. A non-maxima suppression method finds the correct maximum at pixel resolution, but the interpolation method fails, introducing a significant error into the location of the match.

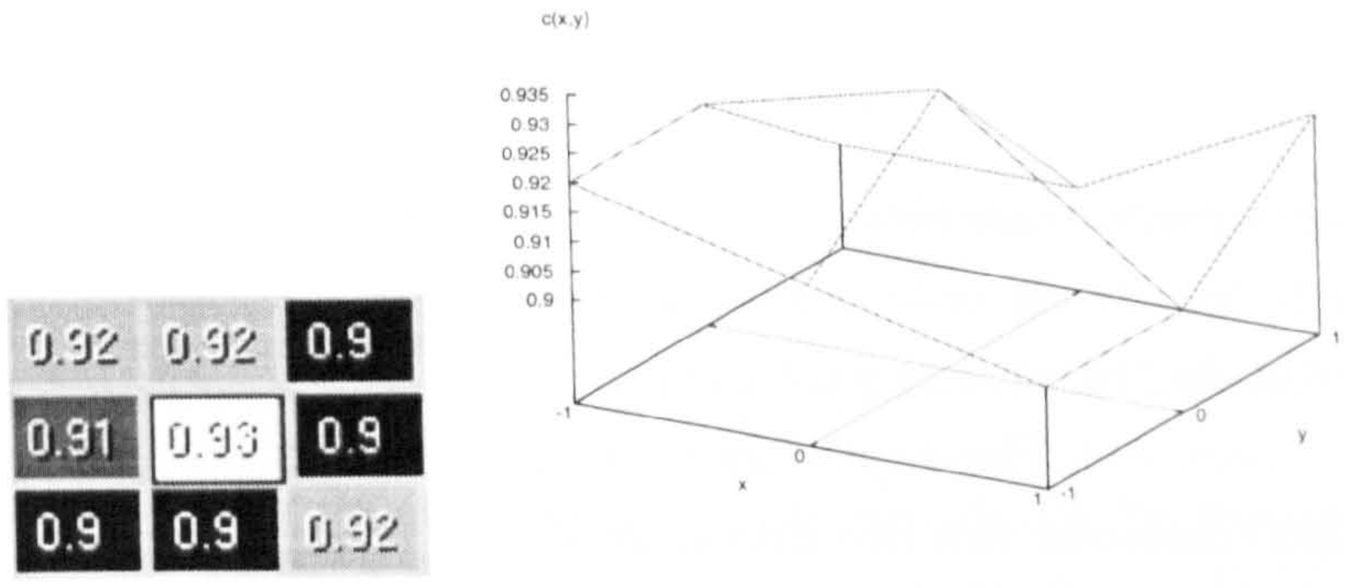
6.5.3.2 Sub-pixel Correlation

An alternative method proposed by Lan and Mohr (1998) is based on a linear sub-pixel correlation. The idea behind the method is that a translation of some signal can be approximated by convolution and the estimation of the convolution mask then provides a sub-pixel translation estimate.

A translation of a signal $f(x)$ by t can be expressed in terms of convolution as

$$f(x - t) = f(x) \star \delta(x - t) = \int_u f(u)\delta(x - t - u)du \tag{6.24}$$

where $\delta(x)$ is a Dirac pulse function. When converted to a discrete domain $(x \rightarrow i)$, Dirac becomes Delta-Kronecker (unit sample) and the integral changes to the sum



(a) correlation values

(b) correlation plane

Figure 6.8: An illustration of unstable sub-pixel location when paraboloid interpolation method is used. Clearly, the maximum is in the middle of the patch. However, the fitted paraboloid has its centre at $(-4, 6.5)$, which is incorrect.

$$f(i-t) = \sum_j f(j)\delta(i-t-j) \tag{6.25}$$

If t is non-integer, there is no exact answer, as the values of f are missing between the samples. However, by assuming Shannon's sampling theorem, a piecewise interpolation is adequate

$$f(i-\epsilon) \approx (1-\epsilon)f(i) + \epsilon f(i-1);\ 0 \leq \epsilon \leq 1 \tag{6.26}$$

This corresponds to convolution with $(1-\epsilon)\delta(x) + \epsilon\delta(x-1)$, which is an approximation of $\delta(x-\epsilon)$. More generally, given $b_i, \lambda_i;\ b_i > 0\ (\forall i)$, such that $\sum_i b_i = 1$, then

$$\sum b_i f(x-\lambda_i) \approx f(x - \sum \lambda_i b_i) \tag{6.27}$$

In other words $\sum_i b_i\delta(x-\lambda_i)$ is an approximation of $\delta(x - \sum_i \lambda_i b_i)$ for a sufficiently smooth function.

Lan defines the matching problem as:

> given f_1 and f_2, find $t \in$ R such that $F_1(x) \sim F_2(x-t) = F_2(x) \star \delta(x-t)$. $F_1$and F_2 are sampled f_1 and f_2. The $\sim$ means 'equal up

to a signal transformation' - for instance, an offset and a scaling in the values of the F_i - and noise, i.e.

$$F_1(x) = SF(x-t) + O + \epsilon \tag{6.28}$$

In case of precise matching, f_1 and f_2 are known and $t \in [-1;1]$ is to be found. Since f_1 and f_2 are only defined at discrete integer values, $\delta(x-t)$ can be approximated by a linear combination $b_{-1}\delta(x-1) + b_0\delta(x) + b_1\delta(x+1)$ where $b_{-1} + b_0 + b_1 = 1$. The displacement can then be estimated as $b_1 - b_{-1}$.

When extended to a two-dimensional case, Equation 6.28 can be rewritten as

$$I_1(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbf{N}} a_{\mathbf{x}} I_2(\mathbf{x}+\mathbf{k}) + O + \varepsilon(\mathbf{x}) \tag{6.29}$$

where $I_1(\mathbf{x})$ and $I_2(\mathbf{x})$ are the two image patches to be exactly matched, $\mathbf{x}$ runs through a chosen window $\mathbf{W}$, $\mathbf{x} \in \mathbf{W} = \{(x,y) \mid s - s_{len} < x < s + s_{len}, t - t_{len} < y < t + t_{len}\}$, $\mathbf{k}$ runs through the neighbourhood $\mathbf{N}$ of $\mathbf{x}$, $\mathbf{k} \in \{(k_x,k_y) \mid -1 \le k_x \le 1, -1 \le k_y \le 1\}$. The neighbourhood $\mathbf{N}$ can either be a 4-connected or an 8-connected neighbourhood. The results in (Lan and Mohr, 1998) show that a 4-connected neighbourhood is sufficient for most cases. While $a_{\mathbf{k}}$ encodes the shift and scaling, O represents possible offset and $\varepsilon(\mathbf{x})$ is uncorrelated white noise. Given two patches in matching windows, I_1 and I_2, $a_{\mathbf{k}}$ (for $\mathbf{k} \in \mathbf{N}$) and O can be estimated from linear least-squares minimisation

$$\min_{a_{\mathbf{k}},O} \sum_{\mathbf{x}\in\mathbf{W}} (I_1(\mathbf{x}) - (\sum_{\mathbf{k}\in\mathbf{N}} a_{\mathbf{k}} I_2(\mathbf{x}+\mathbf{k}) + O))^2 \tag{6.30}$$

The estimated standard deviation of $\varepsilon(\mathbf{x})$ directly follows from Equation 6.29

$$\sigma_\varepsilon = \sqrt{\sum_{\mathbf{x}\in\mathbf{W}} \frac{(I_1(\mathbf{x}) - \sum_{\mathbf{k}\in\mathbf{N}} a_{\mathbf{k}} I_2(\mathbf{x}+\mathbf{k}) - O)^2}{(2s_{len}+1)(2t_{len}+1)}} \tag{6.31}$$

The value σ_ε is used for estimating the uncertainty of displacements.

The problem is symmetric, therefore another displacement estimate can be obtained,

$$I_2(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbf{N}} a'_{\mathbf{k}} I_1(\mathbf{x}+\mathbf{k}) + O' + \varepsilon'(\mathbf{x}) \tag{6.32}$$

The sub-pixel displacements can be determined afterwards from Equations 6.29 and 6.32 as

$$d\mathbf{x}_1 = \sum_{\mathbf{k}\in\mathbf{N}} b_{\mathbf{k}}\mathbf{k} \tag{6.33}$$

and

$$d\mathbf{x}_2 = \sum_{\mathbf{k}\in\mathbf{N}} b'_{\mathbf{k}}\mathbf{k} \tag{6.34}$$

An uncertainty of the estimated displacements can be used to obtain an improved result. If the covariance matrices of both displacements are known as $\mathbf{CV}_1$ and $\mathbf{CV}_2$ then the resulting displacement is given as

$$d\mathbf{x} = \left(\mathbf{CV}_1^{-1} + \mathbf{CV}_2^{-1}\right)^{-1}\left(\mathbf{CV}_1^{-1}d\mathbf{x}_1 + \mathbf{CV}_2^{-1}d\mathbf{x}_2\right) \tag{6.35}$$

This fusion is optimal if the error distribution is Gaussian and uncorrelated (Thacker and Cootes, 1996).

To obtain the solution of Equation 6.30 a matrix notation is used. First of all, Equation 6.29 is written as

$$\mathbf{I}_1 = \mathbf{I}_2\mathbf{T} + \varepsilon(\mathbf{x}) \tag{6.36}$$

where $\mathbf{I}_1$is a column vector with elements $I_1(\mathbf{x}), \mathbf{x} \in \mathbf{W}$, $\mathbf{I}_2$ is a matrix with row vectors $(\{I_2(\mathbf{x}+\mathbf{k}),\ \mathbf{k} \in \mathbf{N}\}, 1)$ and $\mathbf{T} = (\{a_{\mathbf{k}}, \mathbf{k} \in \mathbf{N}\}, 0)^T$. When denoting $\mathbf{L} = (\mathbf{I}_2^T\mathbf{I}_2)^{-1}\mathbf{I}_2^T$, the least-squares solution is obtained

$$\mathbf{T} = \mathbf{L}\mathbf{I}_1 \tag{6.37}$$

Denoting

$$d\mathbf{x} = (dx, dy)$$

$$\mathbf{CV} = \begin{pmatrix} var(dx) & cov(dx,dy) \\ cov(dx,dy) & var(dy) \end{pmatrix}$$

$$\mathbf{L} = \{\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_{n+1}\}^T$$

where n is the number of pixels in neighbourhood $\mathbf{N}$, the displacement and its' covariance matrix for a 4-connected neighbourhood is obtained as

$$(dx, dy) = \left(\frac{a_{(1,0)} - a_{(-1,0)}}{S}, \frac{a_{(0,1)} - a_{(0,-1)}}{S}\right) = \frac{1}{S}((\mathbf{L}_5 - \mathbf{L}_1)\mathbf{I}_1, (\mathbf{L}_4 - \mathbf{L}_2)\mathbf{I}_1) \tag{6.38}$$

where $S = \sum_{\mathbf{k}\in\mathbf{N}} a_{\mathbf{k}}$,

$$var(dx) = \frac{\sigma_\epsilon^2}{S^2}(\mathbf{L}_5 - \mathbf{L}_1)(\mathbf{L}_5 - \mathbf{L}_1)^T \tag{6.39}$$

$$var(dy) = \frac{\sigma_\epsilon^2}{S^2}(\mathbf{L}_4 - \mathbf{L}_2)(\mathbf{L}_4 - \mathbf{L}_2)^T \tag{6.40}$$

$$cov(dx, dy) = \frac{\sigma_\epsilon^2}{S^2}(\mathbf{L}_4 - \mathbf{L}_2)(\mathbf{L}_5 - \mathbf{L}_1)^T \tag{6.41}$$

Because of the symmetry, two displacements with their covariances can be obtained and fused by the means of Equation 6.35.

Lan and Mohr (1998) provides a modified version (called 'robust' algorithm) of the above method (called 'fast' algorithm) that also considers a local affine transformation of the image patches. The method is computationally more expensive and prior knowledge of a dense map is necessary. For minor inter-frame displacements that commonly occur in wide range rigid object tracking the 'fast' method suffices as the optional inter-frame distortions are negligible.

The described method is more robust than paraboloid fitting, it operates directly on the image data so it can be used with any affinity measure. The method also quantifies the uncertainty of the matching that can be employed as a confidence measure in consequent processing steps. The size of the window W is set to 11 pixels in both directions as recommended by Lan and Mohr (1998).

6.6 Spatio-temporal Correspondence Database

The correspondence matching algorithm applied on data coming from the previous modules of the framework determines the temporal correspondence of segments and corners detected in the scene. Every segment detected during the primary segmentation described in Chapter 4 is assigned a segment record that contains all information essential for successful identification and tracking of an object in that segment. Each segment record holds the following information:

- frame number when the segment first appeared in the scene
- frame number when it vanished

- a list of all matched segments with coordinates
- a list of all trace records for traces that were generated when matching corners within the segments

Each trace is described by a trace record similar to a segment record. The trace record holds the following information:

- frame number when the trace was spanned
- frame number when the trace was terminated
- a list of all corners that belong to the trace

Finally, each corner is described by a corner record that contains the following information:

- number of frame in which the corner was detected and assigned to the trace
- detected location of the corner in the current frame
- matched sub-pixel location of the corner in the previous frame
- matched sub-pixel location of the corner in the next frame
- a flag indicating if the corner was detected, matched or predicted

A structure of the segment record is outlined in Figure 6.9. The database of all segment records is maintained through the sequence. To avoid uncontrolled growth of the database size, the segments that vanished before a given number of frames are disposed of. The number of frames for which the unmatched segments are kept is same as the depth of the temporal segment stack described in Section 6.4.2.

The information stored in the database is employed in the following modules of the tracking system for estimation of the motion parameters and for remapping of the image positions into the scene coordinates.

6.7 Structure of Matching Module

The spatio-temporal matching of the geometric feature sets generated by the previous module of the framework consists of four consecutive steps. These

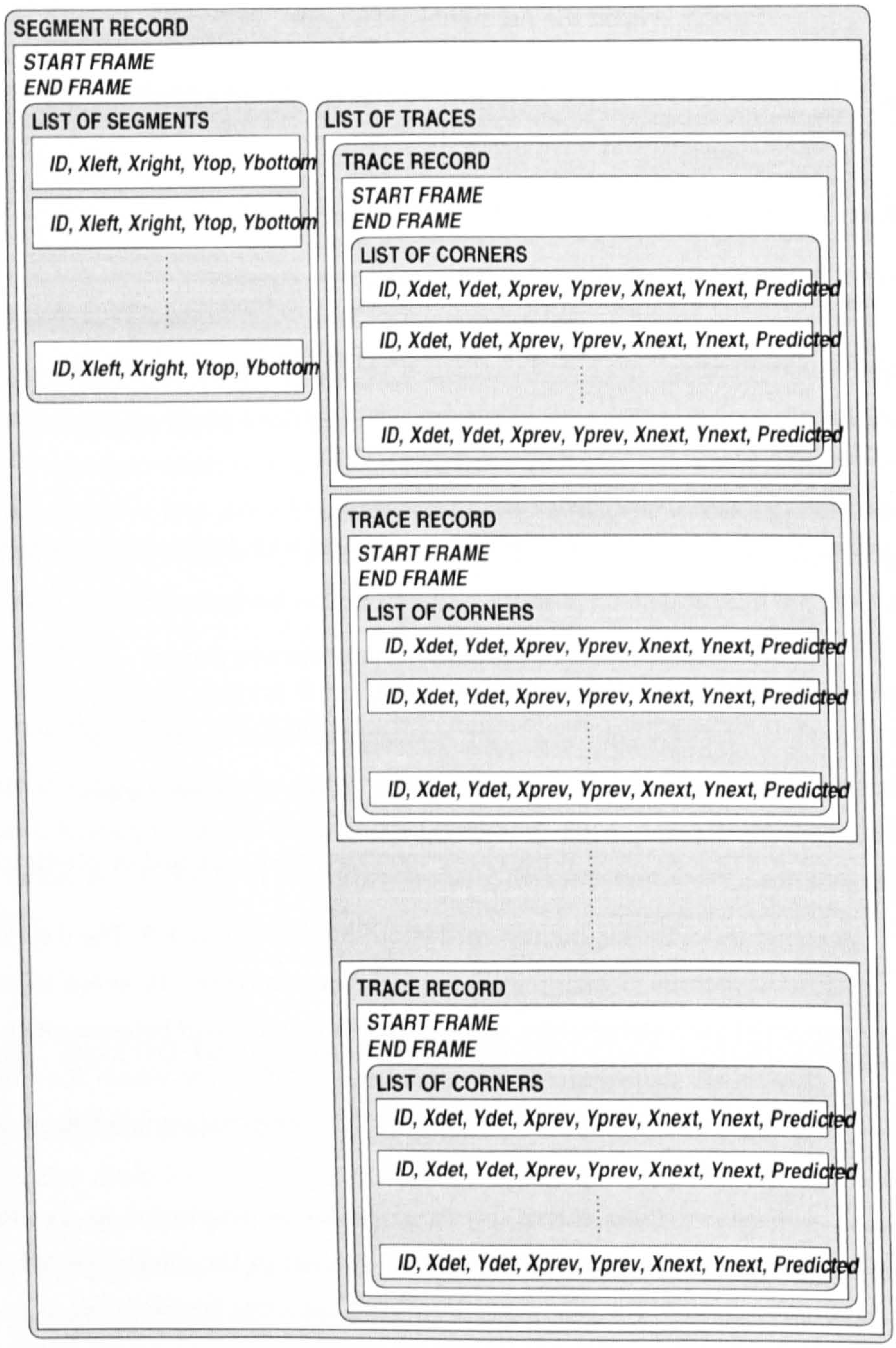


Figure 6.9: Structure of the segment record used for maintaining the spatio-temporal information about the objects in the scene.

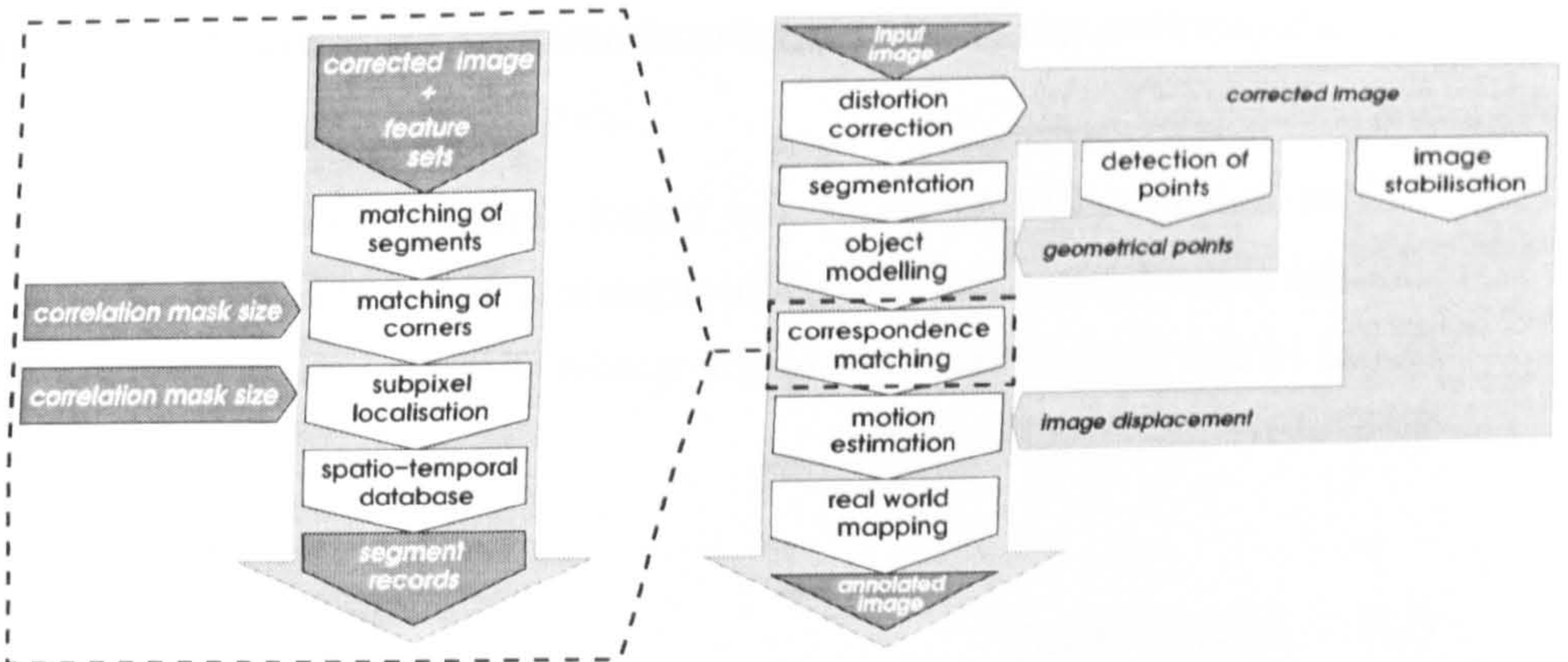


Figure 6.10: The structure of feature matching module.

steps are assembled into the feature matching module. The structure of the module is outlined in Figure 6.10.

The feature matching module takes as inputs the current and previous frames in the sequence together with all feature sets detected in those frames. Following processing steps are applied to the data:

- *Matching of segments*. Primary correspondences between the segments in two consequent frames delimiting the locations of objects are established. The correspondences are based on the area of segment overlap between two consequent frames.

- *Matching of corners*. The matching between two sets of corners belonging to the matched segments is found. The correspondence search employs the correlation between two image patches surrounding the corner candidates. The selected correlation measure is based on the Sum of Squared Differences. A linear prediction scheme maintains the consistency of the matching over multiple frames in case a dropout in the corner detection occurs.

- *Subpixel localisation*. Displacements of the corresponding pixel pairs between the frames in the sequence are refined to sub-pixel levels in order to improve the precision of the consequent motion estimation.

- *Spatio-temporal database*. The history of the presence and motion of objects in the scene is maintained in the spatio-temporal database. The objects are represented by segment records containing the enclosing segments and lists of corners within these segments matched across the sequence.

The motion of objects is implied from the displacements of the matched corners.

The motion parameters of the objects are estimated in the consequent module of the framework. The estimation is based on the corner displacements stored in the segment records of the spatio-temporal database maintained by the feature matching module.

6.8 Summary

An inter-frame correspondence between the geometric points detected at previous stages of processing is established. The correspondence search is an initial step in motion estimation. The correspondence search method is based on a local image registration of intensity patches centred at positions of corners detected in the frames. The Sum of Squared Differences serves as the best performing affinity measure between candidates for correspondence matching, as indicated by evaluations on artificial and real scenes. The Sum of Squared Differences combined with the median fusion of the values detects the displacements in the evaluation sequences with errors below 0.5 pixel.

A two way matching is employed in order to resolve ambiguities in the matching process. The candidates are matched according to their mutual affinity by a modified X-dominant matching algorithm proposed by Sara (1999). For corners that are left unmatched in the recent frame a linear prediction estimates their future position. The linear prediction improves the long-term coherence of the matches by 12% in artificial and by 5% real evaluation scenes. The localisation of the matches is improved by a sub-pixel matching scheme devised by Lan and Mohr (1998). The matched corners are assigned to traces. Traces encode the movement of each corner in time.

Similar to corners, the segments are matched using a modified X-dominant matching algorithm. The affinity measure is based solely on the amount of the overlap of detected segments. To deal with occasional dropouts in the initial segmentation, a temporal stack of all segments detected in a number of previous frames is maintained and used in matching.

Finally, the corresponding segments and corners are stored in a spatio-temporal database with a hierarchical structure - each segment is assigned a record with references to corresponding traces and each trace contains a list of all assigned corners. The information in the database is passed to the

consequent processing modules that estimate the motion of objects and remap the results to scene coordinates.

Chapter 7

Motion Estimation and Tracking

7.1 Introduction

The ultimate goal of a tracking system is to detect, characterise and possibly classify any activity in the scene being surveyed. While for some applications the fact that 'something is moving' might be the required outcome of the tracking process, more specific characterisation of the detected activity is usually desired. Some applications can use the characteristics of the detected activity to distinguish among different types of objects such as cars versus people, (Lipton, 1999). Other applications are not engaged as much in object classification based on the activity characteristics but they attempt to achieve the best possible precision in the estimation of motion parameters, regardless of the type of object, (Dellaert and Thorpe, 1997). For example, in collision avoidance application the type of object that is on the collision course is not as important as its velocity and direction.

In considering the intended application of a maritime tracker, object recognition based on characteristics of detected motion is not the main goal. Furthermore, the varying appearance of maritime objects would make any inference of the type of object somewhat vague. The maritime object classes outlined in Section 2.4.2 are explicitly based on motion characteristics without any direct or unambiguous inference of their type. A floating object can be a

buoy, a wooden log, a mooring vessel, a drowning person, etc. A completely static object can be a pier, a large mooring vessel, a rock, etc.

The aim of the maritime tracking system falls into the second category - to provide estimates of any motion as precise as possible, regardless of the structural characteristics of the objects undergoing the motion. The outcome depends on the precision, completeness and consistency of the data. Although the method used in the correspondence search module of the framework reduces errors by fusing multiple displacements in the segment using median, there remains uncertainty in the obtained values. This uncertainty is due to various effects that cannot be compensated for such as, for example, horizontal fluctuation of the image caused by environmental conditions.

An important issue in outdoor tracking applications is the compensation for errors in localisations of detected features in the image that occur due to uncontrolled oscillations of the imaging device platform. These displacements occur inevitably due to the cross-wind impact on an imaging device harness. The mechanical noise caused by a vessel's engines acts as another source of systematic localisation errors.

A method that compensates for inter-frame global displacements of the image is therefore proposed. The method estimates the displacement of the horizon that represents a strong horizontal feature feasible for tracking. The results are used to compensate for the localisation errors by relating positions of all detected geometric features to the tracked position of the horizon.

As the problem of noisy input data is common in many engineering applications and, especially, in navigation, robust methods to estimate system states treating the data at a stochastic level are typically used. The Kalman filter (Welch and Bishop, 2001) and its extended and modified versions as in (Li et al., 2004) are essential parts of various systems that operate with noisy input data. As the motion of the objects in maritime scenes is generally smooth with only gradually changing parameters, a standard linear version of the Kalman filter is appropriate for tracking purposes.

7.2 Motion Model

Once the corresponding features between consequent frames are established using local registration techniques any motion that might occur can be detected using these spatio-temporal correspondences. The selection of appropriate

motion models is complicated, especially when multiple complex motions are involved. Shapiro (1995) assumes affine motion of rigid objects and he provides a structure from motion analysis framework based on corner tracking. Torr (1998) assumes multiple general transformation models between images including projective one and provides selection criteria based on the maximum likelihood estimation of the parameters of the fitted model.

Both frameworks resolve the problem of motion segmentation by fitting various motion models onto the matched feature pairs without any prior assumption about their correspondence to the structure of the scene. The approach is commonly known as a 'structure from motion', where objects in the scene can be detected and modelled from the motion of the matched features. If the locations of the objects in the scene are known and their corresponding features used in inter-frame matching are detected, the motion estimation is significantly simplified, as the structure of the scene is inferred from the segmentation and does not have to be deduced from motion.

The geometric context of maritime scenes provides constraints essential for the motion estimation. First of all, due to the fact that all objects lie on a horizontal plane the motion is restricted to two dimensions in the scene neglecting vertical displacements due to waves or other natural effects. All objects are considered rigid, which means that there is no motion due to deformation of the objects' structure. The only significant rotation that can occur is parallel to the sea plane, i.e. boat is turning, rotations in other directions are either negligible such as those due to waves or highly unlikely, i.e. boat is turning upside down. Due to the nature of the maritime environment rotations of objects are typically slower than their translations. The rotations can be approximated by many small translations between consequent frames. Thus, it is possible to assume that the majority of detected motion in maritime scenes is translational.

The ego-motion of the camera has an essential effect on estimation of motion of independent objects in many applications (Irani et al., 1994; Cohen and Medioni, 1998). It is necessary to determine and compensate for the ego-motion of the vision system first in order to estimate independent motions of the objects in the scene. If the camera is mounted on a vessel the ego-motion will be mainly translational. Even if the vessel rotates around its axis the rotation will cause a panning effect which is mainly translational for large depths of the scene. A motion relative to the observation point is essential

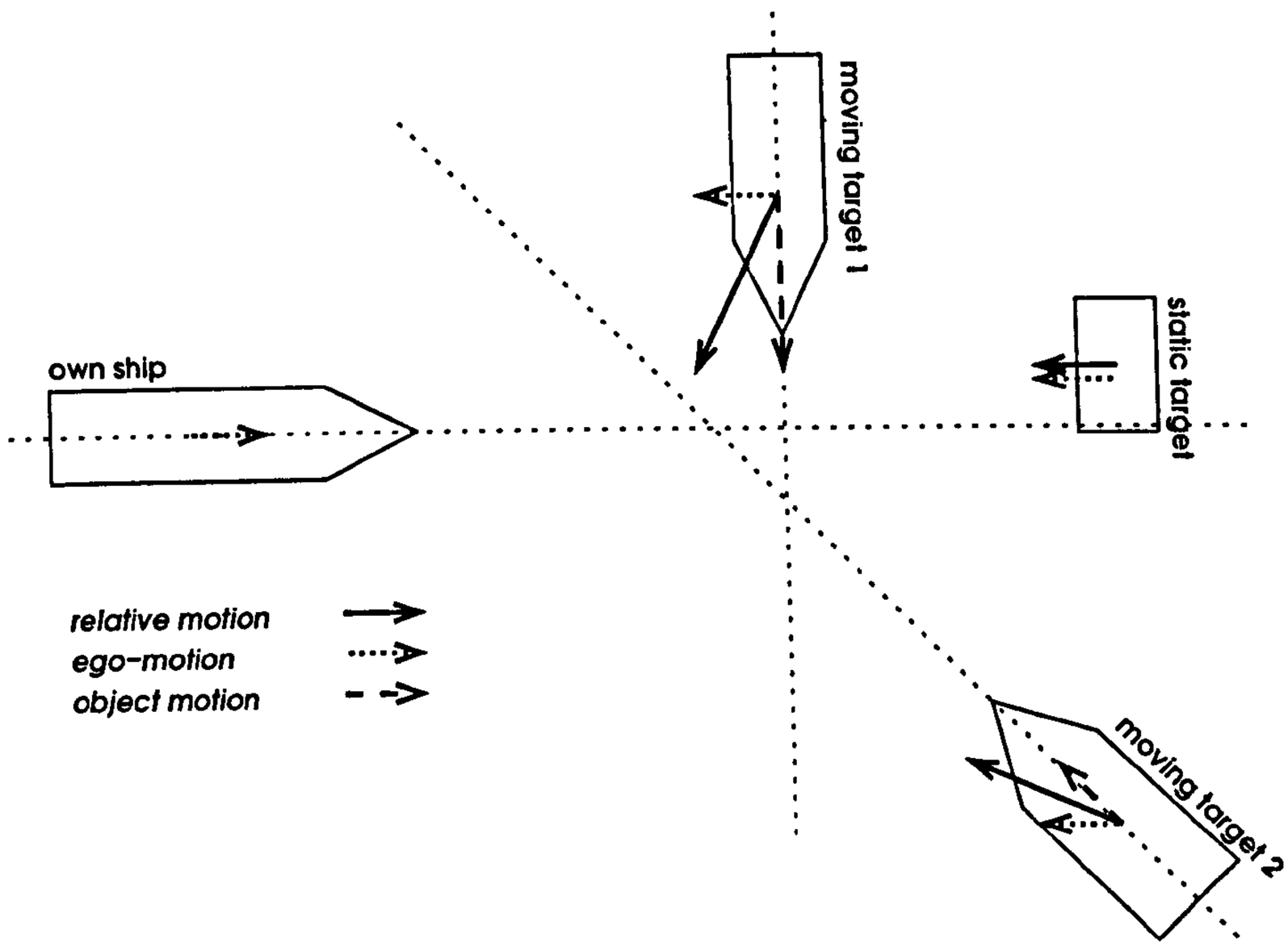


Figure 7.1: Relative motions of the objects in the scene are essential for collision threat assessment. Ego-motion is considered as a component of these relative motions.

for threat assessment applications (see Figure 7.1). The relative motion is a combination of the translational ego-motion and object's motion components. Therefore, the ego-motion component can be considered as a component of independent motions of objects in the scene.

Finally, with the exception of fast moving and highly maneuverable small craft the motion of most maritime objects is uniform or changing gradually. A combination of an adequate value of the proximity constraint introduced in Section 2.4.3 and sufficiently high frame rate of the imaging device allows to capture these gradual changes making it possible to track them using the linear motion model.

7.3 Feature Displacements

Each trace stored in a spatio-temporal database contains a set of corresponding corners detected in consequent frames. The differences in the positions of corresponding corners indicate the displacements of these features between the frames. When divided by the frame duration reciprocal to the frame rate

the actual velocity of each feature is obtained.

To improve on the precision of the displacement a two-way sub-pixel matching scheme is proposed as shown in Figure 7.2. A corner detected in the current frame at position (x_{curr}, y_{curr}) is matched to sub pixel positions in both previous $(x_{prev} + dx_{prev}, y_{prev} + dy_{prev})$ and next $(x_{next} + dx_{next}, y_{next} + dy_{next})$ frames. The difference of these two positions divided by the number of frames N_f across which the correspondence was found determines the average displacement of the feature across a single frame to a sub-pixel precision

$$\Delta x = \frac{x_{next} + dx_{next} - x_{prev} - dx_{prev}}{N_f} \tag{7.1}$$

$$\Delta y = \frac{y_{next} + dy_{next} - y_{prev} - dy_{prev}}{N_f} \tag{7.2}$$

This scheme enables displacements to be determined even when the features were matched across more than two frames ($N_f = 2$). Such a situation occurs when the current segment is matched against a segment deeper in the temporal stack (see Figure 6.5b). In such a case the number of frames $N_f > 2$. The advantage of the two-way matching scheme is that any offset of a corner during the detection in the current frame does not introduce an error into the displacement estimation. The detection error will just cause the centre of the matching patch to be shifted but the average inter-frame displacement of the whole patch is still recovered correctly.

Smith et al. (1998) shows that median provides an adequate estimation of a global inter-frame displacement in the scene from displacements between corresponding features. Similarly to Smith, median value of all displacements in a single segment represents the fused displacement of an object in the segment. A variance of the displacements serves as a confidence measure of the estimation.

Finally, the Kalman filter is associated with each segment. The fused displacement, its variance and position of submersion line represent input data measurements in the process of motion estimation.

7.4 Horizon Tracking

Any machine vision system for outdoor applications faces an ultimate challenge of weather conditions. Most prominent among these is the cross-wind impact on the imaging device platform. Even though the imaging devices used

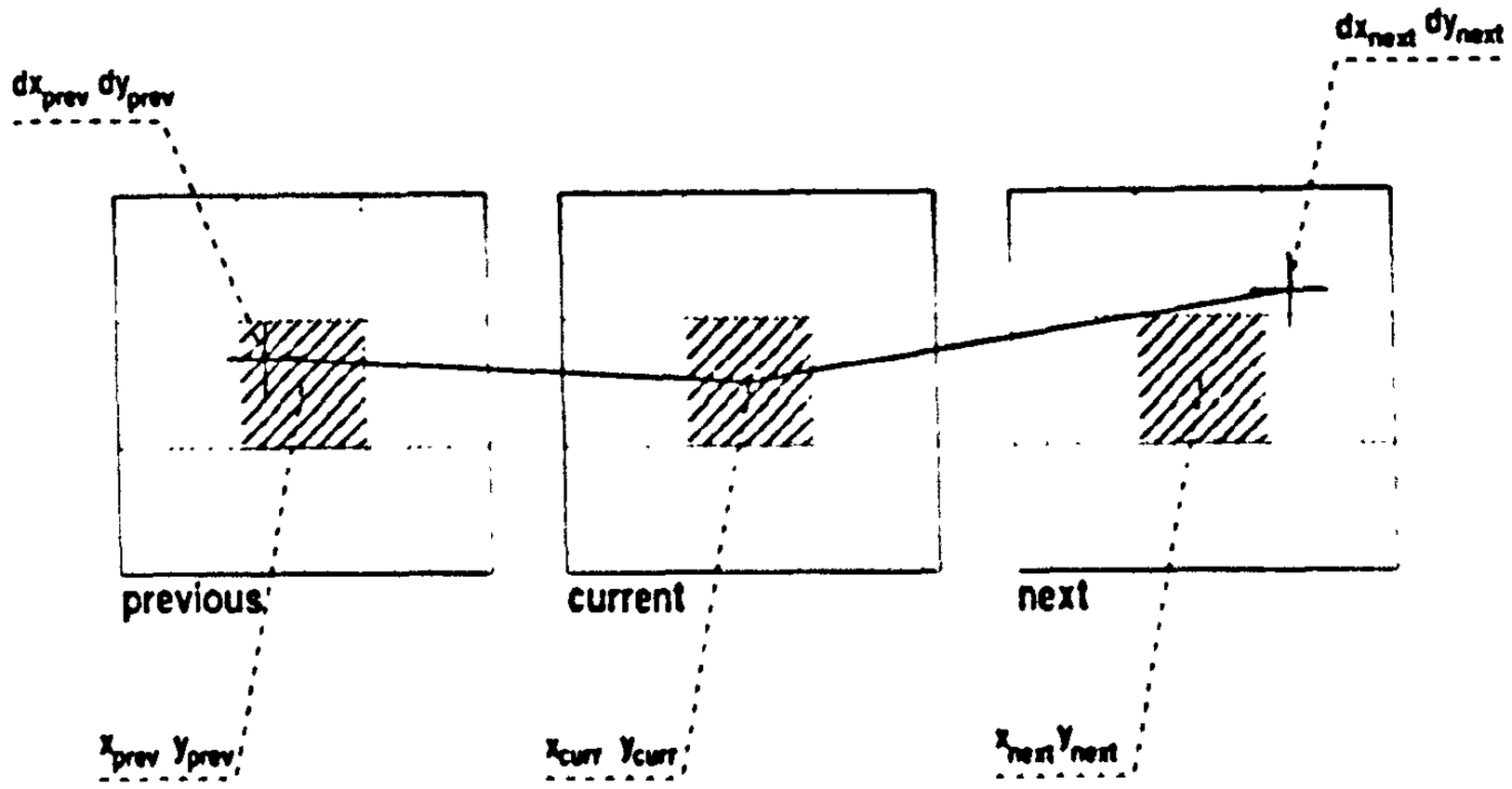


Figure 7.2: Two-way sub-pixel matching of corners: the corner detected to a pixel precision (x_{curr}, y_{curr}) in the current frame is located to a sub-pixel precision in the previous $(x_{prev} + dx_{prev}, y_{prev} + dy_{prev})$ and next $(x_{next} + dx_{next}, y_{next} + dy_{next})$ frames. The difference between the coordinates in previous and next frames provides a sub-pixel displacement between these frames. Matching across more than one frame is possible if the corresponding segment was matched against segment deeper in the segment temporal stack (see Figure 6.5).

in the maritime environment are stabilised for oscillations due to vessel-wave interaction (Vistar Night Vision Limited, 2004a; Vistar Night Vision Limited, 2004b; Vistar Night Vision Limited, 2004c), fluctuations caused by cross-wind impact are too rapid and brief to be captured by an electro-mechanical control system based on a gyroscope and feedback loop. Another source of displacement can originate from vibrations of the vessel itself caused by running engines. These factors cause small movements of the imaging device that result in displacements of the captured frames. When projected onto an image, the fluctuations usually amount to a couple of pixels over variable time spans and are often completely random. Dellaert and Thorpe (1997) model these displacements as first-order Markov process or as time-correlated noise.

A systematic error is introduced into the location of all detected features in the image by these fluctuations. Because the displacement is similar for all features the displacement variance does not increase and only the fused displacement is shifted. The error propagates through the matching and tracking to the remapping where it can cause significant deviations in the scene location and velocity estimates due to a non-linearity of projective mapping. The situation is outlined in principle in Figure 7.3. The original feature position

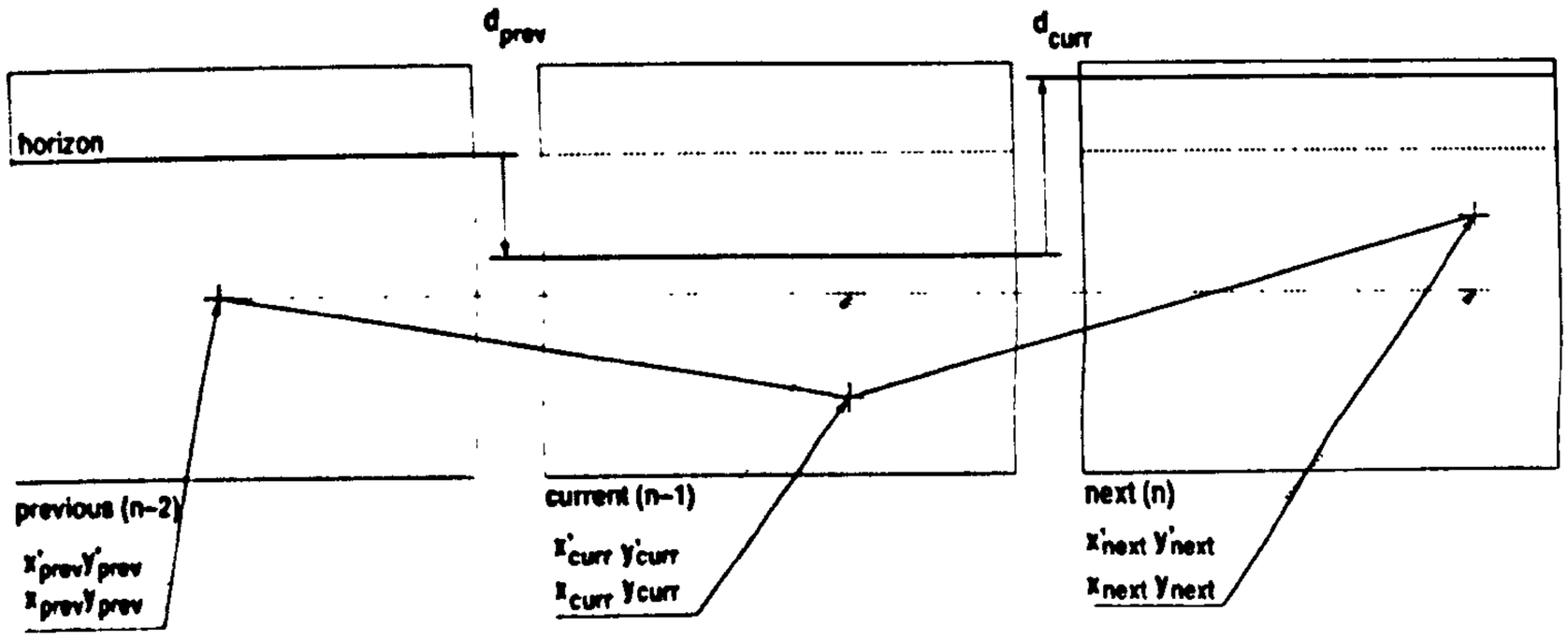


Figure 7.3: The effect of horizon displacement on two-way corner matching. The horizon moved between previous and current frames by amount of d_{prev} and between current and next frames by d_{curr} pixels (the direction of the displacement is given by signs of the values). Ideal positions of the features in consequent frames are indicated as primed symbols. The systematic error in displacement $d_{prev} + d_{curr}$ is same for all detected features.

(x'_{prev}, y'_{prev}) changes with respect to image coordinates even though it remains fixed with respect to the horizon. If the horizon displacement is not accounted for the resulting vertical displacement would not be zero indicating vertical displacement of the feature which is clearly not the case.

Image to scene mapping is non-linear (perspective projection) and the same image displacements are mapped to larger values near horizon as those near the image bottom. This leads to an error in motion estimation that progressively increases towards the horizon. It is necessary to compensate for such systematic errors, namely the offsets caused by vibrations of the imaging device in order to improve the estimations.

The compensation methods for imaging device vibrations are commonly referred to as image stabilisation techniques. Multiple approaches to image stabilisation are used in the photography and imaging domains. One approach called 'optical image stabilisation' (Canon, 2004) is based on complex electro-mechanical devices that compensate for the vibrations by changing the optical properties of the imaging devices. The stabilisation is done prior to the image being captured. This type of stabilisation is mostly available in high-end photography only. The second approach is called 'digital image stabilisation', (Morimoto and Chellappa, 1996; Ko et al., 1999; Erturk, 2003). It is typically used in low-end consumer digital video cameras or in machine vision applications. The stabilisation is done by a registration on a frame to frame basis of some strong directional features detected in the scene.

A method for digital image stabilisation is introduced into the framework that compensates for vertical displacements of the image. A strong feature that is simple to track and can always be present in the scene is the horizon, e.g. dividing line between the water plane and either the shore or sky. A horizon tracking algorithm is applied to the sequence that evaluates the inter-frame vertical displacements caused by the camera vibrations. The displacement in the horizontal direction is not compensated for as a feature suitable for tracking the horizontal motion is not always available.

The initial position of the horizon h in the scene together with a fluctuation ranges Δh_n and Δh_{n-1} are input into the tracking algorithm. A strip surrounding the horizon line that stretches along the image width is used to determine the displacements by a correlation technique. The strip in the current image is matched against the same strip in the previous image. Because only the vertical displacement is of interest, the correlation is done for vertical displacements only. Strips can be expressed as matrices

$$\mathbf{I}_{n,n-1} = \{I_{n,n-1}(r,c);\ h - \Delta h_{n,n-1} \leq r \leq h + \Delta h_{n,n-1};\ 0 < c < C\} \quad (7.3)$$

where $I_{n,n-1}$ are current and previous images that have equal width C and $\Delta h_n > \Delta h_{n-1}$ when matching horizon in frame $n-1$ against the one in frame n (see Figure 7.4). The matching is done by finding the Sum of Squared Differences defined in Equation 6.3 between corresponding columns of the matrices for different shifts d_c

$$M_{SSD}(i,c) = \sum_{r=h-\Delta h_{n-1}}^{h+\Delta h_{n-1}} (\mathbf{I}_n(r+i,c) - \mathbf{I}_{n-1}(r,c))^2 \quad (7.4)$$

where c is the actual column in matrices $\mathbf{I}_{n,n-1}$ and $i = 0,\ldots,\Delta h_n - \Delta h_{n-1} + 1$. The offset is detected at the minimum of $M_{SSD}(d_c,c)$ profile

$$d_c = \arg \min_{i=0,\ldots,\Delta h_n - \Delta h_{n-1}+1} (M_{SSD}(i,c)) \quad (7.5)$$

A sub-pixel refinement to improve the displacement estimate of the offset is done for every column. A same technique by Lan and Mohr (1998) used in corner matching in Section 6.5.3 is applied in a single dimension to every column in the matching strip. The one-dimensional matching follows the same principles as the two-dimensional one. In fact, two-dimensional sub-pixel matching is an extension of the one-dimensional one, (Lan and Mohr, 1998).

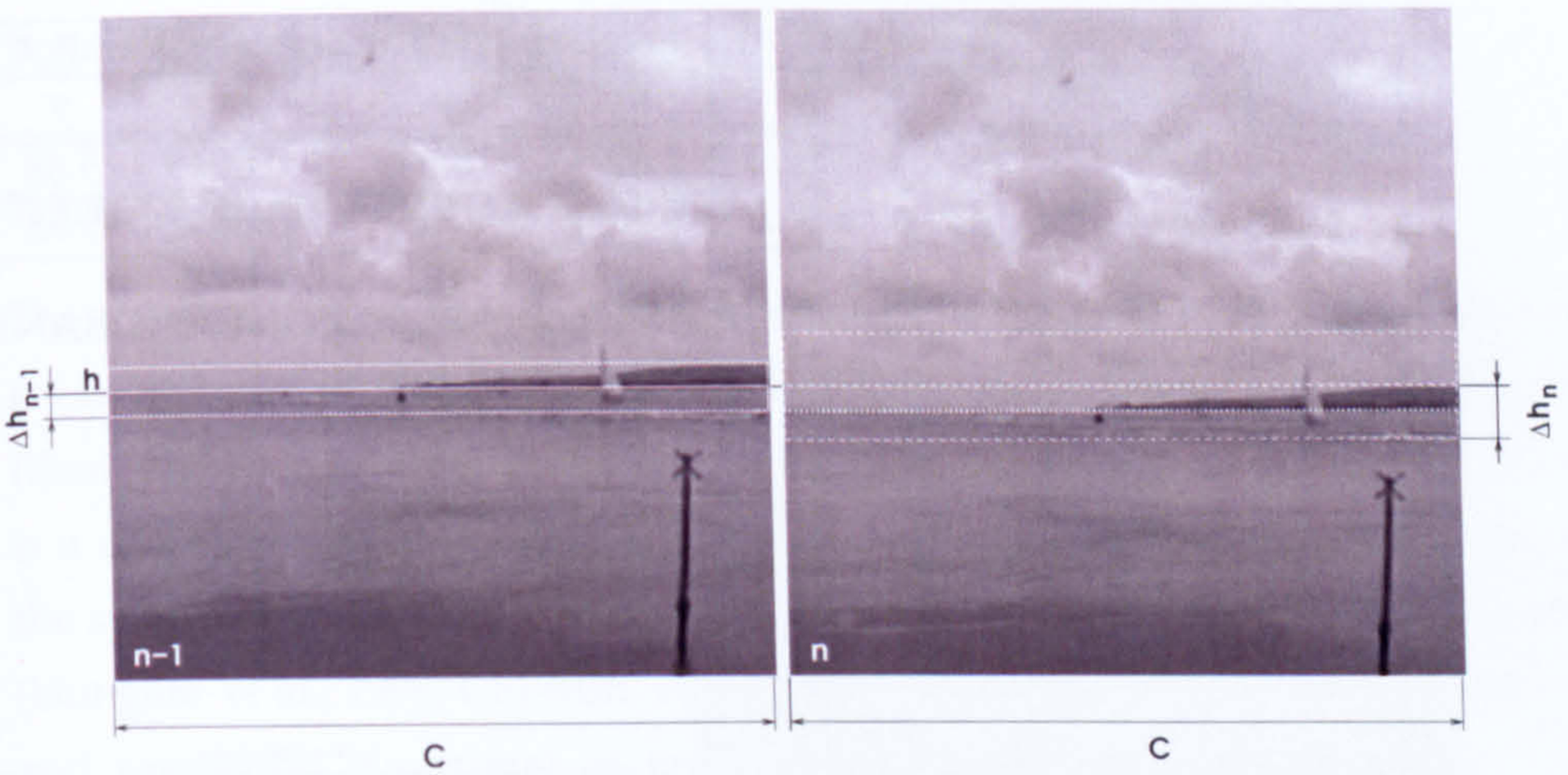


Figure 7.4: Detection of horizon displacement between the previous frame $n-1$ and the current frame n. The detection is based on vertical correlation of the strip in frame $n-1$ with the strip in in frame n.

The resulting overall vertical shift d_h is delimited as a median value of all column offsets d_c; $c = 1, \ldots, C$. This avoids a contribution of any outliers in offsets caused by noise or change in the scene structure.

Two-dimensional matching is effectively divided into C one-dimensional matching procedures. Such division has numerous advantages. First of all, a parallel implementation of the process is possible. Secondly, by performing C independent matches and fusing the results through a median any likely inconsistency in individual correlations does not influence the result.

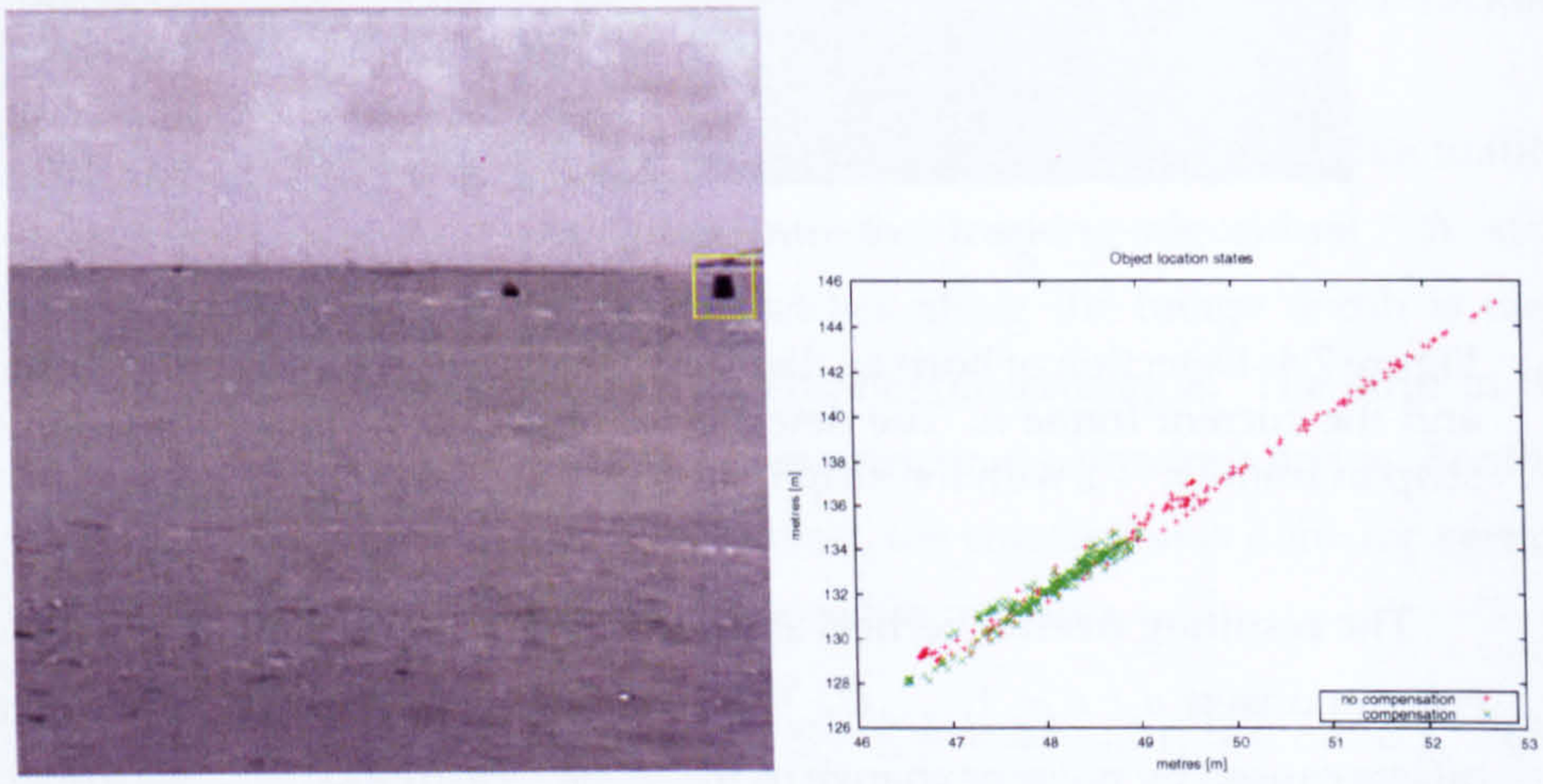
The stabilisation is done by relating locations of all detected submersion lines and corners to the position of horizon projection rather than to the image boundaries. Any changes to the locations of features are then likely due to the motion in the scene rather than the displacement of the camera.

An example of the compensation for horizon oscillations applied to a sample sequence is shown in Figure 7.5. A SANDBANKS2Q sequence shows a channel marking buoy approximately 150 metres away from the camera. Even though the sequence has been taken on a relatively calm day small vibrations of the camera can be registered throughout the sequence. Estimations of location of the buoy for the first 200 frames of the sequence have been determined with and without the horizon tracking.

The results summarised in Table 7.1 show the average location, it's variance and estimate of the state variance obtained by Kalman tracking. The example clearly indicates the benefits of the horizon tracking in increased estimation

	avg. location		location variance		state variance	
compensation	x	y	x	y	x	y
no	49.1	135.2	2.62	18.63	0.0045	5.88
yes	48	132.2	0.35	2.49	0.0042	5.53

Table 7.1: Evaluation of horizon oscillation compensation in SANDBANKS2Q sequence for the first 200 frames.The values are in pixels.



(a) frame 114 from SANDBANKS2Q sequence with channel marking buoy highlighted

(b) position states

Figure 7.5: Compensation for the horizon oscillation. Channel marking buoy (a) highlighted in the sample frame is tracked over the first 200 frames of the SANDBANKS2Q sequence with and without compensation for a horizon shift. The resulting estimated locations (b) for both cases clearly indicate the benefits of the compensation.

precision. The similar values of the state variances after Kalman tracking in both cases indicate that the systematic error due to the horizon shift is indeed undetectable. The presence of the error is indicated by the high value of the location variance when the shift is uncompensated.

7.5 Kalman Tracking

7.5.1 Linear Kalman Tracker

The translational motion is characteristic for most objects in maritime scenes. It can be estimated as a linear process using a discrete version of the Kalman filter. The Kalman filter (Maybeck, 1979; Reid, 2002; Welch and Bishop, 2001) is a common tool in tracking and navigation applications where the state of the system is updated by a combining the prediction and noisy measurements, (Dungate et al., 1999; Chi-Min et al., 1994). The combination of measurement and prediction is optimal in terms of residual mean squared error. The standard Kalman filter models linear systems. For other cases the principle of the filter can be extended to suit non-linear models, (Li et al., 2004).

If the actual position and velocity of a target in the image represent a state of linear system then the relation between previous and current states can be written as

$$\mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{w}_k \tag{7.6}$$

where $\mathbf{x}_k, \mathbf{x}_{k+1}$ are the previous and current state, $\mathbf{F}_k$ is the state transition matrix and $\mathbf{w}_k$ is additive noise of the system process. The observation of the states is done through an observation system represented by linear equation

$$\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k \tag{7.7}$$

where $\mathbf{z}_k$ is the observation or the measurement at time k, $\mathbf{x}_k$ is the state at time k, $\mathbf{H}_k$ is the observation matrix and $\mathbf{v}_k$ is additive measurement noise. The following assumptions are made:

- $\mathbf{w}_k$ and $\mathbf{v}_k$ are uncorrelated, zero-mean white-noise processes with known covariance matrices $\mathbf{Q}_k$ and $\mathbf{R}_k$. Both matrices are symmetric, positive and semi-definite.
- initial state $\mathbf{x}_0$ is a random vector that is uncorrelated with both system and measurement processes.
- initial state estimate $\hat{\mathbf{x}}_0$ is known and it has known covariance matrix $\mathbf{P}_0$.

The task is to obtain an optimal state estimate $\hat{\mathbf{x}}_{k+1}$ given the observations $\mathbf{z}_1, \ldots, \mathbf{z}_k$ that minimises the expectation of the squared error between the actual state and it's estimate, $E\left[\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}\|^2\right]$.

The solution is provided in recursive steps defined by the following equations. The index k corresponds to the previous step and the $k+1$ corresponds to the current step. The index $k+1|k$ represents the transition from previous to the current steps.

- Prediction step (time update):

$$\hat{\mathbf{x}}_{k+1|k} = \mathbf{F}_k \hat{\mathbf{x}}_k \tag{7.8}$$

$$\mathbf{P}_{k+1|k} = \mathbf{F}_k \mathbf{P}_k \mathbf{F}_k^T + \mathbf{Q}_k \tag{7.9}$$

- Update step (measurement update):

$$\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_{k+1|k} + \mathbf{K}_{k+1}[\hat{\mathbf{z}}_{k+1} - \mathbf{H}_{k+1}\hat{\mathbf{x}}_{k+1|k}] \tag{7.10}$$

$$\mathbf{P}_{k+1} = (\mathbf{I} - \mathbf{K}_{k+1}\mathbf{H}_{k+1})\mathbf{P}_{k+1|k}(\mathbf{I} - \mathbf{K}_{k+1}\mathbf{H}_{k+1})^T + \mathbf{K}_{k+1}\mathbf{R}_{k+1}\mathbf{K}_{k+1}^T \tag{7.11}$$

where $\mathbf{K}_{k+1}$ is Kalman gain matrix defined as

$$\mathbf{K}_{k+1} = \mathbf{P}_{k+1|k}\mathbf{H}_{k+1}^T[\mathbf{H}_{k+1}\mathbf{P}_{k+1|k}\mathbf{H}_{k+1}^T + \mathbf{R}_{k+1}]^{-1} \tag{7.12}$$

7.5.2 Kalman Tracking in Maritime Scenes

For motion estimation of a rigid object in a maritime scene, the state vector contains the following data

$$\hat{\mathbf{x}}_k = \begin{bmatrix} \hat{p}_{x,k} & \hat{v}_{x,k} & \hat{p}_{y,k} & \hat{v}_{y,k} \end{bmatrix}^T \tag{7.13}$$

where $\hat{\mathbf{p}}_k$, $\hat{\mathbf{v}}_k$ are the position and velocity estimates at time k and x, y denote horizontal and vertical components of the vectors.

The input measurement vector contains the data obtained from submersion line detection and corner matching

$$\mathbf{z}_k = \begin{bmatrix} p_{x,k} & v_{x,k} & p_{y,k} & v_{y,k} \end{bmatrix}^T \tag{7.14}$$

where $p_{x,k}$ is the centre of the submersion line and $p_{y,k}$ is the horizontal location of the submersion line as detected in frame k. Instead of velocities,

the average per-frame displacements $d_{x,k}$ and $d_{y,k}$ are used as measures. Both displacements are obtained from Equations 7.1 and 7.2 determined for each segment.

In many applications the measurement covariance matrix $\mathbf{R}_k$ is kept fixed in time. This is not necessary because the variance of the $d_{x,y}$ measurement is obtained as a part of the displacement fusion.

The vertical location measurement p_y corresponds to the position of the submersion line and the horizontal location p_x is at a centre of the segment width. The minimum change of the size of the segment is at least a couple of pixels given by the overlap of the segmentation grid. The change occurs when an object moves over the boundary of the segment. The change increases towards the bottom of the image as the resolution of segmentation grid decreases. This would cause a step change in the horizontal location measurement. In addition the size of the segment can fluctuate despite the object being still. This occurs when the distance value from Equation 4.24 for the particular segment is close to the classification threshold in the initial segmentation. Such fluctuations can cause false changes to the location measurements.

To reduce these effects a median value together with a variance taken over multiple frames are given as input location measurements. If the segment remains stable, i.e. its size and position does not change over time, the measurement variance would become zero, which would indicate infinitely large confidence in the measurement. Such over-confidence would then cause the displacement measurement values to be ignored for slowly moving objects. To enable the filter to estimate motion of slow objects a fixed value of two pixels is added to the variance of location measurement values. The value matches a typical localisation error obtained for sample sequences as presented in Section 7.5.5 (see Table 7.2).

The $\mathbf{R}_k$ matrix has a following structure

$$\mathbf{R}_k = \begin{bmatrix} Var_k(p_x) & 0 & 0 & 0 \\ 0 & Var_k(d_x) & 0 & 0 \\ 0 & 0 & Var_k(p_y) & 0 \\ 0 & 0 & 0 & Var_k(d_y) \end{bmatrix} \tag{7.15}$$

where the positions and velocities are considered uncorrelated (hence zero elements in R_k at corresponding positions) and

$$Var_k(u) = (1 - \frac{1}{k})Var_{k-1}(u) + \frac{1}{k}Var(u) \tag{7.16}$$

are temporal averages of all previous values of variances up to current time k and u is any of p_x, p_y, d_x or d_y. The averaging enables the noise levels of the measurements to be identified and stabilised. The state transition matrix $\mathbf{F}_k$ is defined as

$$\mathbf{F}_k = \begin{bmatrix} 1 & dT & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & dT \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where dT is a duration of a single frame. The measurement matrix $\mathbf{H}_k$ is defined as

$$\mathbf{H}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & h_k & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & h_k \end{bmatrix}$$

where h_k is either equal to dT if the displacement data are available or 0 otherwise. Changing the $\mathbf{H}_k$ matrix in such way allows to estimate the object state with only a partial observation available. When displacement data in the observation are unavailable the $Var_k(d_x)$ and $Var_k(d_y)$ are both set to infinity.

Finally, the process noise covariance matrix is defined by Reid (2002) as

$$\mathbf{Q}_k = q \begin{bmatrix} \frac{dT^3}{3} & \frac{dT^2}{2} & 0 & 0 \\ \frac{dT^2}{2} & dT & 0 & 0 \\ 0 & 0 & \frac{dT^3}{3} & \frac{dT^2}{2} \\ 0 & 0 & \frac{dT^2}{2} & dT \end{bmatrix}$$

where q is a constant set to $q = 0.1$. The values gives satisfactory results for the sequences used in the development.

The initial state vector and its covariance matrix are obtained by averaging corresponding measurement values over multiple initial frames. The state is updated every frame, together with the $\mathbf{R}_k$ matrix and the $\mathbf{z}_k$ input measurement vector. The output of the estimation is the state comprising the actual position and velocity of the object in the segment. An example of Kalman tracker applied to data in WEYMOUTH2A sequence is shown in Figure 7.6.

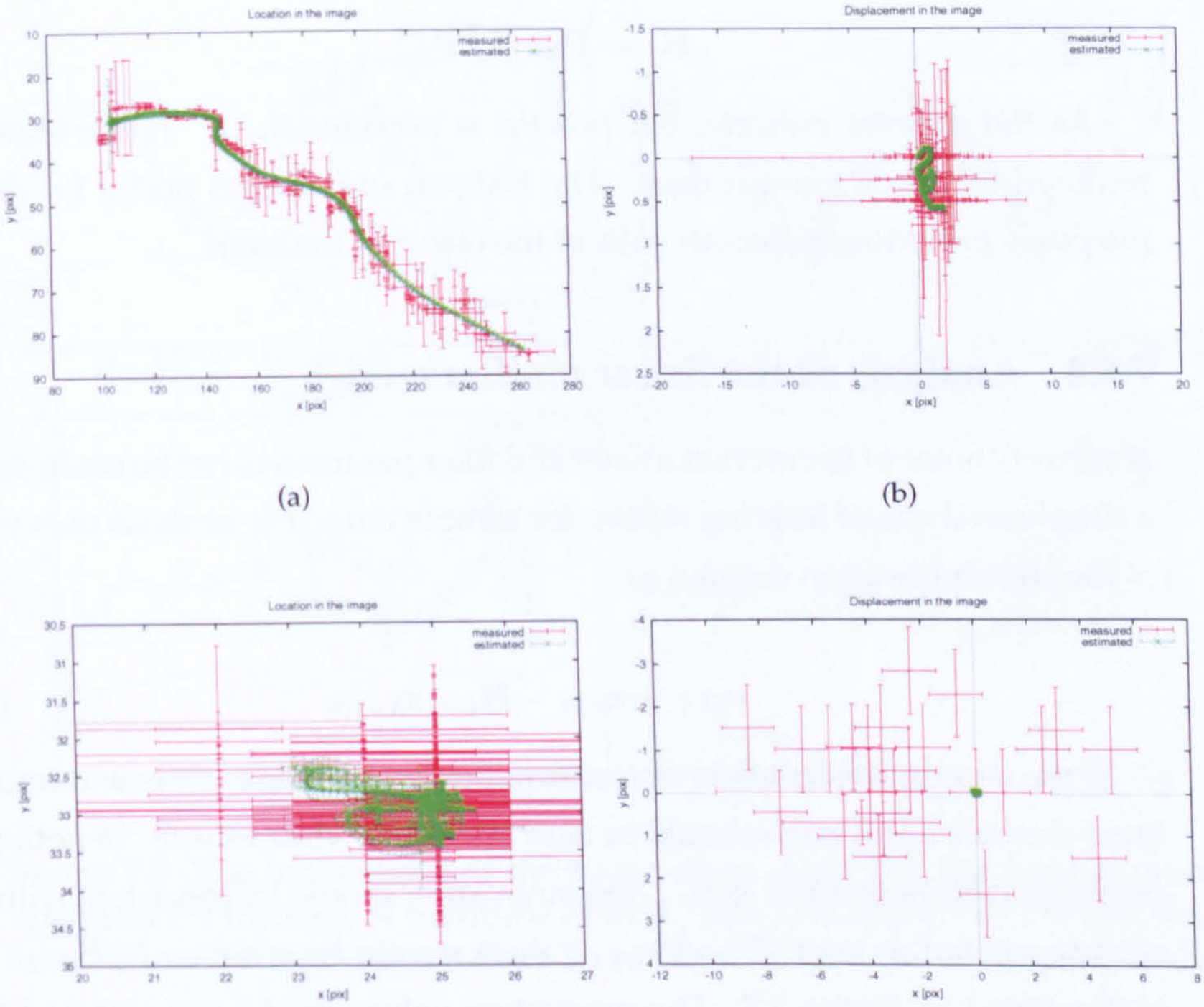


Figure 7.6: Kalman tracking of moving and static objects. The state is split into two parts: (a),(c) position, (b),(d) velocity of the objects being tracked. Both, measured and estimated, states are shown. The uncertainties of the measurements are outlined by the red crosses.

7.5.3 Kalman Smoother

To further minimise the error of state estimates found by the Kalman tracker a recursive method proposed by Anderson and Moore (1979) that updates the state vector and its covariance matrix in reverse order is employed. If N states are already estimated, then for every state $\mathbf{x}_k$ where $k = N, \ldots, 2$ a smoothed state vector $\widetilde{\mathbf{x}}_{k-1}$ and it's covariance matrix $\widetilde{\mathbf{P}}_{k-1}$ are obtained from the following relations

$$\widetilde{\mathbf{x}}_{k-1} = \mathbf{x}_{k-1} + \widetilde{\mathbf{K}}_k(\widetilde{\mathbf{x}}_k - \mathbf{F}_k\mathbf{x}_{k-1}) \tag{7.17}$$

$$\widetilde{\mathbf{P}}_{k-1} = \mathbf{P}_{k-1} - \widetilde{\mathbf{K}}_k(\widetilde{\mathbf{P}}_k - \mathbf{P}_{k|k-1})\widetilde{\mathbf{K}}_k^T \tag{7.18}$$

$$\tilde{\mathbf{K}}_k = \mathbf{P}_{k-1}\mathbf{F}_k^T\mathbf{P}_{k|k-1}^{-1} \tag{7.19}$$

As the indexes indicate, the process is non-causal, i.e. future states are required to obtain current ones. The Kalman smoother is useful for display purposes providing a smooth path of the object in the scene.

7.5.4 Analysis of the linear motion model

A correct choice of the motion model and filter parameters can be confirmed by a simple analysis of filtering results for sample data. The analysis uses values of the state innovation defined as

$$\nu_{k+1} = \mathbf{z}_{k+1} - \mathbf{H}_{k+1}\hat{\mathbf{x}}_{k+1|k} \tag{7.20}$$

If the assumption is made of measurement noise being white and uncorrelated then the innovation should be zero mean and white as well. Denoting the covariance of innovation as $\mathbf{S}_{k+1}$the innovation should be consistent with this covariance and at least 95% of ν_{k+1} values should lie within $\pm 2\sqrt{\mathbf{S}_{k+1}}$. This is illustrated in Figure 7.7. The innovation values used were obtained from the same sequence as in Figure 7.6 which contains a typical sequence of a boat approaching the camera. This simple test indicates that the filter is adequate for modelling the motion of objects in maritime scenes in the image plane as a clear majority of innovation values lies well within the standard deviation boundaries.

7.5.5 Detection of Occlusions

Occlusions often occur in open world scenes containing numerous moving objects with crossing paths. General perception is that occlusions are difficult to resolve, especially in complex structured environments, (Mirmehdi et al., 1996; Lipton et al., 1998). Maritime scenes are no exception. In theory, the strong geometric constraint of horizontal ground plane enables to distinguish which object is in front of the other from vertical positions of their image projections. The feature-based object characterisation, however, does not provide sufficient information about the object structure and appearance that could be utilised in resolution of the occlusion. If an object is occluded some of it's salient features become hidden and new features might occur at the

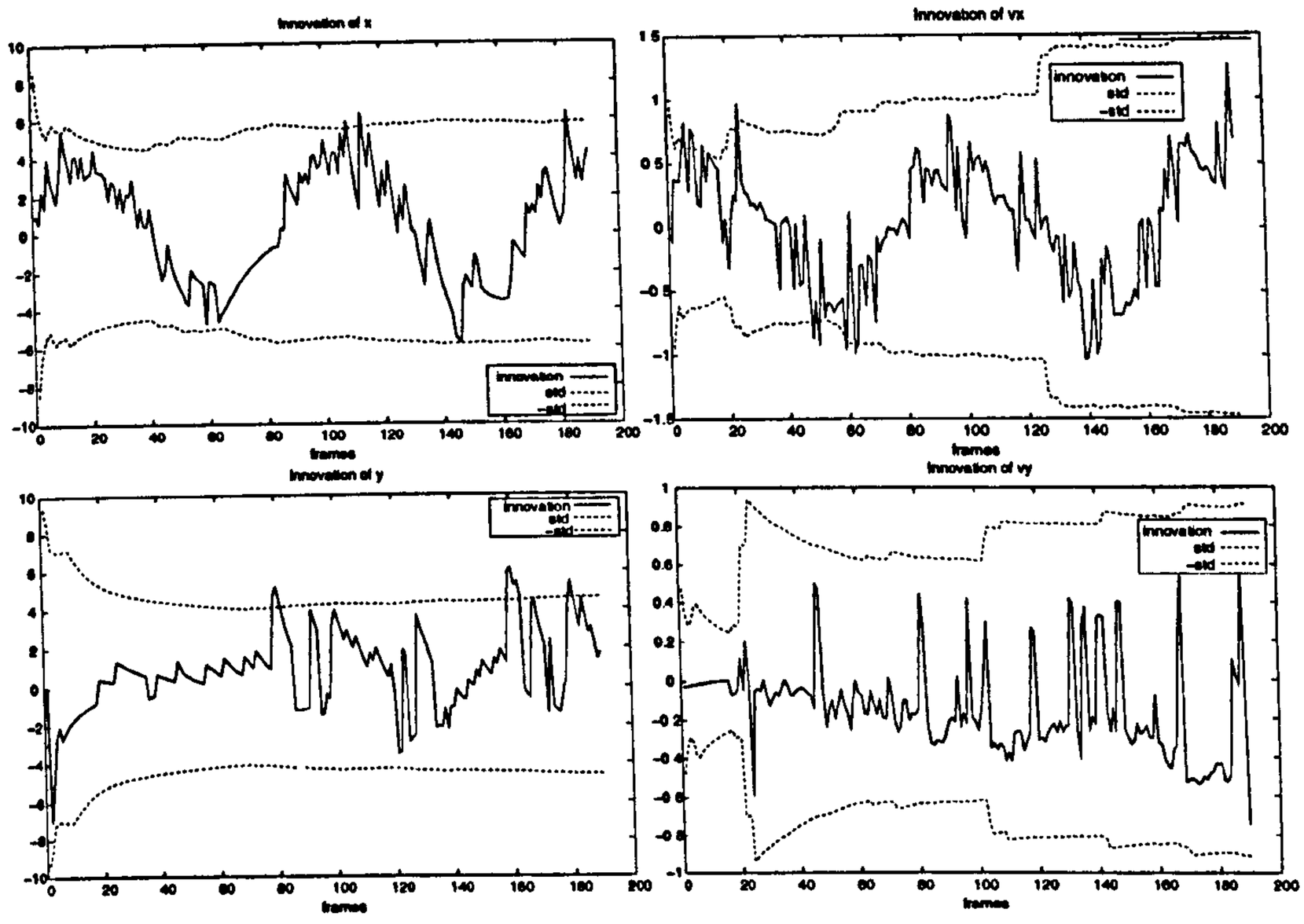


Figure 7.7: The innovations of the state values show that the Kalman filter is adequate for tracking of maritime objects. This is indicated by the fact that at least 95% of the innovation values lie within the standard deviation boundary.

occlusion boundaries. The features can be mismatched, producing wrong displacement estimates.

The segmentation presented in Chapter 4 is unable to separate objects that are closer than approximately the amount of overlap of the segments in the grid. Such close objects are treated as a single one. When two objects are moving towards each other their corresponding segments will join into a single one. The line of submersion will be detected and the objects will be treated as a single one. The detection of a submersion line always detects the line which is nearest to the camera. When the objects move apart, the segment will split and the tracking of two independent objects will be initiated.

The measurements of location and displacement during the occlusion are less reliable and their variance increases. An average standard deviation of the measurements embodies the overall certainty of the measurements. It is calculated as

$$\sigma_{avg} = \sqrt{R_y + R_{dx} + R_{dy}} \tag{7.21}$$

where R_y, R_{dx} and R_{dy} are variances of vertical location given by the submersion line and horizontal and vertical displacements. The horizontal

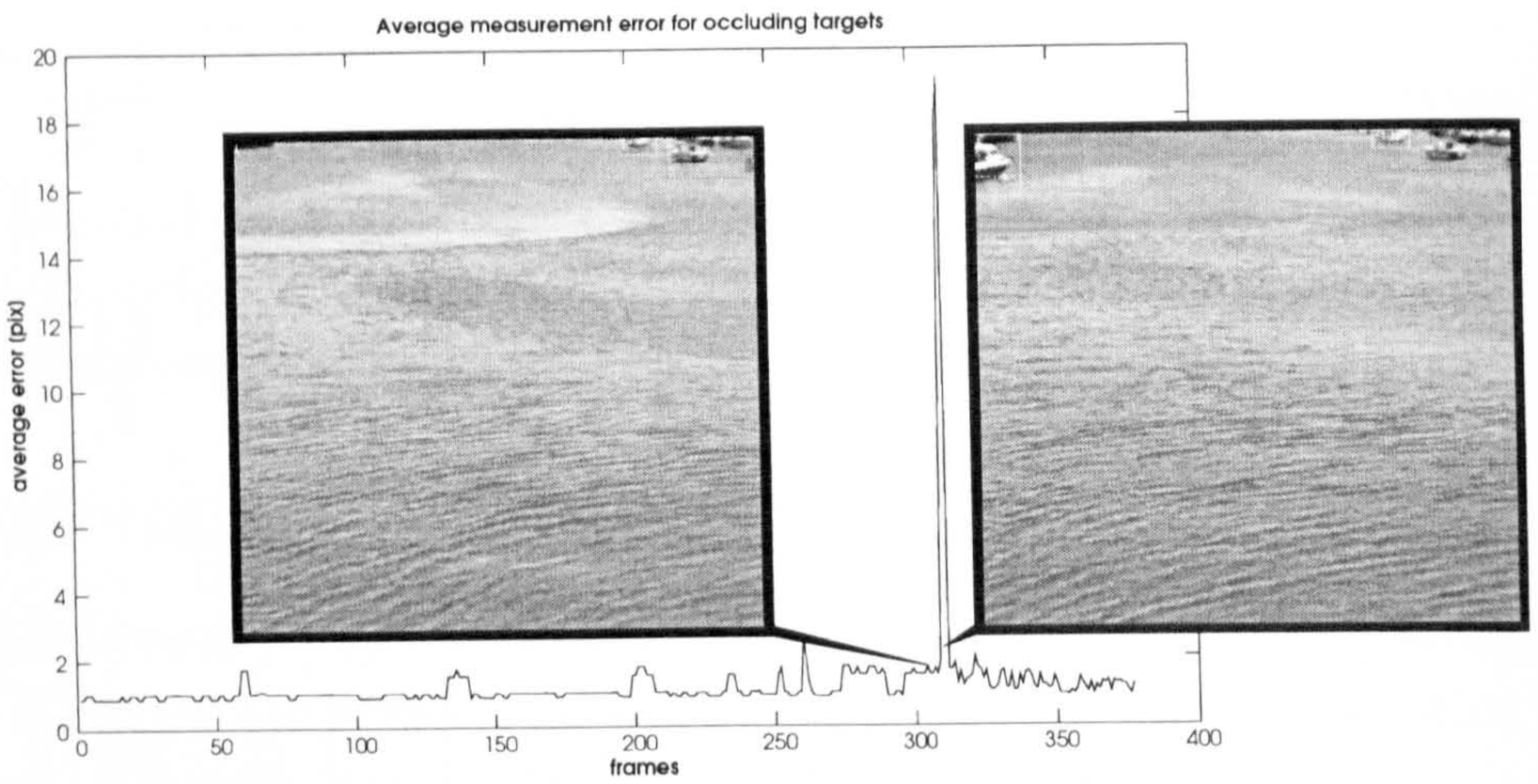


Figure 7.8: Change in average measurement error for occluding objects. The error for the pier on the left remains approximately static during tracking. When the pier becomes occluded by a boat coming from the left the measurement error increases instantly indicating sudden change in the structure within the segment. The Kalman filter is reset at this stage and it locks to the new object.

position is excluded from the calculation as it is derived from the segment width and it does not reflect the motion of the object.

An example of the average measurement deviation when occlusion occurs is shown in Figure 7.8. The occlusion occurs in the frame 311 of the WEYMOUTH2A sequence. A boat leaves the port on the left occluding the pier. The segmentation extends the segment covering the pier to include the boat as well. A submersion line closest to the camera is found in the segment. Because the position of the submersion line changes suddenly by a significant amount of pixels the measurement becomes uncertain. This uncertainty projects into the average deviation as a sharp peak. The uncertainty decreases back to a stable level after a very short time as the new location is confirmed in multiple consequent frames. Such 'jumps' in measurements would de-stabilise the Kalman filter producing unrealistic estimates. It is, therefore, necessary to detect these changes and re-initialise the filter accordingly.

A fixed threshold is applied to σ_{avg}. The value of the threshold is set to 5 pixels and it has been determined from an average standard deviation of measurements for a set of sample scenes containing targets of various complexity and dynamics. The whole unoccluded paths of objects from each sequence have been evaluated. The length of these sequences is between 150 to 400 frames. Table 7.2 shows that the average error is 3 ± 2.5 pixels.

Sequence	Target	Localisation Error	
		Average	Standard Deviation
WEYMOUTH2A	static	0.87	0.25
WEYMOUTH2A	moving	1.30	0.40
WEYMOUTH2A	static	1.25	0.21
POOLEH1	moving	1.87	0.54
POOLEH1	moving	1.47	0.40
POOLEH1	fluctuating	2.86	2.29
POOLEH1	fluctuating	1.88	0.69
WEYMOUTH2D	static	1.27	0.44
WEYMOUTH2D	moving	1.33	0.16

Table 7.2: The average standard deviation of the measurements of vertical location and displacements for sample maritime scenes and objects. The objects varied in their appearance, scale and dynamics.

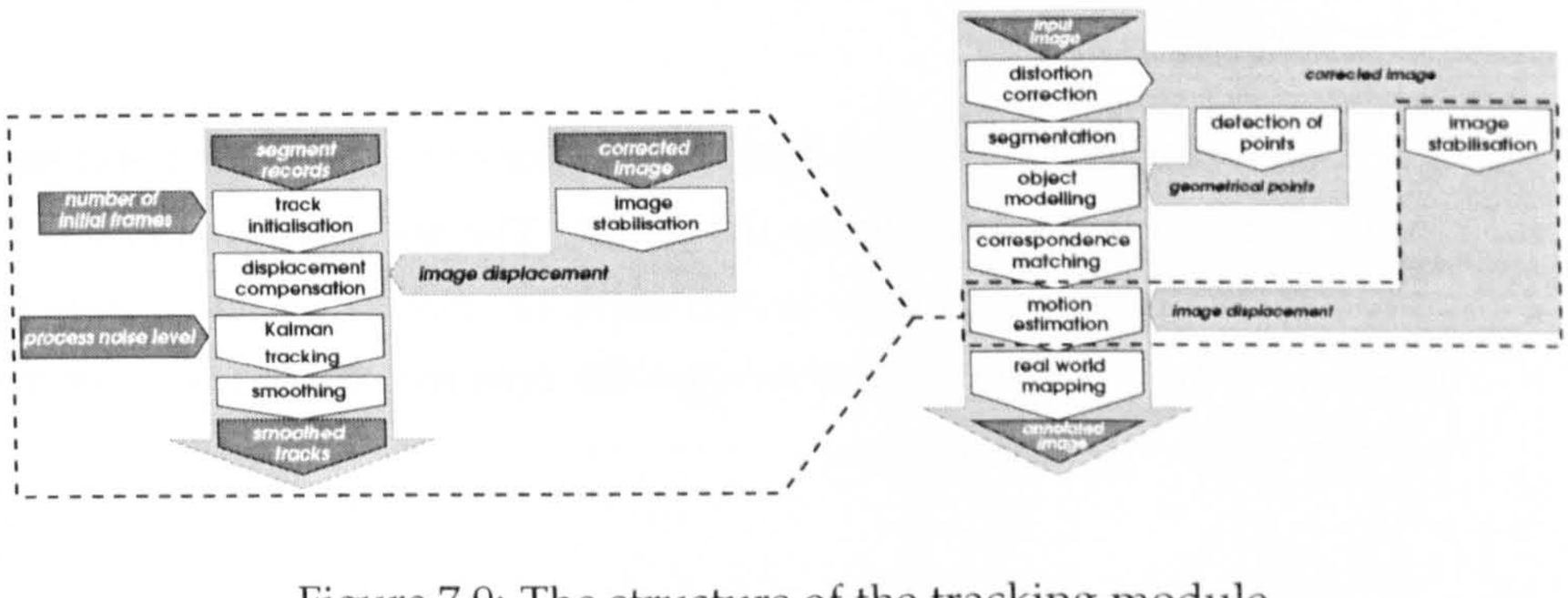


Figure 7.9: The structure of the tracking module.

7.6 Structure of Tracking Module

The tracking module estimates the states of objects within the scene. The state consists of the location and velocity of an object. The state estimation is done by Kalman tracking that operates on the displacement data in the segment records generated and maintained by the matching module. The structure of the tracking module is outlined in Figure 7.9.

The estimation consists of the following steps:

- *Track initialisation.* A Kalman tracker is initialised and associated with each object in the scene detected and matched in a specified number of initial frames. The initial estimates are obtained by averaging the input data in segment records over the specified number of initial frames.

- *Displacement compensation.* Systematic errors in the location and displacement data are filtered out. These systematic errors are caused by vibrations of the imaging platform due to environmental conditions such

as cross-wind impact. The amplitudes of the vertical oscillations are detected in the *Image stabilisation* section of the module. The amplitudes are determined by a correlation based registration of an image patch surrounding the horizon projection.

- *Kalman tracking*. The filtered location and displacement data enter the linear Kalman tracking. Possible occlusions are detected by a fixed thresholding of the current standard deviations of the input data. The Kalman tracking is re-initialised whenever an occlusion is detected.

- *Smoothing*. The variances of the state estimates are further reduced by the Kalman smoothing. The smoothing is non-causal and therefore it has mainly the purpose of improving the visualisation of the tracks of the objects.

The output of the tracking module consists of the estimated and smoothed states of the objects detected in the scene. The states consist of the location and velocity data in the image image units, i.e. pixels. The data enter the final module of the framework that remaps the data into the scene units using an inverse perspective projection.

7.7 Summary

A methodology for estimating the location and motion parameters of objects detected in the scene is presented. The estimation is based on displacements of geometric features detected in previous module of the framework. The tracks of objects are initialised by the data from the spatio-temporal database using a specified number of initial frames.

Geometric features for motion estimation are detected up to a certain precision and confidence. Even though the majority of the errors in detection can be regarded as uncorrelated noise, there are numerous sources of errors that are systematic and that can be filtered out. One type of the systematic error in maritime scenes originates from the vibration of the imaging device. An image stabilisation scheme is devised, based on image registration, that detects the vertical displacement of the horizon on a frame-to-frame basis.

The positions of the submersion line, displacements and their variances represent measurements and their uncertainties that enter the Kalman tracking. A discrete linear Kalman tracker is used for estimation of linear motion

parameters from noisy measurements. The suitability of the tracker is confirmed by results of an analysis of filter state innovations.

To avoid instability of the tracker in case the objects become occluded, the partial trace of the measurement matrix is compared against a preset threshold. The filter is re-initialised whenever the trace drops below the threshold, indicating that an occlusion occurred.

Finally, the results of the Kalman tracking enter a non-causal Kalman smoothing that further minimises the residual error of the estimation. The smoothing improves the tracks for the display purposes.

The results of the tracking and motion estimation enter the final module of the framework that relates the data to the real-world units and coordinates.

Chapter 8

Remapping

8.1 Introduction

Up to this stage the framework modules operate on a two-dimensional image data. Objects are detected, tracked and their motions estimated in units of image coordinates. The final module of the framework converts the information acquired from the image into a real world coordinate system and units. This will enable all activity detected in the image to be related to the real world structure. For a human operator the information is more comprehensible using real world coordinates and units than image-based ones.

The image to scene transformation is devised from a general perspective projection (Shapiro, 1995) by imposing geometry constraints applicable in maritime scenes as discussed in Section 2.3. Transformed information is then used for frame annotation. The original image is annotated and augmented with a radar-like view of the scene with all estimated and transformed parameters displayed.

Detection of possible collisions using a collision zone surrounding the observation point serves as an example of detection and assessment of predefined scenarios from the remapped data. In case a collision is detected a time to contact is determined which is a crucial information in decision making related to navigation of vessels. The collision zone surrounding the observation point represents just one of numerous early warning scenarios which can be configured.

It is important to analyse the precision and resolution of the results obtained

as, inevitably, restrictions on the detection range and precision will be inherent due to the discrete nature of the imaging device (Borman et al., 1999). Closely related to the topic of precision and resolution is camera calibration, (Clarke and Fryer, 1998). Most machine vision applications are intended for indoor use and majority of calibration methods are designed with that fact in mind. An alternative approach to calibration based on vanishing points detected in images of architecture is considered as a camera calibration alternative for outdoor applications (Cipolla et al., 1999).

8.2 Projective Transformation

The mapping of a real world scene to image coordinates is typically expressed in terms of a projection matrix. A general projection from scene to image can be written as, (Shapiro, 1995; Pettofrezzo, 1978; Mohr and Triggs, 1996)

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ T_{21} & T_{22} & T_{23} & T_{24} \\ T_{31} & T_{32} & T_{33} & T_{34} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} \tag{8.1}$$

where (x_1, x_2, x_3) and $(X_1, X_2, X_3, X_4)$are homogeneous coordinates related to image and world coordinates as

$(x, y) = (\frac{x_1}{x_3}, \frac{x_2}{x_3})$ and $(X, Y, Z) = (\frac{X_1}{X_4}, \frac{X_2}{X_4}, \frac{X_3}{X_4})$. The transformation matrix $\mathbf{T} = [T_{ij}]$ can be decomposed into

$$\mathbf{T} = \mathbf{CPG} = \begin{bmatrix} \xi f & 0 & o_x \\ 0 & f & o_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R_{xx} & R_{xy} & R_{xz} & D_x \\ R_{yx} & R_{yy} & R_{yz} & D_y \\ R_{zx} & R_{zy} & R_{zz} & D_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8.2}$$

where $\mathbf{C}$ is a calibration matrix accounting for intrinsic camera parameters (f is a focal length, ξ is an aspect ratio, (o_x, o_y) is the location of the principal point). As a convenience, the focal length is often expressed as two separate parameters, f_x and f_y instead of using intrinsic scale parameter ξ. Matrix $\mathbf{P}$ is the projection and matrix $\mathbf{G}$ accounts for extrinsic camera parameters by encoding a relative transform between the world and camera coordinate

systems. The elements of $\mathbf{G}$ correspond to rotation $\{\mathbf{R}_1^T, \mathbf{R}_2^T, \mathbf{R}_3^T\}$ and translation $[D_x, D_y, D_z]^T$ of the coordinate system.

From there, the relation between image point $\mathbf{x}$ and scene point $\mathbf{X}$ can be written as

$$\mathbf{x} = f \begin{bmatrix} \frac{\xi(\mathbf{R}_1\mathbf{X}+D_x)}{\mathbf{R}_3\mathbf{X}+D_z} \\ \frac{\mathbf{R}_2\mathbf{X}+D_y}{\mathbf{R}_3\mathbf{X}+D_z} \end{bmatrix} + \begin{bmatrix} o_x \\ o_y \end{bmatrix} \tag{8.3}$$

A simple example with the $\mathbf{R} = \mathbf{I}_3$ and $\mathbf{D} = 0$, so that image and scene coordinates aligned, and $\xi = 1$ and $(o_x, o_y) = (0, 0)$ gives

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{f}{Z} \begin{bmatrix} X \\ Y \end{bmatrix} \tag{8.4}$$

This transformation is clearly ambiguous, i.e. the relation holds for an infinite number of scene points lying on a line passing through principal point and image point (x, y). The relation makes a recovery of scene structure possible only up to an arbitrary scale.

Nevertheless, if the points in the scene are planar and the position of the plane relative to the camera is known, the planar structure can be recovered unambiguously. To illustrate this a real world coordinate system is provided such that the position of the camera is known up to the tilt angle ω around the $X-$axis (see Figure 2.3). All scene points are considered to lie on the $XZ-$plane. The camera is placed at the point $(0, H, 0)$ and oriented in such a way that it overlooks the $XZ-$plane. The Z-axis is assumed to always point in the direction of the optical axis of the camera. The edge of the image plane is assumed co-linear with the $XZ-$plane so that the yaw and roll angles κ, ϕ are both zero. The rotation matrix elements can be expressed using Euler's angles, (Jain et al., 1995)

$$
\begin{aligned}
r_{xx} &= \cos\phi\cos\kappa \\
r_{xy} &= \sin\omega\sin\phi\cos\kappa + \cos\omega\sin\kappa \\
r_{xz} &= -\cos\omega\sin\phi\cos\kappa + \sin\omega\sin\kappa \\
r_{yx} &= -\cos\phi\sin\kappa \\
r_{yy} &= -\sin\omega\sin\phi\sin\kappa + \cos\omega\cos\kappa \\
r_{yz} &= \cos\omega\sin\phi\sin\kappa + \sin\omega\cos\kappa \\
r_{zx} &= \sin\phi \\
r_{zy} &= -\sin\omega\cos\phi \\
r_{zz} &= \cos\omega\cos\phi
\end{aligned}
\tag{8.5}
$$

From there, the matrix $\mathbf{G}$ can be written as

$$
\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\omega & -\sin\omega & 0 \\ 0 & \sin\omega & \cos\omega & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8.6}
$$

By obtaining the intrinsic camera parameters from calibration the transformation matrix $\mathbf{T}$ in 8.2 becomes

$$
\mathbf{T} = \begin{bmatrix} f_x & o_x\sin\omega & o_y\cos\omega & 0 \\ 0 & f_y\cos\omega + o_y\sin\omega & -f_y\sin\omega + o_y\cos\omega & 0 \\ 0 & \sin\omega & \cos\omega & 0 \end{bmatrix} \tag{8.7}
$$

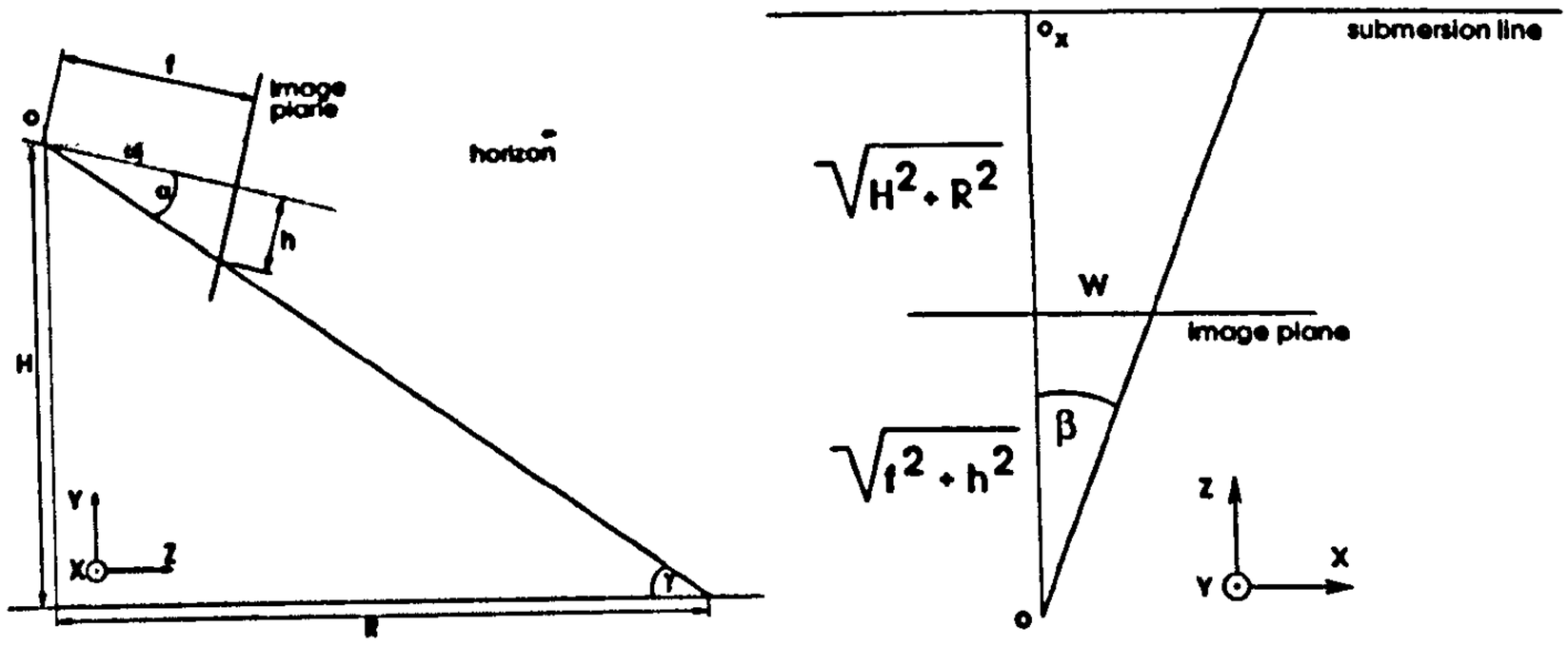
Making $\mathbf{T} = \{\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3\}^T$, a relation between image point $\mathbf{X} = (X, Y, Z)$ and scene point $\mathbf{x} = (x, y)$ can be written as

$$
x = \frac{\mathbf{T}_1(\mathbf{X}\ 1)^T}{\mathbf{T}_3(\mathbf{X}\ 1)^T} = \frac{f_x X}{Y\sin\omega + Z\cos\omega} + o_x \tag{8.8}
$$

$$
y = \frac{\mathbf{T}_2(\mathbf{X}\ 1)^T}{\mathbf{T}_3(\mathbf{X}\ 1)^T} = \frac{f_y(Y\cos\omega - Z\sin\omega)}{Y\sin\omega + Z\cos\omega} + o_y \tag{8.9}
$$

Assuming that for scene points lying on the plane $Y = H$, Equations 8.8 and 8.9 become

$$
x = \frac{f_x X}{H sin\omega + Z cos\omega} + o_x \tag{8.10}
$$



(a) Triangulation for range

(b) Triangulation for horizontal offset

Figure 8.1: Triangulations used to determine the range R and horizontal offset W of an object in the scene from image measurements h and w.

$$y = \frac{f_y(H\cos\omega - Z\sin\omega)}{H\sin\omega + Z\cos\omega} + o_y \tag{8.11}$$

It is clear from Equation 8.11 that there is a one-to-one inverse mapping between the position in the image and the Z-coordinate of the point in the scene. By inverting Equation 8.11 a mapping from image y-coordinate to the scene Z-coordinate is obtained

$$Z = \frac{H}{\tan[\omega + \arctan\frac{y-o_y}{f_y}]} \tag{8.12}$$

By inverting Equation 8.10 the mapping from image x-coordinate to the scene X-coordinate is obtained

$$X = \frac{x - o_x}{f_x}(Z\cos\omega + H\sin\omega) \tag{8.13}$$

The same expressions can be obtained from a simple triangulation. The configuration of the setup where a camera overlooks a planar scene from a height H above the plane under a tilt angle ω is shown in Figure 8.1a. The objective is to obtain the range R and the offset W on the scene plane in terms of an image measurement h and w. A main assumption is that the scene plane stretches to infinity in all directions, in which case the line connecting the horizon at infinity and the principal point are parallel to the scene plane. If

the location of the horizon projection in the image and the camera height H are known it is possible to obtain the range R assuming that

$$\gamma = \omega + \alpha \tag{8.14}$$

$$\cot\gamma = \frac{R}{H} \tag{8.15}$$

$$\tan\alpha = \frac{h}{f} \tag{8.16}$$

then R can be expressed as

$$R = \frac{H}{\tan(\omega + \arctan\frac{h}{f})} \tag{8.17}$$

It is the same relation as Equation 8.12, considering $h = y - o_y$, $f \equiv f_y$ and $R \equiv Z$.

The offset W is obtained from a simple triangulation outlined in Figure 8.1b,

$$W = w\sqrt{\frac{H^2 + R^2}{h^2 + f^2}} \tag{8.18}$$

It can be shown that the relation is equivalent to Equation 8.13 by substituting for $h = y - o_y$ from Equation 8.11 and considering $w = x - o_x$.

The f_x, f_y and H are all obtained directly from the camera calibration, the tilt angle ω is determined from the projection of the horizon. Assuming that the sea plane area under the observation of the camera is ideally horizontal and flat and that it stretches to infinity in all directions, the horizon can be placed at infinity. The tilt angle ω can be determined from the projection of the horizon by inverting Equation 8.17

$$\tan(\omega + \arctan\frac{h'}{f}) = \lim_{R'\to\infty}\frac{H}{R'} = 0 \Rightarrow \omega = -\arctan\frac{h'}{f} \tag{8.19}$$

where R' is the range of the horizon and h' is the projection of the horizon in the image.

8.3 Camera Calibration

The projection back to scene coordinates depends on the camera parameters that have to be obtained from a camera calibration process. These parameters

can be divided into two groups - intrinsic and extrinsic. The only extrinsic parameter that is crucial and that has to be obtained off-line is the height H of the camera above the sea plane. The value is easily obtainable as the structure of the vessel caring the camera is usually known to a great precision.

There are, however, numerous intrinsic parameters that influence the remapping accuracy as well. Most calibration methods for machine vision are built on well-established methods developed for aerial surveying, dating back to World War I, (Clarke and Fryer, 1998). The main principle of the calibration techniques is to determine the parameters of camera devices such as focal length, principal point, pixel aspect ratio and distortions due to the imperfections of lenses and structure of the cameras. This is best achieved from the projection of a target with known geometry. Many methods have been developed, most of them are based on the projection of 3D calibration targets with known geometrical pattern, (Stein, 1997; Heikkila and Silven, 1997; Bakstein, 1999), or 2D planar calibration pattern, (Bakstein, 1999; Brand et al., 1996; Zhang, 1998, 1999; Stein, 1993). For applications in robotics, the calibration of cameras with zoom lenses is often needed, (Li and Lavest, 1995; Li, 1994).

The intrinsic parameters that are usually the subjects of calibration are

- focal length in pixels (f_x, f_y)
- principal point (o_x, o_y) - point where optical axis intersects the image plane
- skew coefficient (α) - the angle between x and y pixel axes
- distortions $(k_1, \ldots, k_5)$ - radial and tangential lens distortion parameters

The lens distortion model as specified, for example, in (Heikkila and Silven, 1997), can be written as

$$\begin{bmatrix} x_d \\ y_d \end{bmatrix} = (1 + k_1 \mathrm{r}^2 + k_2 \mathrm{r}^4 + k_5 \mathrm{r}^6) \begin{bmatrix} x_n \\ y_n \end{bmatrix} + \begin{bmatrix} 2k_3 x_n y_n + k_4(\mathrm{r}^2 + 2x_n^2) \\ k_3(\mathrm{r}^2 + 2y_n^2) + 2k_4 x_n y_n \end{bmatrix} \tag{8.20}$$

where (x_n, y_n) are undistorted coordinates from a pinhole camera projection, (x_d, y_d) are distorted coordinates and $\mathrm{r} = \sqrt{x_n^2 + y_n^2}$ is the radial distance from the projection centre (assumed $(0, 0)$). The first part of Equation 8.20 corresponds to radial distortion and the second part represents the tangential

distortion. For wide angle lenses these distortions are more significant and they would cause a substantial systematic error in image to scene mapping. In order to avoid propagation of the systematic error through the processing chain of the framework all images entering the framework are corrected for the lens distortions. The distortion consists of a pixel-based geometric transform of the image that reverts the distortion given by Equation 8.20. Efficient implementations of the lens distortion correction can be found, for example, in (Tsai, 1987; Heikkila and Silven, 1997; Bouguet, 2004)

8.3.1 Calibration for Long-Range Imaging

Stein (1993) points out an important aspect of camera calibration: the calibration should be done for the range and depth of field at which the camera will operate in the actual application. The reason is, that the parameters established for a certain focal length do not remain the same for other focal lengths. Their changes are not linear and they cannot be easily extrapolated, (Li and Lavest, 1995).

Another assumption is that for a complete estimation of lens distortion parameters the target should cover the majority of the image used in calibration. For a near-range imaging this does not pose a significant problem as it is possible to build a target that would suit such constraints. However, for outdoor applications and longer imaging ranges any purpose-built target would be impractical. When calibrating aerial cameras (Clarke and Fryer, 1998), natural targets like stars and frozen lakes were often used as calibration targets. Another option is calibration using aligned collimators available in photogrammetric laboratories. However, such equipment is very expensive and of limited availability. The main drawback of all these methods is that they are designed for cameras that are projecting onto a real film which still has resolution much higher than the resolution of an off-the-shelf CCD device.

An alternative option for calibration for long range imaging applications arises from the use of architectural structures such as buildings where strong geometrical features of straight and parallel lines are inherent. Such a method is presented by van den Heuvel (1999). It uses vanishing points and straight lines (van den Heuvel, 1998) to estimate focal length, principal point and the first radial distortion coefficient. Even though van den Heuvel (1999) argues that the prime goal of the method is not a precise camera calibration for 3D reconstruction, it seems feasible as a less precise alternative to laboratory

methods using collimators.

8.3.2 Estimation Precision

An important issue closely related to the camera calibration is the precision achievable when transforming from the image to the scene. Because the image is a discretisation of the actual scene and the transformation is not linear any error in estimation also becomes non-linear. This is indicated in Figure 8.2. The resolution of a single pixel decreases with the range in the scene.

For example, for focal length of 1000 pixels and range 100 metres one pixel corresponds to approximately 2 metres, which means that the average error is 2% per pixel. However, for 1000 metres the error is almost 100% per pixel and the estimation of the range is impossible. This is the major drawback of using a standard imaging device for range estimation in long range applications (Reilly et al., 1999).

There are two options how to improve the resolution: larger focal length and higher image resolution. The relative resolution of the image is increased by increasing the focal length as the same field of view projects onto a larger area of the image. Because the area of the image is limited the increase in the focal length leads to the reduction of the overall field of view. A smaller field of view contains less of the scene structure and, therefore, it reduces the chance of the system detecting possible threatening objects.

The second option is to increase the resolution of the image. This is done by increasing the size of the CCD chip inside the imaging device. There is no reduction in the field of view. The only limiting factor is the cost of such an enhancement as larger CCD chips are more expensive and require more powerful processing platforms to allow the increased number of pixels to be processed in a required amount of time.

8.3.3 Calibration for Development Sequences

The camera used in acquisition of the sequences for development of the framework was calibrated off-line, in laboratory conditions using a planar target and calibration toolbox by Bouguet (2004). A sequence of 20 calibration frames of the planar target grabbed from differing positions was used. The depth of field was set to infinity and focal length preset to fixed values. Two values are used as various sequences were grabbed at two different focal

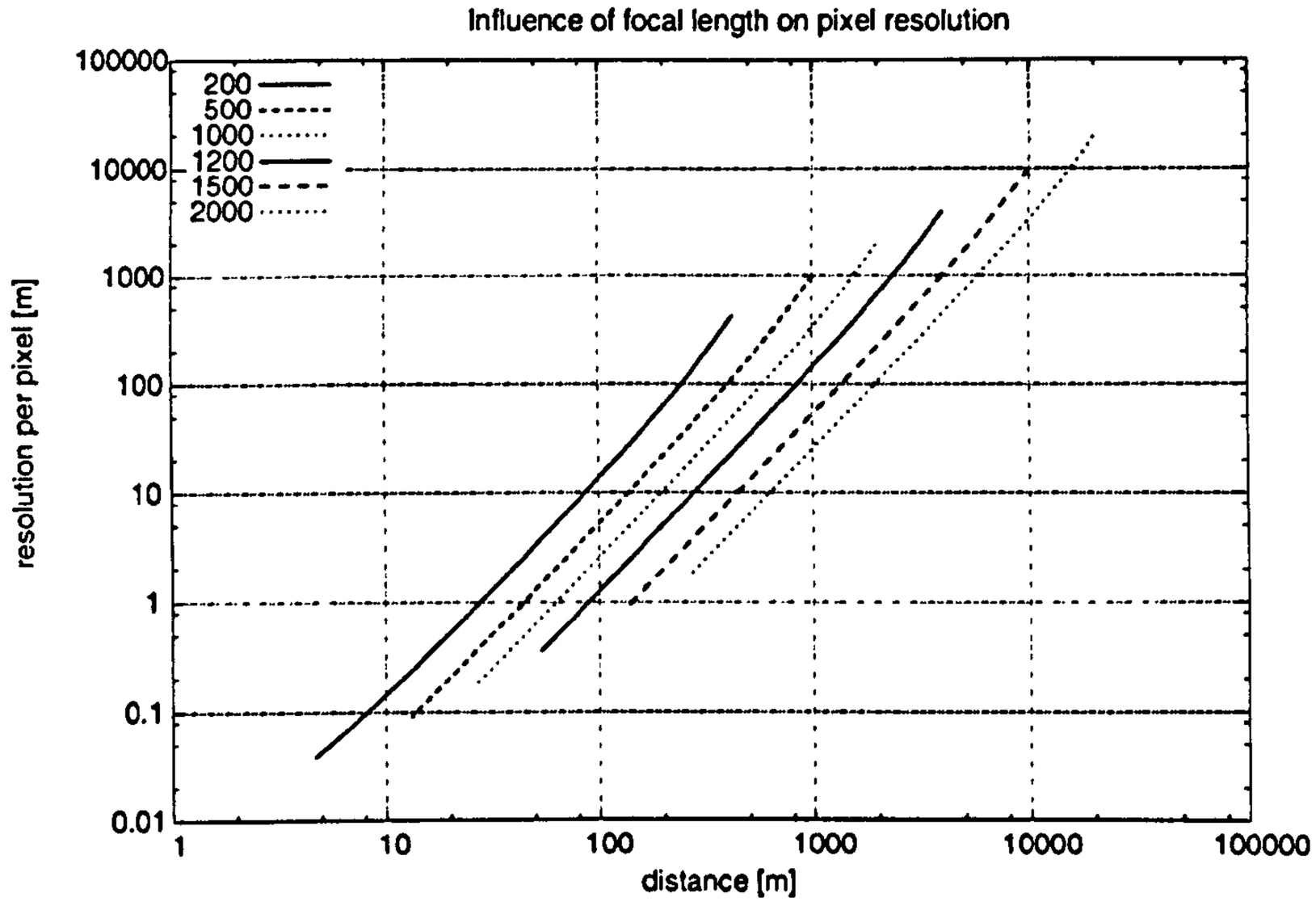


Figure 8.2: Pixel resolution as a function of distance in the scene where camera overlooks planar scene. Note that both axes are logarithmic. The focal lengths are given in pixels. As the plot indicates, due to the non-linear projective transformation and discretisation of the scene in image, the resolution is non-linear. The resolution is approximately reciprocal to the range. This restricts the possible precision of range estimation towards horizon. One solution, as indicated in the plot, is to increase the focal length. This, however, reduces the field of view.

f_x[pix]	f_y[pix]	o_x[pix]	o_y[pix]	k_1	k_2	k_3	k_4
843	840.2	387.7	279	-0.2912	0.2141	-0.0001	-0.0016
940	939	370	280	0.07	-0.07	0	0

Table 8.1: Calibration parameters. The calibration frames are 736×560 pixels.

lengths.

The results are shown in the Table 8.1. The obtained parameters are approximations of the values that could be obtained when using a calibration target matching the above stated requirements. The calibration parameters are put to use at the beginning of the processing chain. All frames are corrected for lens distortions prior to any further processing by a transform inverse to Equation 8.20. Focal lengths and principal point are substituted into Equations 8.10 and 8.11 and image coordinates are transformed to real world coordinates.

8.4 Image Annotation

The final stage of the framework processing chain is the presentation of the results in a comprehensible form to the human operator. The important cues (Hitchcock et al., 2003) for collision detection are the position, velocity and direction of motion of the object. It is necessary to draw the attention of the operator to any activity in the scene that might result in an accident such as targets on collision course towards predefined zones. For objects on a collision course a time to contact is crucial information in the decision making process. Provision of such an information is an essential purpose of any semi-automated surveillance system that is designed to help the operator in situation assessment.

8.4.1 Events of Interest

Events of interest are user-specified heuristic rules that trigger various pre-defined responses whenever location or motion data of objects meet the conditions specified by these rules. Three quantities enter the evaluation of the rules - position, velocity magnitude and velocity direction. The rules can either test the individual quantities or their combinations. The specification and representation of the rules defining events of interest is application-specific. The following examples are some illustrations of how the events of interest can be defined.

Security zones A security surveillance application monitoring a harbour might require detection of any craft entering private moorings. The entrance to the moorings is overlooked by a camera on a fixed, elevated platform. A security zone is set up that covers the mooring entrance. The security zone is specified by an interval of ranges and bearings detected with respect to the location of camera. An event-specifying rule is based on the evaluation of the objects' current location data. The rule specifies that "any detected object with current location within the intervals specified by the security zone triggers the response". The response can be a notification of an operator, an alarm, launching of an automatic video logging facility, etc.

Collision Avoidance A camera mounted on a vessel overlooks the sea in front and detects any moving objects in the scene. Event of interest is specified

as detection of any object on a collision course. A collision zone surrounding the point of view is defined. The width of the zone corresponds to the minimum allowed distance between the vessel and any object passing it. The event-specifying rule is based on the evaluation of the velocity direction. The rule states that "any detected object with velocity vector pointing inside the collision zone is on a collision course and it triggers the response." The response can be a notification of an operator, an alarm, an automated change of course, etc.

Speed camera A camera is monitoring a busy confined area with imposed speed limit such as a harbour entrance. The event of interest involves any craft exceeding the speed limit. A simple rule based on the evaluation of the velocity magnitude of each detected object states that "any object with velocity magnitude greater than specified limit triggers the response". The response can be a notification of an operator, an alarm, launching of an automatic video logging facility for evidence gathering, etc.

The rules specifying the events of interest are not restricted to the cases presented above. A more complex scenarios can be built up depending on the targeted applications.

8.4.2 Time to Contact Estimation

In addition to the targeted events of interest discussed above a time to contact (TTC) is estimated in collision avoidance scenarios. Time to contact is a crucial information necessary in the process of decision making. Sufficient time ahead is necessary when planning any collision avoidance maneuvers. The maritime objects move with considerably larger inertia due to a low friction of the water. Any change of direction must be initiated well in advance.

The TTC is determined as the time for an object to reach the collision zone when moving at the current speed along the line connecting the centres of the submersion line and the collision zone. This connection line represents the line of sight between the object and the collision zone. The Kalman tracker produces estimates of object location and velocity states together with their uncertainties. These uncertainties can be employed in the estimation of the TTC as well.

The situation of a target approaching the collision zone is illustrated in Figure 8.3. The target moves towards the collision zone with velocity vector

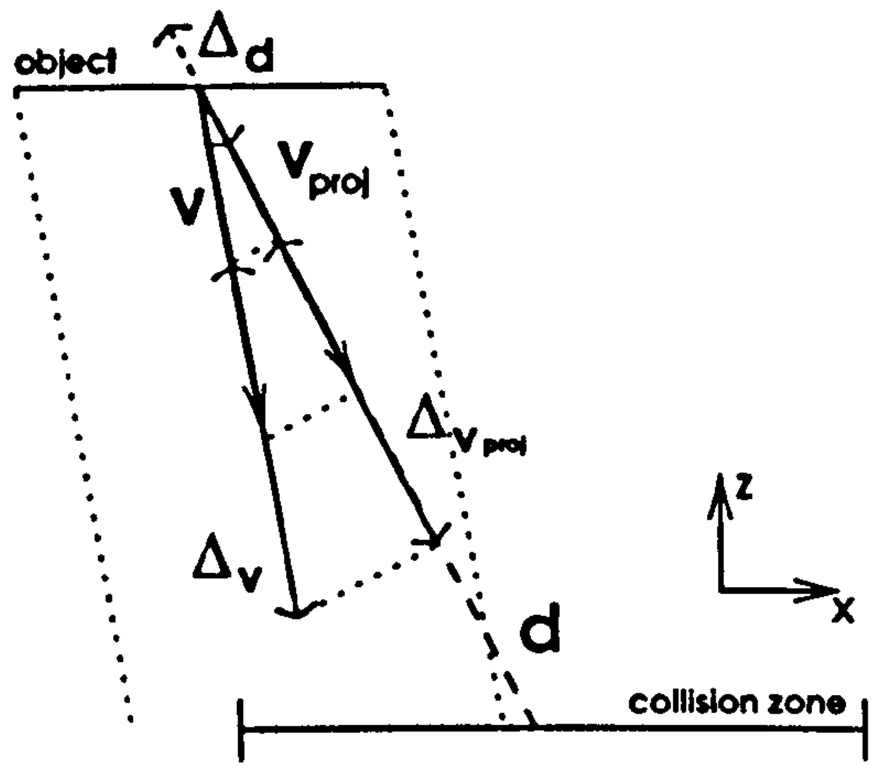


Figure 8.3: A target is on a collision course moving towards a collision zone. The Time To Contact estimates are determined from the projections of velocity, it's uncertainty and uncertainty of the location onto the direction of the line of sight. The line of sight connects centres of the object's submersion line and the collision zone.

$\mathbf{v}$ in such a way that its profile projected along the line of sight overlaps with the collision zone. The TTC is determined in the following steps

1. The length of a line of sight d connecting the submersion line and the zone centres is calculated.

2. The uncertainty of the location estimation in X and Z coordinates of the scene is projected onto the direction of the line of sight. The length of the projection Δ_d is calculated.

3. A projection $\mathbf{v}_{proj}$ of the velocity vector $\mathbf{v}$ onto the direction of the line of sight is obtained.

4. The uncertainty of the velocity estimation in X and Z coordinates is projected onto the direction of the line of sight. The length of the projection $\Delta_{\mathbf{v}_{proj}}$ is calculated.

Three values of the TTC are determined, pessimistic, centre and optimistic as defined by

$$\tau_p = \frac{d - \Delta_d}{|\mathbf{v}_{proj}| + \Delta_{\mathbf{v}_{proj}}} \tag{8.21}$$

$$\tau_c = \frac{d}{|\mathbf{v}_{proj}|} \tag{8.22}$$

$$\tau_o = \frac{d + \Delta_d}{|\mathbf{v}_{proj}| - \Delta_{\mathbf{v}_{proj}}} \tag{8.23}$$

The τ_p indicates how much time is left if the target moves at the highest speed over the shortest distance, the centre estimate τ_c is the most probable estimate and τ_o is the time left if the target moves at the lowest speed over the longest distance. All three values provide the operator with crucial information needed in planning of any action.

8.4.3 Image Annotation

The results are presented to the operator on a convenient radar-like chart due to the prevalence of radar applications in maritime traffic domain. Each frame of the sequence is annotated with detected segments highlighted and labelled by unique numbers. The segments contain the detected submersion lines as well. The centre of projection is marked at the bottom of the frame by a short vertical line. The line marks the origin of horizontal coordinate in the scene.

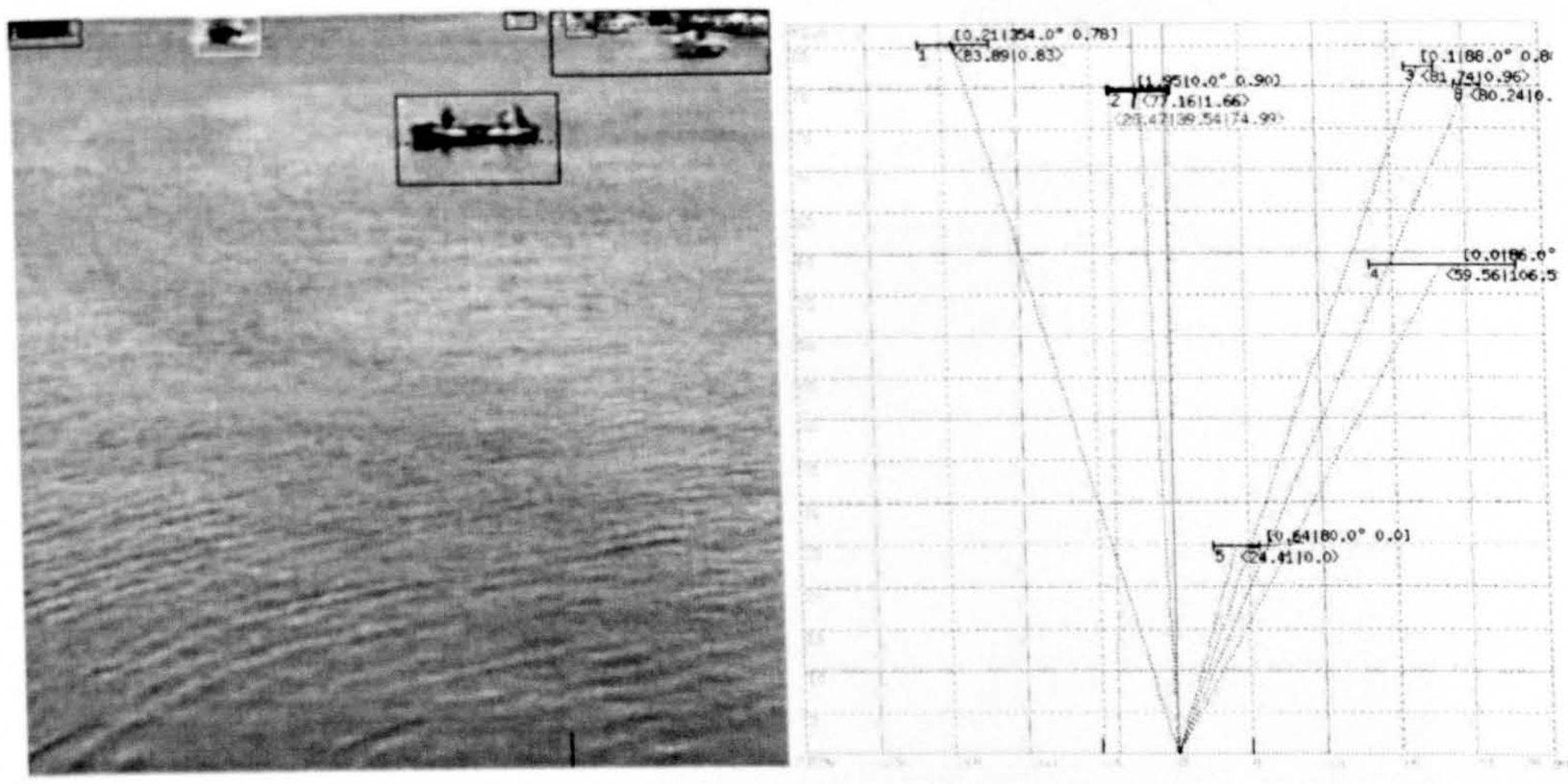
The image is augmented with a radar-like chart of the scene monitored by the camera with targets located at estimated locations. Each target is assigned the same number of the corresponding segment in the frame.

An additional information about the range and velocity of the target is displayed on the radar together with estimation uncertainties. If the target has been tracked for a longer period of time, it's smoothed track is shown as well.

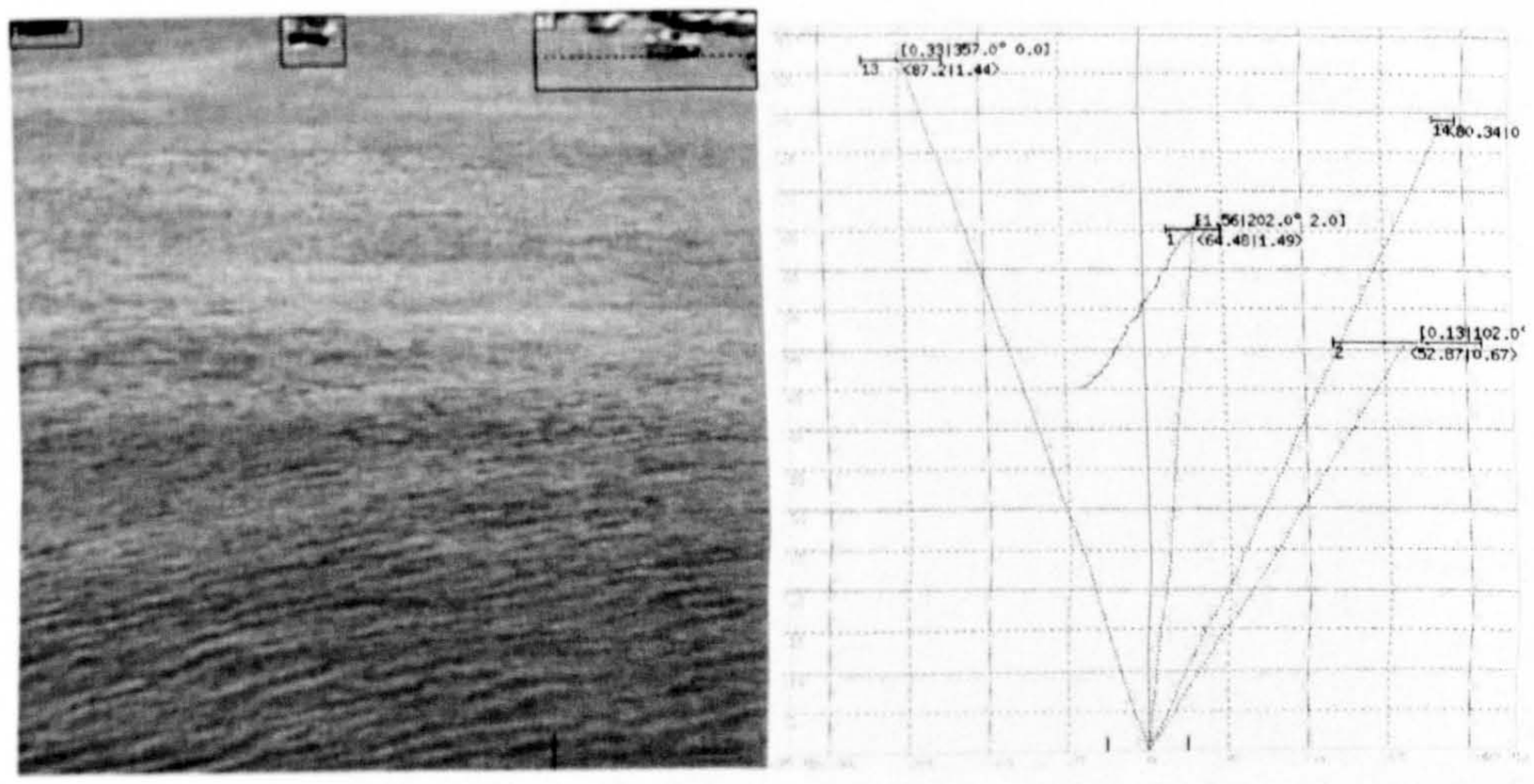
A collision zone is outlined on the radar as line on each side of the centre of coordinates. This zone has a predefined width and it represents the minimum distance from the observation point where the passing by is still considered as safe.

Objects on collision course are highlighted in the annotated frame and on the radar. The highlighting provides a visual cue to the operator that there is a collision likely to occur. A TTC values obtained from Equations 8.21-8.23 are displayed on the radar as well.

Outputs of the annotation process for sample sequences are shown in Figures 8.4,8.5.

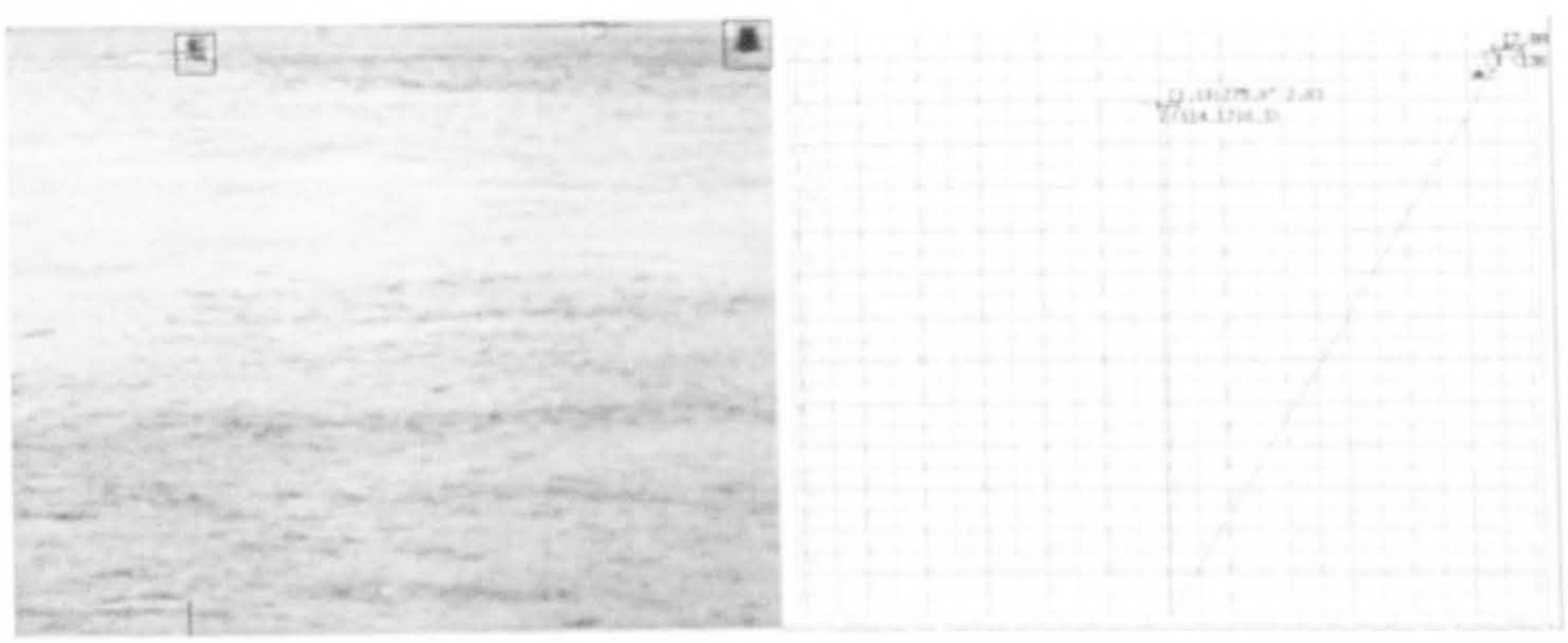


(a) frame 90

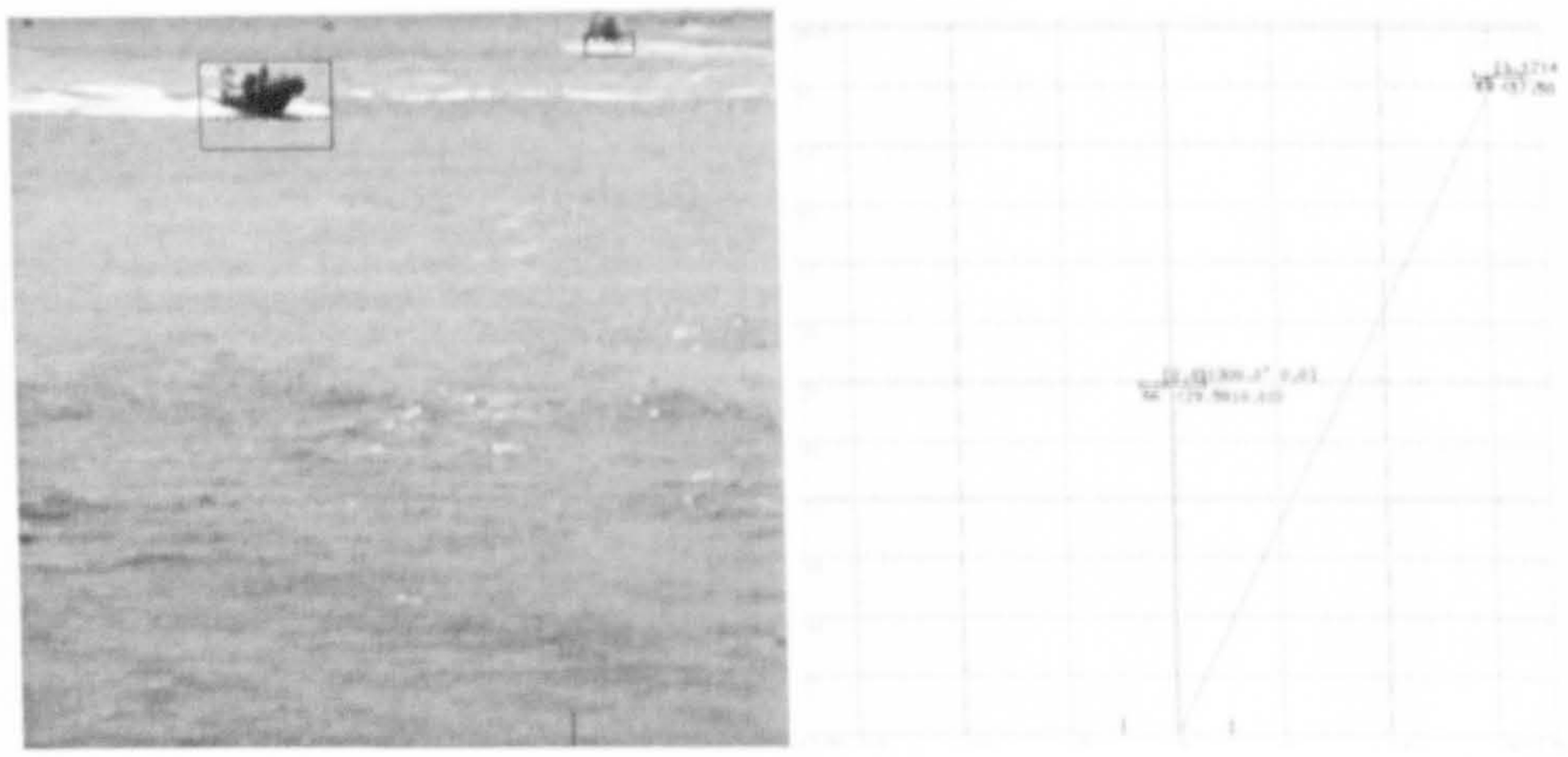


(b) frame 396

Figure 8.4: Tracking results for WEYMOUTH2A sequence. All data are in metres, velocities are in metres per second. The collision zone is outlined by two vertical lines at the bottom of the radar image.The first image (a) shows the target approaching the observation point that is on the collision course. It is highlighted in the original frame as well as on the radar overview.



(a) frame 120



(b) frame 411

Figure 8.5: Tracking results for WEYMOUTH2J,R sequences.

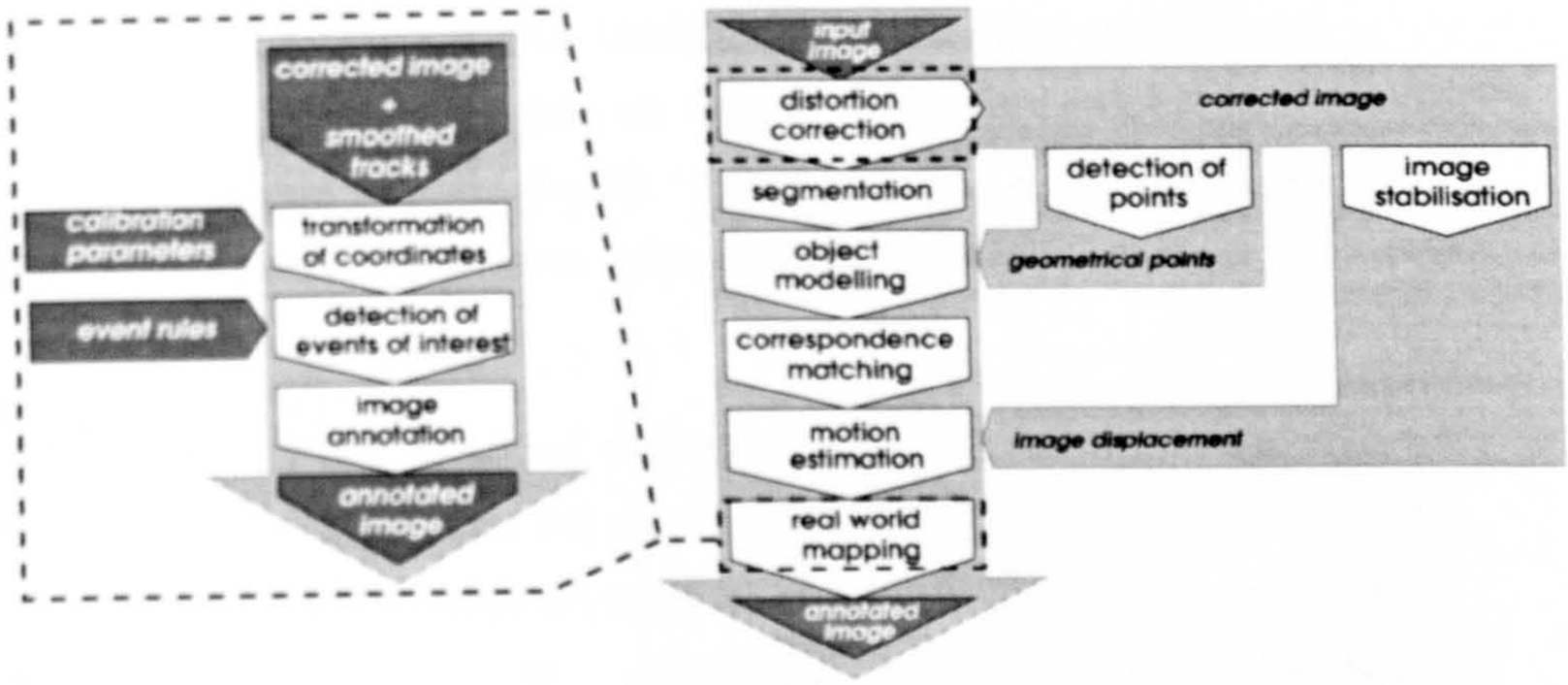


Figure 8.6: Structure of the remapping module.

8.5 Structure of Remapping Module

The remapping module is the last part of the framework. It takes the data provided by the tracking module as an input and re-projects them back to the scene co-ordinates and units. The data are also evaluated for any user-specified events of interest. The processing concludes by annotation of the original sequence and notification of the operator about any events requiring his attention such as collision threats. The structure of the remapping module is outlined in Figure 8.6.

The module consists of the following parts:

- *Transformation of coordinates*. The location and velocity data provided by the tracking module are transformed back to the scene co-ordinates and units in order to provide the operator with comprehensible information about the activity in the scene. The transformation is based on an inverse projective transform that is constrained by the Ground Plane Constraint, (Worrall et al., 1994). The parameters of the transform are obtained during an off-line calibration of the camera. The parameters are also utilised in the *Distortion correction* module of the framework that compensates input images for the lens distortion.

- *Detection of events of interest*. Events of interest are detected by evaluating the remapped location and velocity data against the set of predefined rules. The framework triggers a response whenever the conditions specified by the rules are met.

- *Image annotation*. The final step in the processing is the annotation of

the original sequence. The detected objects are highlighted and labelled. A radar-like overview of the monitored scene is generated with all important data such as speeds, directions and ranges of objects in the scene. Any objects involved in the events of interest are highlighted as well.

The annotation of the input sequence is one of many possible utilisations of the output of the framework. There are other applications including an integration of the framework into a more complex environments such VTS as discussed in the introducing Chapter 1.

8.6 Summary

The remapping module (shown in Figure 8.6) of the framework comprises of the inverse mapping, assessments of events and image annotation. The inverse mapping relates all the estimates obtained in the two-dimensional image to the three-dimensional scene structure. An unambiguous relation between the locations in the image and on the sea plane is obtained either from a general projective transform or by a simple triangulation.

Issues of camera calibration for long range imaging are discussed. An inherent limitation of the precision of the inverse mapping is analysed. The limitation is caused by non-linear projection of the sea plane onto a discrete image plane with a finite resolution. An improvement is suggested, either by increasing the resolution of the sensor or by increasing the focal length.

The output of the framework is presented to a human operator in a comprehensible form. Each original frame is annotated to indicate the detected objects. A radar-like chart is plotted and regularly updated. It contains all the important information such as the location and velocity of objects in the scene.

An early warning functionality is illustrated using a simple scenario of a collision zone surrounding the observation point. Any object moving towards the collision zone is highlighted by the system as a possible threat. A Time To Contact (TTC) is introduced as a measure crucial in decision making and action planning. It is determined from the location and motion of the threatening object relative to the observation point. Objects that are on collision course are highlighted in the original image as well as on the radar chart.

Chapter 9

Framework Cross-validation

9.1 Introduction

The cross-validation of the complete framework is done on two maritime sequences previously unused in the development of the framework in order to test whether the framework possesses scene and object independence. Both sequences are captured at scene conditions varying from those in development scenes. In addition, both evaluation sequences are taken from a moving camera mounted on a passenger ferry which differs from the development scenes that were captured by a static camera.

Both evaluation scenes contain objects that are varying in their appearances and motion characteristics. This will allow confirmation that the parameter values and thresholds are chosen by logical and experimental means and that they work for the majority of scenes.

The system is evaluated in the following three categories: object detection and tracking, motion estimation and inverse mapping. The evaluation methodologies used in each category are detailed in following sections.

9.1.1 Object Detection and Tracking

The system is evaluated for its ability to detect and continuously track genuine objects while minimising false negatives and positives. False negatives correspond to objects that are undetected in the scene. False positives correspond to detected regions that do not correspond to any actual object or it's part.

Detected objects are tracked by means of a feature-based correspondence matching. Motion parameters are estimated based on the matching. All tracked objects are also continuously assessed for predefined events of interest by evaluating the location and motion parameters using a set of predefined rules. The choice of the rules generally depends on the requirements of a particular application. The scenario proposed here is the one of collision detection where a collision zone surrounding the observation point is defined.

The performance of all detection, tracking and threat assessment is illustrated by an activity chart. The activity chart clearly shows if and when an object is detected, tracked or obeys the scenario rules. Detection and tracking of false positives is charted as well. The activity chart is used to determine time periods of tracking and threats relative to the periods of detection of each object. The chart also shows detection and tracking of false negatives and positives. The data in the chart express the detection sensitivity and tracking robustness of the tracker.

9.1.2 Motion Estimation

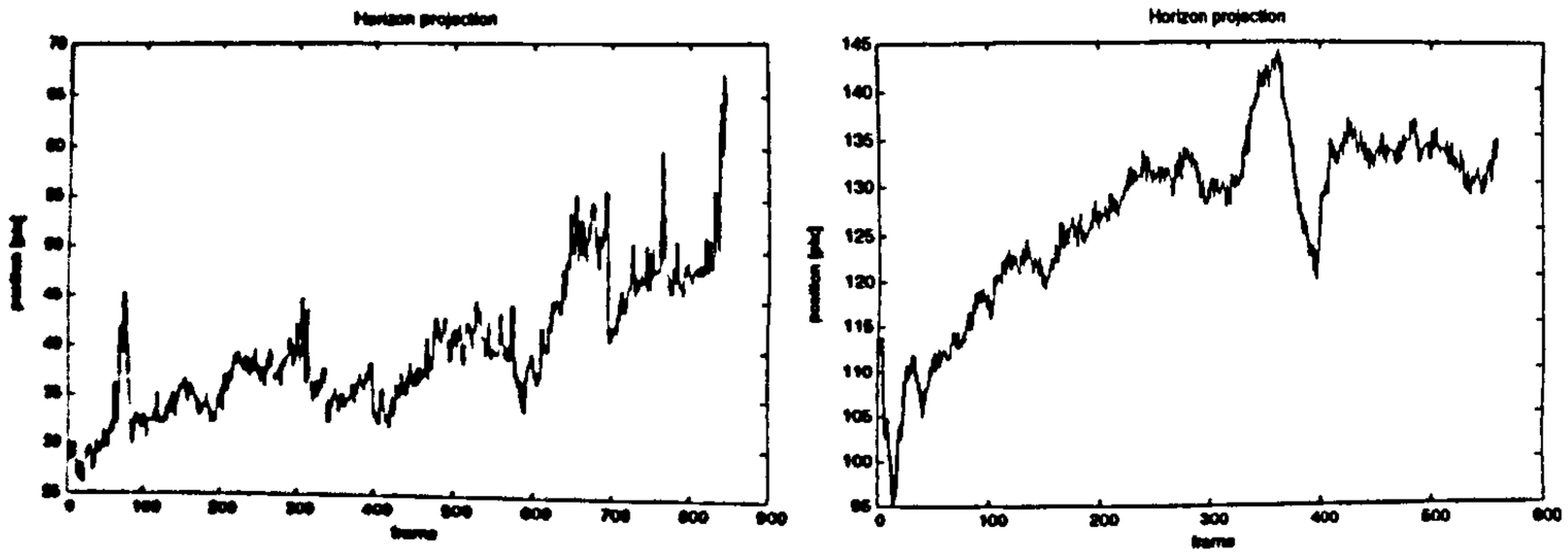
The stability and accuracy of the estimation of location and motion of tracked objects are evaluated. The level and consistency of estimation errors during tracking are also evaluated.

The evaluations are only in relative terms as the ground truth for the evaluation scenes is unavailable. Paths generated by Kalman tracker/smoother for objects in the scenes are plotted together with velocity vectors and location uncertainties.

9.1.3 Inverse Mapping

The consistency of the remapped values throughout the scene is evaluated. The hypothesis is that the estimated dimensions of an object in the scene are similar and independent of the position of the object in the image. The evaluation tests the hypothesis by determining heights of the buoys in the scene coordinates along their paths and checking the consistency of these values.

The buoys are chosen for the consistency evaluation as it is only their size that changes through the sequences. They appear as homogeneous regions of low intensity compared to surrounding sea which makes them simple to segment. A binarisation method by Otsu (1979) is applied to every detected



(a) Sequence A (b) Sequence B

Figure 9.1: Horizon oscillations in evaluation sequences. The apparent increasing trend in the height of the detected horizon is due to the fact that the camera has been hand-held during the acquisition process.

segment containing the buoy being tracked. The height of the buoy in the image is given as a vertical size of the corresponding blob.

The actual height of the buoy is obtained from the inverse projective mapping. Camera calibration parameters indicate square pixels as the difference between f_x and f_y is negligible. It is, therefore, possible to assume that horizontal and vertical resolutions are same at a particular depth of the scene. The actual height of the buoy is given as a product of height in pixels and horizontal resolution at the location of the buoy in the scene.

9.2 Evaluation Sequences

Both sequences were taken from a moving passenger ferry that travels between Cowes and Southampton Ports. The camera was positioned approximately 7 metres above the sea surface. In both scenes the horizon is projected within the visible area. This enables the horizon tracking to compensate for the oscillations of the camera. The detected oscillations are quite significant in both sequences (see Figure 9.1). There is an apparent increasing trend in the detected displacements of the camera. This is due to the fact that the camera was not mounted on a fixed platform during the capture but it was hand-held. Slight continual relaxing of the muscles due to the prolonged holding of the camera caused a vertical decline from the initial camera position.

The errors in position estimation of the objects in the scene due to the

presence of the waves are analysed in Section 2.3.2.2. However, the errors are not considered significant as the sea conditions in either of the scenes are calm with negligible significant wave heights.

The intrinsic camera parameters were not available at the time of a capture except for the focal length which was fixed to 100 mm in both scenes. The parameters were approximated from parameters obtained for a different camera of a similar type. The approximated parameters were

- $f_x \doteq f_y = 940$ pixels
- $o_x = 312$ pixels, $o_y = 267$ pixels
- $k_1 = -0.02$ and $k_2 = 0.17$.

The frame rate in both sequences was 12.5 frames per second which is a half of standard 25 frames per second, the frame size was 720×576 pixels.

9.2.1 Sequence A

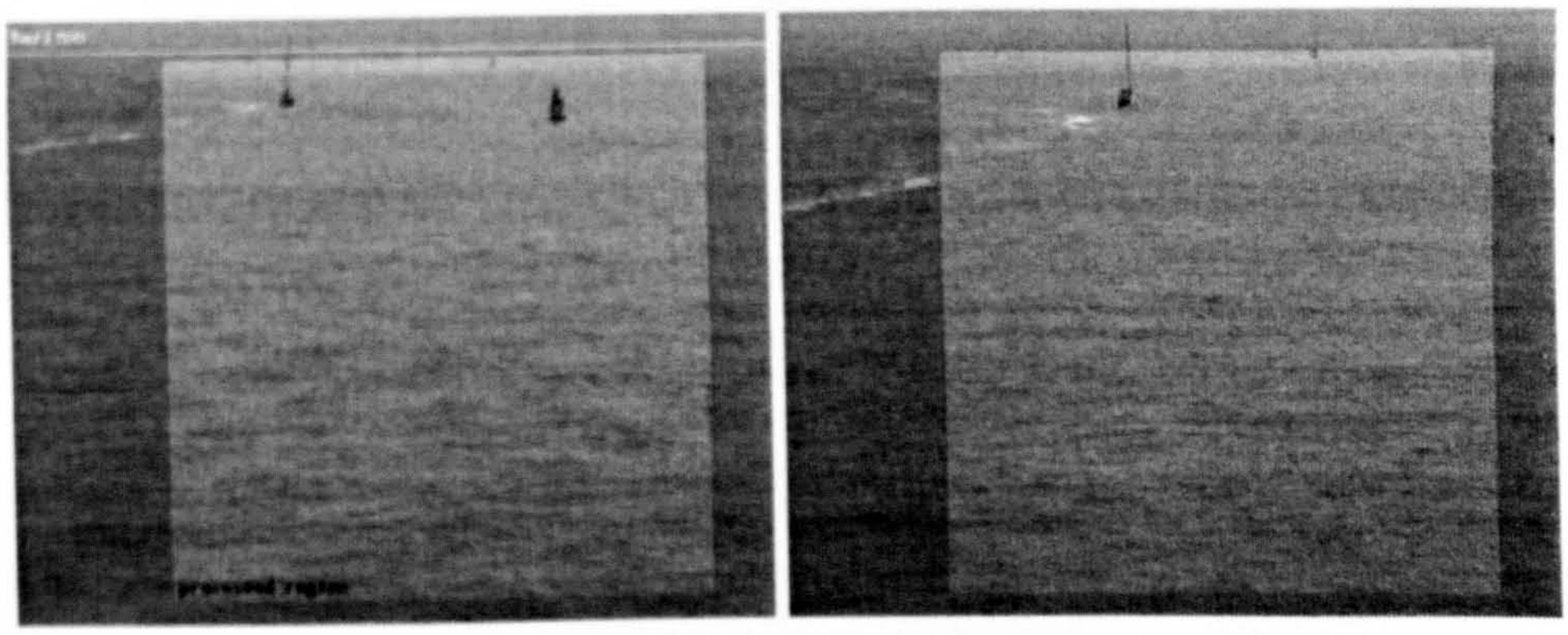
The sequence is 847 frames long. The initial horizon position is estimated at 30 pixels from the top edge of the frame. The processed region is 512×512 pixels, starting at position (150, 40) in the original frame (see Figure 9.2a).

The scene contains a single large channel marking buoy on the right, moving down and out of the image. There is a yacht moving from the top left to the middle right of the image. It is accompanied by a large and bright wake moving in a same direction. A second channel marking buoy appears on the right, near the top edge of the processed region. Channel marking buoys and the yacht all start within the scene.

The near channel marking buoy leaves the scene first, after approximately 210 frames. The yacht follows, after approximately 790 frames. The far channel marking buoy remains in the scene throughout the sequence. A large bright wake develops through the scene that travels towards the bottom edge of the image. It changes structure along it's way and breaks into smaller wakes as it reaches the observation point towards the end of the sequence. Figures 9.2a-d show the initial, intermediate and final states of the scene.

9.2.2 Sequence B

The sequence is 560 frames long. The initial horizon position is estimated at 112 pixels from the top edge of the frame. The processed region is 512×436



(a) frame 001

(b) frame 205

(c) frame 513

(d) frame 843

Figure 9.2: Evaluation sequence A. Initial (a), intermediate (b),(c) and final (d) states of the scene.

Sequence A	Large buoy	Yacht	Small buoy	Large wake	Other wakes
Detected	198	788	633	827	302
Tracked	194	729	593	400	112
Threat	0	86	1	138	9
Tracked	100%	93%	94%	49%	38%
Threat	0%	11%	0%	17%	3%

Table 9.1: The evaluation of object detections, trackings and threats for sequence A.

pixels, starting at position (200, 140) in the original frame (see Figure 9.3a).

The scene contains small channel marking buoy moving from the centre towards the left. A boat enters the scene at the top right corner of the image travelling at a constant speed and direction across the scene towards the left and centre of the image. A distant yacht and a second channel marking buoy appear near the centre and on the right of the top edge of the processed region approximately one third through the sequence. They both remain visible in most of the frames through the rest of the sequence.

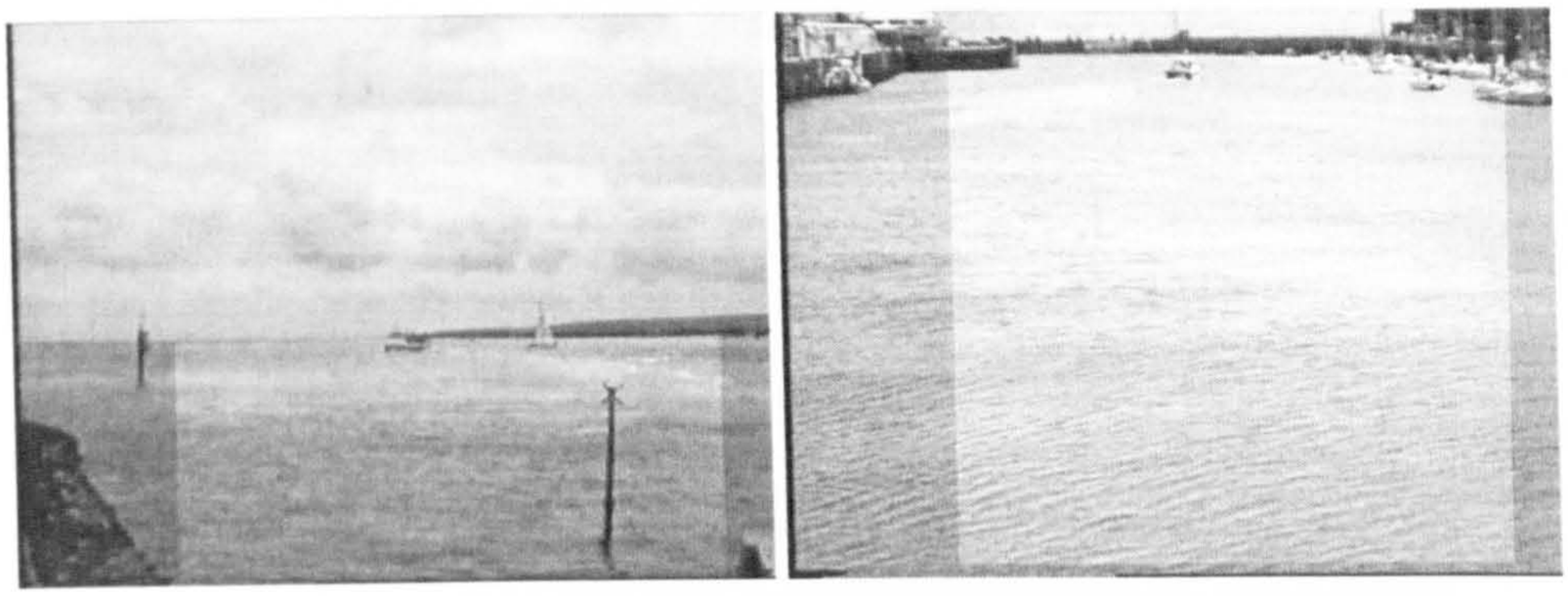
The first channel marking buoy starts within the scene and leaves the scene after approximately 74 frames. The boat enters the scene at frame 23 and leaves the scene at frame 308. The yacht and the second buoy enter the scene at approximately frame 232. They both remain in the scene till the end of the sequence with an occasional dropout caused by a horizon oscillation.

9.3 Evaluation Results

9.3.1 Detection and Tracking

Activity charts for both sequences are shown in Figures 9.4a,b. The charts indicate whether and when objects are detected, tracked and considered as threats. A single black vertical line corresponds to a single occurrence of one of the three events - detection, tracking and threat - in a single frame. Multiple occurrences in a single frame are colour-coded.

Tables 9.1 and 9.2 summarise the results presented in activity charts. The periods of tracking and threat relative to detection periods are provided in the last two rows of each table. The values are calculated for detection periods shortened by five frames necessary for initialisation of the Kalman tracker.



(a) SANDBANKS2H, fr. 109 (b) WEYMOUTH2A, fr. 796

Figure 9.5: Evaluated development sequences used to compare the consistency of the detection and tracking. The processed regions are highlighted.

SANDBANKS2H	BUOY	BOAT A	BOAT B	POLE	WAKES
Detected	100%	93%	97%	100%	6%
Tracked	100%	24%	86%	39%	0%
WEYMOUTH2A	PIER	BOAT A	BOAT B	FERRY	MOORINGS
Detected	100%	100%	100%	100%	100%
Tracked	92%	97%	100%	100%	81%

Table 9.3: The evaluation results of the detection and tracking obtained for the sequences used in the development of the framework. The values are relative to the number of the frames in which the objects are present in the scenes (a whole sequence is considered in case of the wakes).

9.3.1.1 Detection

No false negatives are encountered in detection of any of the objects in both sequences which indicates adequate sensitivity, scene and object independence and robustness of the initial segmentation algorithm. Objects are detected as soon as they enter the scene and detection continues until they leave the scene. Dropouts in the detection of certain objects are not due to any failure of the segmentation but they are caused by occlusions. An occlusion occurs in the sequence A where the yacht moving across the scene occludes the small buoy near horizon in the last third of the sequence. This is indicated by a gap in the activity chart in Figure 9.4a (labelled 'SMALL BUOY'). These results are consistent with the results for sample development scenes SANDBANKS2H and WEYMOUTH2A (see Figure 9.5) presented in the Table 9.3.1. The maximum amount of 7% of false negatives is detected in the SANDBANKS2H

sequence for a small target of a significantly low contrast at the horizon. The WEYMOUTH2A sequence contains no false negatives.

False positives are detected in 39% of the frames in the sequence as indicated in the last row of activity charts (labelled 'OTHER WAKES'). These false positives correspond to wakes that occur due to motion of man-made objects in both scenes.

The question is, should the wakes be generally ignored by the system or are they significant in threat assessment ? The wakes are mainly caused by presence or motion of rigid objects that can become subject to a collision. The large, bright wake in sequence A, for example, has well defined sharp contours and homogeneous texture and one can imagine a piece of debris of a similar appearance floating on the water. Smaller objects such as jet skis can be easily distinguished by the wake they generate, even over a longer distance. The conclusion is that wakes are usually associated with a presence and activity of objects that might become potential threat candidates.

The small wakes that appear in the sequence A are caused by breaking up of the large wake. The large wake moves towards the point of observation which is indicated by a high number of frames in which the wake is considered as a threat.

The wakes in the sequence B are caused by the motion of the boat across the scene. The wakes start to appear when the boat leaves the scene. Their number gradually decreases as the trail wake generated by the boat disintegrates.

9.3.1.2 Tracking

Once the object is detected, associated Kalman tracker and smoother are first initialised by data from five frames. The data from these frames are necessary to estimate the initial object state. Tables 9.1 and 9.2 show that all objects in both sequences are tracked at least in 93% of frames in which their presence is detected with two noticeable exceptions - the YACHT in the sequence B is tracked in 85% of the frames and SMALL BUOY II in the sequence B is tracked in 55% of the frames. There are two causes of the reduced tracking periods:

- no corners are detected - if an object is small or it hasn't got any salient features it is not possible to determine it's displacement.
- detected submersion line is unreliable - the position of the line detected over multiple frames changes by more than allowed amount of pixels.

This is a common issue with wakes and occlusions as the structure within the segment changes rapidly. It also happens with small objects near the horizon with low contrast and blurred appearance. The difference in intensity distributions covered by the detection mask is usually not enough to trigger a response of the line detector.

As soon as the geometric features are reliably detected again the tracking resumes. A gap of five frames is, however, necessary for tracker initialisation.

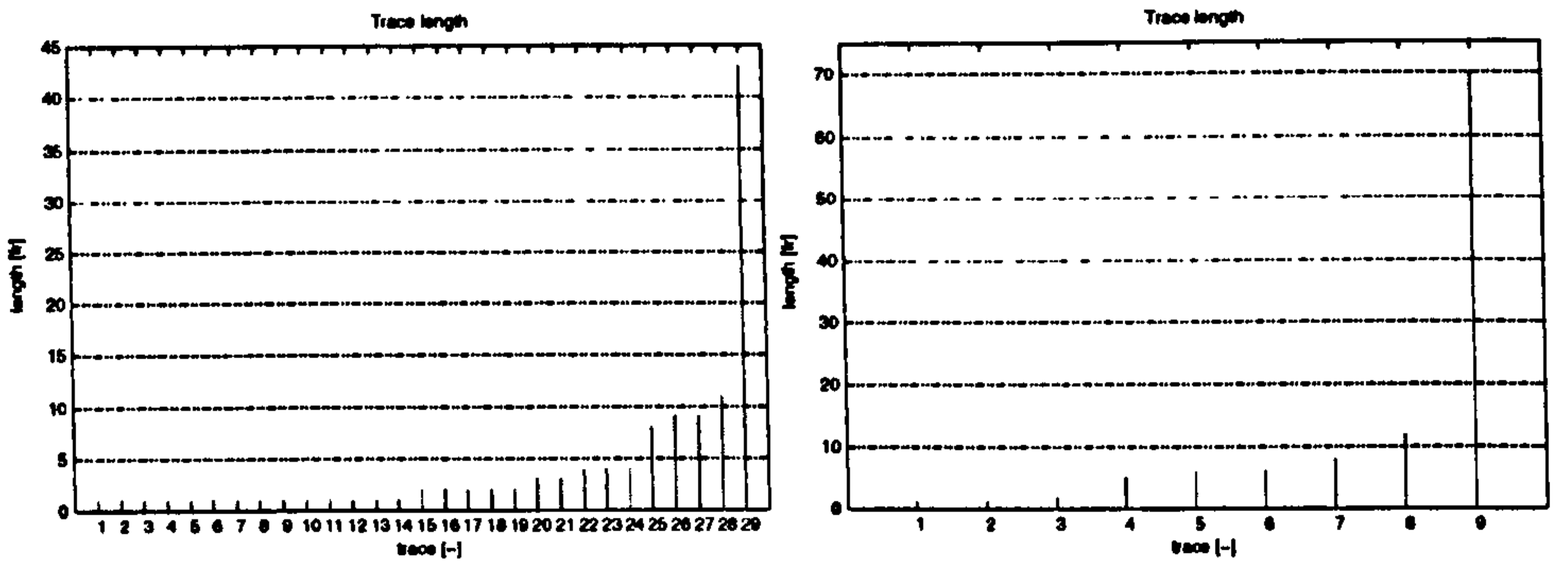
Occasional dropouts in tracking occur in the sequence A as the large wake moves closer to the yacht. Both are segmented as a single object causing the tracker to reset. The small buoy in sequence A is located close to the horizon. The buoy partially disappears from the scene and the tracker is reset as the horizon drifts upwards during the sequence. The large wake breaks up towards the end of the sequence making the detection of the submersion line unreliable. Many dropouts in the tracking occur as a consequence.

The large buoy and the boat in sequence B are tracked without any dropouts for most of the time. The tracking of the boat drops out initially for a couple of frames even though the boat is being detected. This is due to the fact that the boat is segmented as a part of a wake detected previously. The tracking resumes after five frames. The second dropout follows shortly after the wake and the boat separate.

The dropouts in tracking of the remaining two objects (the yacht and the buoy) are caused by the oscillating horizon. Both objects are very close to the top edge of the processed region and they disappear from the region on several occasions.

A number of small wakes are tracked in both sequences. The period of tracking for most wakes is relatively short due to their transient nature. If the wake is more persistent then it is probably associated with a presence and activity of an object and it is tracked for longer. Such is the case of the wake following the boat in the sequence B. Figure 9.6 shows tracking periods for various small wakes detected in the sequences. As the plots indicate most wakes are tracked for less than ten frames.

These results are consistent with those obtained for the development sequences listed in the Table 9.3.1 with two following exceptions. The BOAT A is near the horizon and there is not enough visible structure on which to detect any trackable corner. The POLE has a strong reflection in the water that prohibits a reliable detection of the line of submersion.



(a) Sequence A

(b) Sequence B

Figure 9.6: Tracking periods (trace lengths) of various small wakes detected in both evaluation sequences. As the plots indicate most of the periods are below ten frames. The longest traces are associated with motions of real objects in the scene (wake following the boat).

9.3.1.3 Threat Assessment

All tracked objects are assessed for a scenario of a collision threat. If the current velocity of an object points inside the collision zone surrounding the observation point the object is highlighted as a potential threat. The scenario of the collision threat is established by defining a 5 metres collision zone surrounding the observation point. An object is marked as a threat if it's motion vector points inside the collision zone.

The significant threats in the sequence A come from the yacht and the large wake as both subjects move across the scene and close to the observation point. The threat from the yacht occurs in the first half of the tracking period as the yacht moves from top left corner towards the centre of the image. The threat diminishes as the yacht passes by the image centre and continues to move towards the right edge of the image. The threat from the large wake is more persistent as the wake moves directly towards the observation point.

The only major threat in the sequence B comes from the boat moving from right to left across the scene. The threat is detected at the beginning of the tracking period. The threat diminishes as soon as the boat passes by the image centre. The secondary threat occurs when wakes generated by the passing boat move towards the observation point. The threat from these wakes does not last for a long, as the values in Table 9.2 indicate.

9.3.2 Motion Estimation

Figures B.1 - B.6 in Appendix B show the outputs of the Kalman tracking in image and scene coordinates. The results for objects that are tracked in at least 80% of frames are provided. The locations and uncertainties are plotted together with every tenth motion vector placed at the corresponding location.

Table 9.4 summarises the detection errors for each tracked object. The values are median standard deviations. The median is used to avoid influence of states with high variances that occur during the initial transient phase of the Kalman tracking.

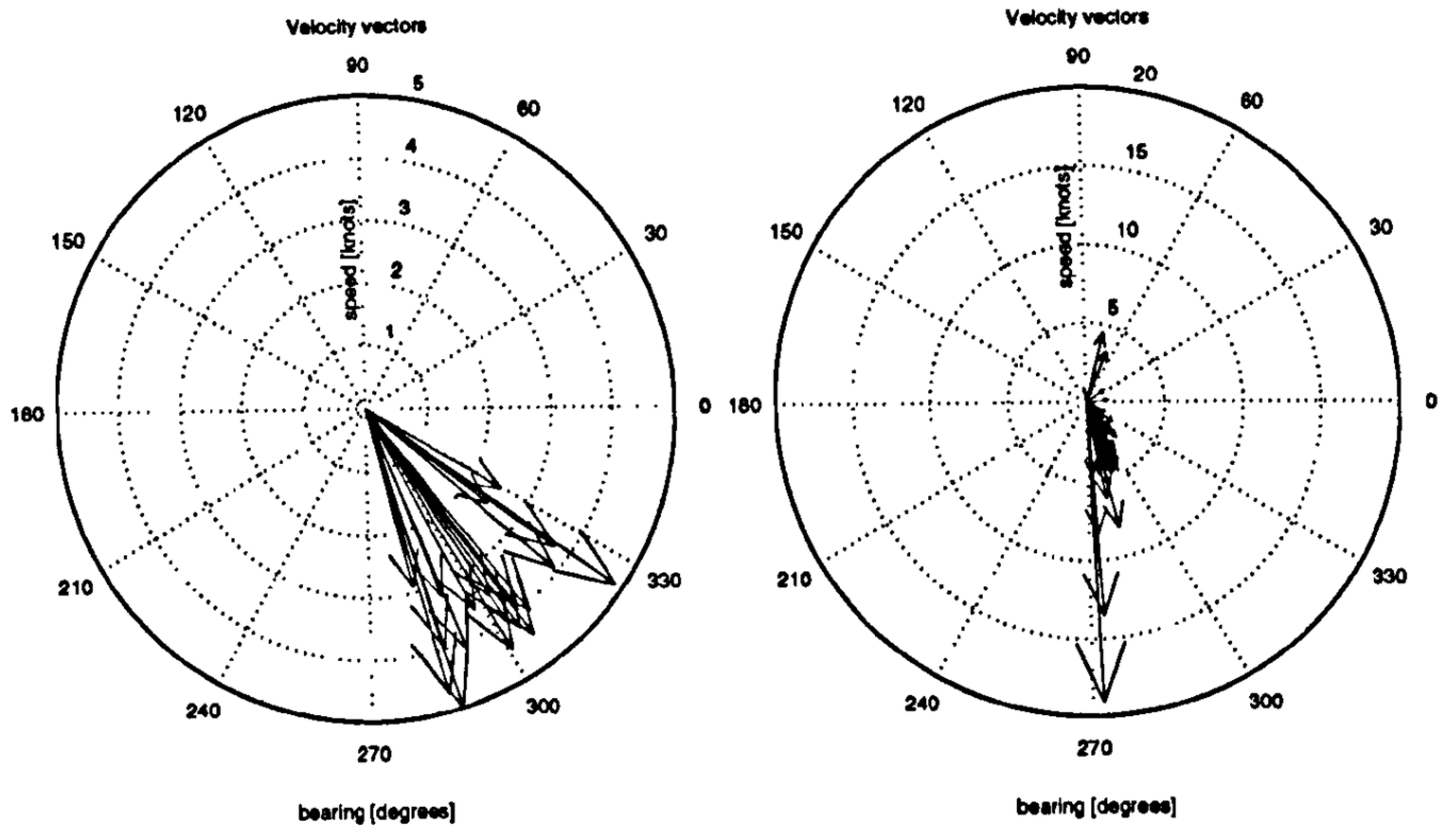
The transient phase of the Kalman tracker is characterised by a high variance of the estimated states. It occurs at the beginning of the tracking over a couple of frames as illustrated in Figures B.2a,b, for example. The transient state estimates are close to the actual states of an object despite initial high variance.

The non-linear mapping from image to scene coordinates dramatically increases the estimation errors for objects close to the horizon. This is illustrated in Figures B.3, B.6 as well as indicated by values in Table 9.4. Both objects are tracked in the image with errors less than one pixel. The inverse mapping causes approximately ten times larger errors than for other objects due to the fact that both objects are close to the horizon where the resolution per pixel decreases. The results confirm the limits of the system's precision caused by camera resolution as discussed in Section 8.3.2.

Figure 9.7 shows polar plots of velocity vectors for some of the objects being tracked. The speed is in knots (1 knot (international) = 0.51444457 m/s or 1 knot (UK) = 0.51477004 m/s) and the bearing is in degrees. The average speed estimates together with standard deviations are summarised in Table 9.5. Relatively high standard deviation is caused by inclusion of all motion vectors along the path of the tracking.

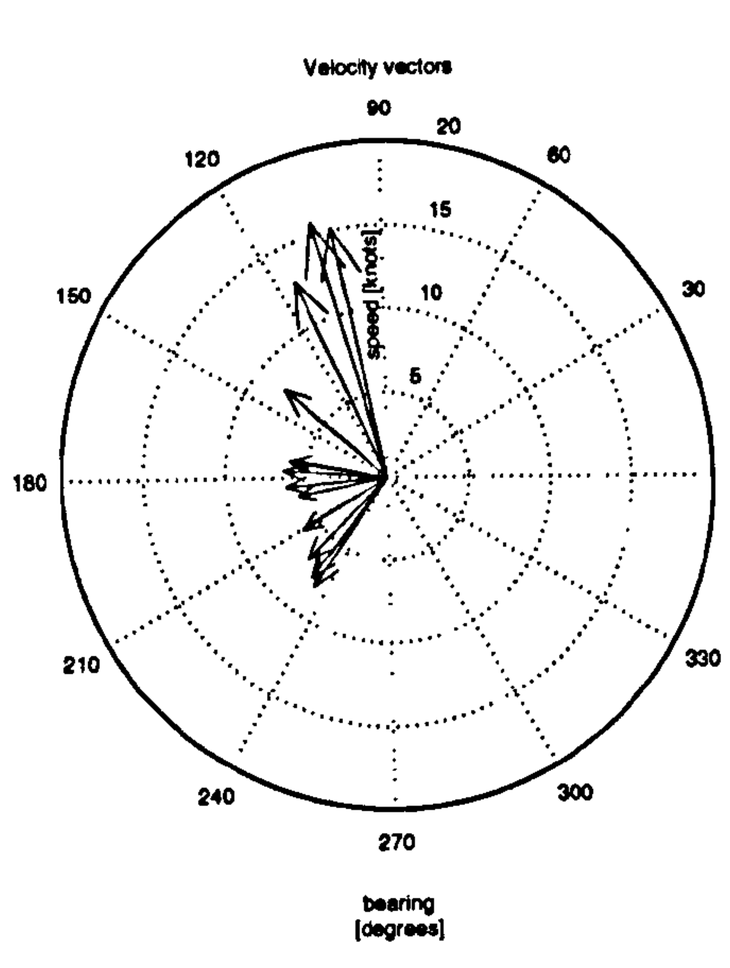
9.3.3 Inverse Mapping

The final evaluation tests the consistency of the inverse projective mapping by checking the heights of buoys in both the image and the scene. Both sequences contain channel marking buoys that are well-suited for the evaluation described above. Each buoy seemingly moves due to the self-motion of the ferry. The buoy in the sequence A moves predominantly towards the ferry,

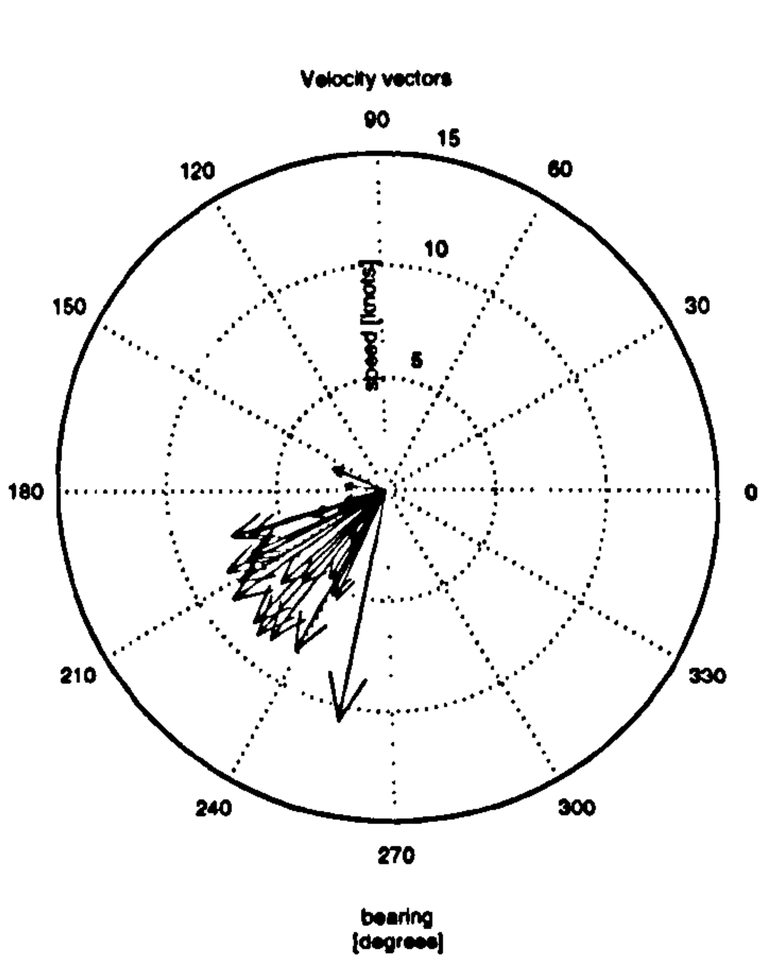


(a) LARGE BUOY (seq. A)

(b) YACHT (seq. A)



(c) LARGE BUOY (seq. B)



(d) BOAT (seq. B)

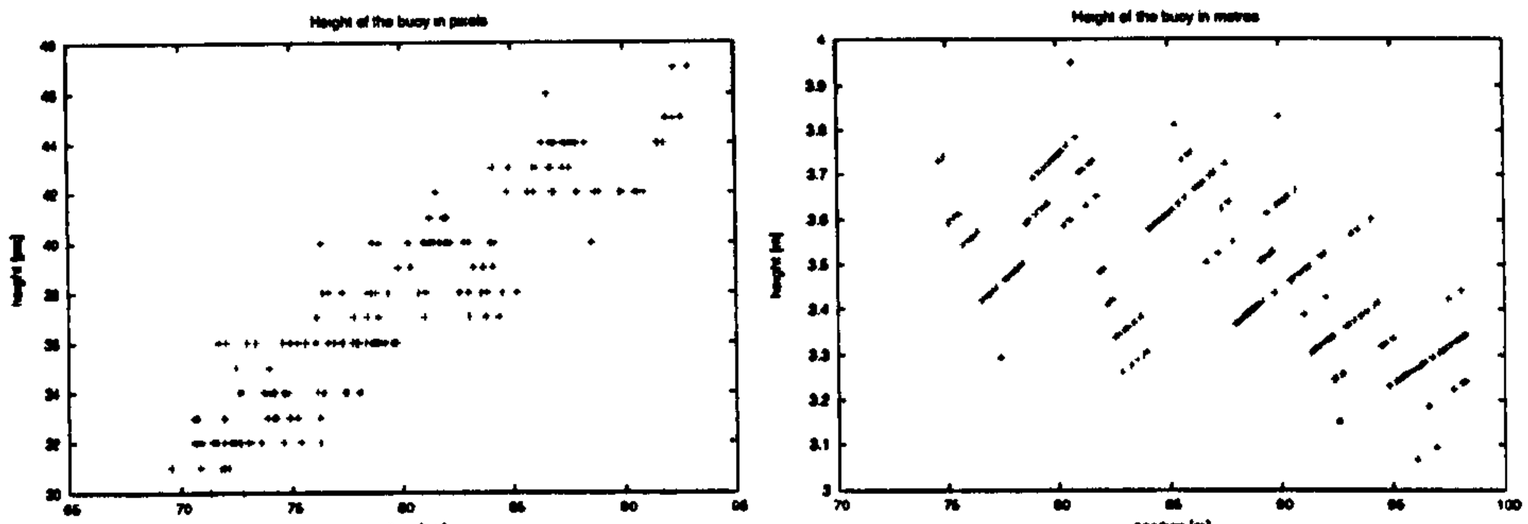
Figure 9.7: Velocity vectors of tracked sample objects. The speed is in knots.

	Image			
	location [pix]		velocity [pix/s]	
	x	y	x	y
LARGE BUOY (seq. A)	0.7	0.5	1.0	0.9
YACHT (seq. A)	1.1	0.8	1.1	1.0
SMALL BUOY (seq. A)	0.5	0.7	0.9	1.0
LARGE BUOY (seq B)	0.7	0.6	1.0	0.9
BOAT (seq. B)	1.6	0.4	1.3	0.9
YACHT (seq. B)	0.5	0.9	0.9	1.1
	Scene			
	location [m]		velocity [m/s]	
	x	y	x	y
LARGE BUOY (seq. A)	0.1	0.5	0.1	1.0
YACHT (seq. A)	0.1	0.9	0.1	1.1
SMALL BUOY (seq. A)	0.3	25.6	0.5	35.9
LARGE BUOY (seq B)	0.1	1.8	0.2	3.0
BOAT (seq. B)	0.2	0.6	0.1	1.1
YACHT (seq. B)	0.2	25.1	0.4	29.6

Table 9.4: Medians of errors (standard deviations) of state estimates for objects in evaluation sequences.

Object	velocity [knots]	standard deviation [knots]
Buoy (A)	2.0	0.5
Yacht (A)	2.0	1.6
Buoy (B)	8.3	3.3
Boat (B)	5.9	2.8

Table 9.5: Summary of velocity estimation in evaluation sequences.



(a) in image (b) in scene

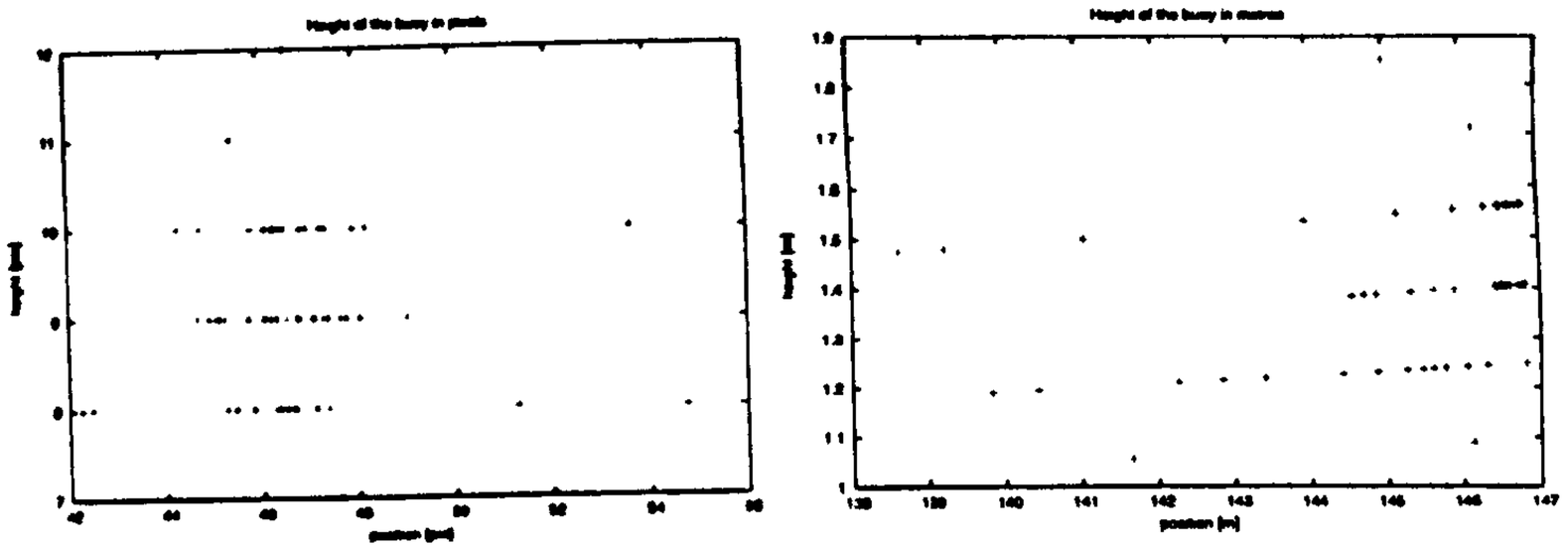
Figure 9.8: The estimated height of LARGE BUOY in sequence A.

while the buoy in the sequence B moves along the horizontal direction.

The binarised segments containing the buoys are listed in Figures B.7,B.8 in Appendix B. The heights of each buoy in the image and the scene are plotted in Figures 9.8, 9.9. Tables 9.6 and 9.7 contain average heights, standard deviations, extreme values and distance ranges for both buoys.

The nomenclature of U.S. Coast Guard (G-SEC-2, 2003) recognises different types of buoys according to their dimensions. The types are distinguished by labels such as '9X35' where the numbers represent the total width and height of the buoy in feet. These values can be considered as a ground truth assuming that there is no significant difference between the US and UK regulations. Doolin (2003) points out that only about one half of the height of the buoy is above the water.

The height of the buoy in the sequence A is estimated at 3.5 metres (11 feet 5 in) with 0.2 metres (7 in) standard deviation. This would correspond to one of 8X26 or 8x21 buoy types designed for open locations according to G-SEC-2 (2003). The height of the buoy in the sequence B is estimated at 1.4 metres (4 feet 7 in) with 0.2 metres (7 in) standard deviation. This would correspond to one of 7X15 or 5x11 buoy types that are designed for semi-exposed and protected locations. Both heights are estimated with the same standard deviation which confirms the consistency of the inverse mapping. These estimates are only approximations due to the absence of a ground truth information and due to approximation of camera calibration parameters.



(a) in image (b) in scene

Figure 9.9: The estimated height of LARGE BUOY in sequence B.

LARGE BUOY (A)	Height				Range	
	average	std. dev.	min.	max.	min.	max.
image [pix]	37.5	4.2	30	47	92.9	69.6
scene [m]	3.5	0.2	3.1	3.9	74.5	98.3

Table 9.6: Buoy tracking results for sequence A.

9.4 Summary

The system is evaluated in three categories: detection and tracking, motion estimation and inverse mapping. The evaluation in the first category shows the ability of the system to detect various objects in the scene regardless of their size and appearance. The system attempts to track each detected object by estimating it's motion parameters. The state of the object is assessed for a threat and an alarm is raised if the motion parameters fit the conditions for threatening object.

The results show that all objects in the scene are detected with no false negatives present in either of the scenes. The tracking of the objects is more than 90% of their presence in the scene. The remaining 10% of time without tracking corresponds to occlusions and drop-outs in corner detection in case of

LARGE BUOY (B)	Height				Range	
	average	std. dev.	min.	max.	min.	max.
image [pix]	9.1	1	7	12	92.3	54.8
scene [m]	1.4	0.2	1.1	1.9	138.6	146.8

Table 9.7: Buoy tracking results for sequence B.

small objects near the horizon. These values are consistent with those obtained from evaluation of the framework using the development sequences. The threat is established for all objects that match the conditions of a threat.

The evaluation in the second category investigates the error of the location and motion estimates. All objects are tracked with an error (standard deviation) in the image which is less than two pixels in either direction excluding the transient states at the beginning of the tracking. The errors of estimates remapped back to the scene are significantly larger for objects near horizon which is expected as the resolution of the camera is limited.

The estimated average speeds of some selected objects are about 2 knots for sequence A and 6 and 8 knots for the sequence B. These values can be considered in accordance with the real situation despite the absence of a ground truth for either sequence.

The last evaluation category tests the hypothesis that inverse mapping should produce consistent values that are independent of location and velocity of the objects. The hypothesis is tested using projections of channel marking buoys in the sequences. Each sequence contains a buoy that is tracked and its height in pixels is determined from binarisation of the corresponding segment. The heights of the buoys are estimated at 3.5 metres with relative 5% standard deviation in the sequence A and at 1.4 metres with 11% relative standard deviation in the sequence B. Both buoys can be associated with actual buoy classes specified in (G-SEC-2, 2003) nomenclature.

Chapter 10

Conclusions

10.1 Summary of Results

The work presented in this thesis has identified and addressed the problem of automated visual surveillance systems for the maritime transportation domain. The main objective has been to deliver a machine vision framework that analyses the maritime scene, identifies and tracks any objects in it, assesses their activities for specified events of interest and delivers the information to a human operator in a comprehensible form. The framework operates on monocular, monochrome, visible range video sequences captured by a camera monitoring the sea plane from several metres above.

The general contribution of the thesis consists in bringing the technology of conventional land-based surveillance and tracking systems to the maritime sector. Until now the surveillance task in maritime sector relied on the limited vigilance of the human operator. The proposed framework automates the surveillance process and offers new areas of vision applications. The potential integration of the framework with existing navigational aids can solve some of the persisting problems associated with the use of the radar. As a complementary technology it can provide additional information that can facilitate resolution of the navigational tasks.

In order to achieve that the following two primary contributions are delivered: the novel segmentation method based on a scene and object independent texture analysis of the individual sequence frames and the introduction of a new geometric feature corresponding to the submersion line

of an object in the scene that allows an unambiguous localisation of the object in the scene coordinates from a single calibrated camera view. Both methods are essential elements in the presented framework that consists of the following parts:

Texture-based Scene Segmentation Main emphasis of the research is on the segmentation part of the framework as there is a perceptible absence of robust segmentation methods that can specifically target the maritime scenes. Previously proposed segmentation methods of maritime scenes are of limited use due to their specialisation. Methods that operate on infra-red or airborne sequences can automatically detect only targets that are small, weak (Messer and Kittler, 2000; Messer et al., 1999) or have a specific colour (Sumimoto et al., 2000; Yamamoto et al., 1999). Other methods can detect moving targets only (Sanderson et al., 1997; Sanderson et al., 1999) or they rely on restrictive assumptions (Smith and Teal, 1999; Smith et al., 2003). The proposed segmentation algorithm addresses all these issues and provides a method of detection of a variety of static and moving objects at extended ranges in maritime scenes with varying environmental conditions.

The algorithm proposed in Chapter 4 is a scene and object independent segmentation algorithm that perceives the maritime scene as a planar texture under perspective projection where objects correspond to local variations in the texture. The structure of the segmentation grid adapts to the projection so as to minimise a bias in texture characterisation. By accounting for perspectiveness in the scene, the segmentation can detect objects at extensive ranges. The texture is described by a set of statistical moments commonly used for characterisation of natural textures.

The proposed thresholding method based on principles of non-parametric clustering enables to find objects in the scene without any prior knowledge of their appearance. No training is necessary and the thresholding scheme used for selection of objects in the scene is temporally adaptive. All values of parameters of the segmentation algorithm have been obtained either by evaluations using multiple sets of typical maritime scenes or derived empirically.

The cross-validation of the segmentation algorithm carried out in Section 4.9.5 shows that the relative number of objects being missed in the sequence is below 1%. The final cross-validation of the complete framework in Chapter 9 confirms the result as all objects in both evaluation scenes are detected.

Object Characterisation The proposed object characterisation using geometric features avoids segmentation of the object at pixel level that often assumes object's homogeneity and plain structure, thus restricting it's use. Instead, each detected object in the scene is assigned a set of geometric features that characterise it. Salient features such as corners prove to be a competent representation of the structure due to the man-made nature of majority of objects in maritime scenes.

A new geometric feature corresponding to a submersion line that is specific for objects in maritime environment is introduced in Section 5.2. The submersion line is a competent estimate of the position of the nearest point of an object with respect to the camera. This nearest point is convenient for a collision detection as it also represents the point of the possible impact. Any object detected in the scene is completely specified by the segment enclosing it, its submersion line and a set of corners detected within the segment. This allows to locate and track the object without the need to resolve it's exact 2D shape or 3D structure.

Object Tracking A feature-based motion tracking algorithm is described in Chapters 6 and 7. A search for inter-frame correspondence between corners based on the criterion of similarity between small intensity patches surrounding each corner provides an estimate for inter-frame displacement of the object. The Sum of Squared Differences in combination with median fusing provide the minimum error in displacement estimations. The error is below 0.5 pixels in both directions for the artificial and real evaluation sequences.

A simple linear predictor propagates the tracking in case of dropouts in corner detection, as suggested by Shapiro (1995). The predictor increases the coherence of the tracking by 12% for artificial evaluation sequence and by 5% for real evaluation sequence.

A sub-pixel localisation improves the precision of matches. A horizon tracking scheme is proposed to avoid systematic error in localisation of the matches caused by vertical oscillations due to a cross-wind impact on the imaging device.

Based on an assumption that motion of objects in maritime scenes is piecewise linear the estimation of motion parameters is done by a linear Kalman filtering. Possible occlusions are detected from the values of the measurement covariance matrix. The Kalman tracker is re-initialised in case the target

becomes occluded.

Inverse Mapping The information about objects and their activities obtained from the 2D image is related to the 3D structure of the monitored scene. The relation is characterised by a perspective projection between the image and sea planes. Two extrinsic camera calibration parameters are necessary to find the inverse transform: projection of the horizon and height of the camera above the sea surface. The first one is directly available from the image. The second one is obtained from an off-line camera calibration.

The re-mapped locations and velocities are tested against a set of heuristic rules specifying criteria of events of interest such as collision threat, intrusion detection, violation of traffic regulations, etc. Any object that meets the criteria is highlighted and a pre-defined response such as an alarm is triggered.

The output of the system is in a form of annotated image of the original scene with objects being marked, labelled and optionally highlighted. A radar-like chart of the scene with details of objects positions and velocities is generated as well. In such a way the result is comprehensible to a human operator.

10.2 Future Work

The framework presented in the thesis has been developed using image sequences acquired by a camera operating in visible light range. The applicability of the system is restricted to daytime hours. In addition, severe weather conditions such as heavy rain or thick fog would reduce the visibility and, consequently, restrict the operational range of the framework. Shifting the wavelength range of the imaging device towards infra-red would enable to deploy the framework in less favourable lighting conditions as the temperature of objects generally differs from the temperature of the sea. Even though this option hasn't been considered in the research the fact that the image segmentation is illumination independent suggests that there is a possibility that the system would operate on these sequences without necessarily re-designing the framework structure and derived algorithms.

There is a substantial increase in the tracking uncertainty for small and homogeneous objects located near the horizon. Not enough salient features are usually detected making the estimate of displacement unreliable. An

alternative displacement detection method such as registration of the whole segment containing the object would improve the tracking results.

The collision detection and estimation of the time to contact are done only with respect to fixed pre-defined zones in the scene. A natural extension is to estimate collision in between any detected objects in the scene. Such functionality would further extend the possible applications of the system.

The limitations of camera calibration should be addressed in the future research. The precision of location and velocity estimates is directly related to the precision of intrinsic camera parameters. The 'rule of thumb' for camera calibration states that the camera should be calibrated for the same depth of scene at which it will operate. Calibration of cameras for wide range imaging cannot adhere to this rule as any calibration target would be unrealistically large. An option is to extrapolate the calibration results. The relation between scene depth and calibration parameters, however, is not always predictable. A realistic approach to the calibration process for wide range imaging is to use a calibration target as large as practically manageable with camera lens focused to infinity. A representative statistical sample of the calibration parameters can be obtained by repeating the calibration process several times at different ranges and configurations of the target. Finally, the obtained values can be checked for consistency by, for example, a method using architectural features of buildings as proposed by van den Heuvel (1999).

Finally, the exploration of possible applications of the framework in real world scenarios is desirable. The specific demands of the maritime industry regarding the surveillance and navigation tasks should be identified and incorporated into the framework. In such a way an attractive solution that catches the attention of the maritime industry can be provided.

References

Ablavsky, V. (2003). Background Models For Tracking Objects in Water, In: *The Proceedings of The International Conference on Image Processing*, Vol. 2, pp. III – 125–8.

Achard, C., Bigorgne, E. and Devars, J. (2000). A sub-pixel and multispectral corner detector, In: *The Proceedings of The 15th International Conference on Pattern Recognition*, Vol. 3, Barcelona, 3-8 September, pp. 959–962.

Amiel, L. (2000). VTDS - Total Information System Integration, *Port Technology*.

Anderson, B. and Moore, J. (1979). *Optimal Filtering*, number 0-13-638122-7, Prentice-Hall.

Argenti, F., Alparone, L. and Benelli, G. (1990). Fast algorithms for texture analysis using co-occurrence matrices, *IEE Proceedings on Radar and Signal Processing* 137(F/6): 443–448.

Australian Transport Safety Bureau (2004). Ships and Fishing Vessels, Safety Bulletin.

Bakstein, H. (1999). A complete DLT-based Camera Calibration with a Virtual 3D Calibration Object.

Banerjee, S. (2002). Camera Models and Affine Multiple Views Geometry, Online Tutorial.

Barron, J., Fleet, D. and Beauchemin, S. (1994). Performance of optical flow techniques, *International Journal of Computer Vision* 12(1): 43–77.

Batlle, J., Casals, A., Freixenet, J. and Marti, J. (2000). A review on strategies for recognizing natural objects in colour images of outdoor scenes, *Image and Vision Computing* 18: 515–530.

Beauchemin, S. and Barron, J. (1995). The Computation of Optical Flow, *ACM Computing Surveys* 27(3): 433–467.

Belmont, M. and Morris, E. (1994). Adaptive measurement and signal processing strategies associated with deterministic sea wave prediction, In: *The Proceedings of the 6th International Conference on Electronic Engineering in Oceanography*, pp. 181 –188.

Black, J. and Ellis, T. (2002). Multi-camera image measurement and correspondence, *Measurement* (32): 61–71.

Borman, S., Robertson, M. A. and Stevenson, R. L. (1999). Block-Matching Sub-Pixel Motion Estimation from Noisy, Under-Sampled Frames - An Empirical Performance Evaluation, In: *The Proceedings of the SPIE Visual Communications and Image Processing '99*, Vol. 3653.

Bouguet, J.-Y. (2004). Camera Calibration Toolbox for Matlab.

Brand, P., Courtney, P., de Paoli, S. and Plancke, P. (1996). Performance Evaluation of Camera Calibration for Space Applications.

Broggi, A. (1995). Parallel and Local Feature Extraction: A real-Time Approach to Road Boundary Detection, *IEEE Transactions on Image Processing* 4(2): 217–223.

Broggi, A. and Berte, S. (1995). Vision-based Road Detection in Automotive Systems: A real-Time Expectation-Driven Approach, *Journal of Artificial Intelligence Research* 3: 325–348.

Buluswar, S. and Draper, B. (1994). Non-parametric Classfication of Pixels Under Varying Outdoor Illumination, In: *ARPA '94*, Vol. II, pp. 1619–1625.

Campbell, N. W. and Thomas, B. T. (1996). Segmentation of natural images using self organising feature maps, In: *British Machine Vision Conference Proceedings*, University of Edingburgh, Edingburgh, UK, September, pp. 223–232.

Canon (2004). What is Image Stabiliser.

Capitao, R. and de Carvalho, M. (2000). Flume simulation of water surface profiles using a general software package, In: *In the Proceedings of OCEANS 2000 MTS/IEEE Conference and Exhibition*, Vol. 2, pp. 927 –933.

Changming, S. (2002). Fast Stereo Matching Using Rectangular Subregioning and 3D maximum-Surface Techniques, *International Journal of Computer Vision* 47(1/2/3): 99–117.

Chen, J.-L. and Kundu, A. (1995). Unsupervised Texture Segmentation Using Multichannel Decomposition and Hidden Markov Models, *IEEE Transactions on Image Processing* 4(5): 603–619.

Chetverikov, D. and Haralick, R. M. (1995). Texture Anisotropy, Symmetry, Regularity: Recovering Structure and Orientation from Interaction Maps, In: *British Machine Vision Conference Proceedings*, The University of Birmingham, Birmingham, UK, pp. 57–66.

Chi-Min, L., Kuo-Guan, W. and Jer-Heh, S. (1994). A single Kalman filtering algorithm for maneuvering target tracking, In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. III/193 –III/196.

Cipolla, R., Drummond, T. and Robertson, D. (1999). Camera calibration from vanishing points in images of architectural scenes, In: *British Machine Vision Conference 1999 Proceedings*, Vol. 2, The University of Nottingham, Nottingham, UK, pp. 382–391.

Clarke, T. and Fryer, J. (1998). The Development of Camera Calibration Methods and Models, *Photogrammetric Record* 16(91): 51–66.

Cohen, I. and Medioni, G. (1998). Detecting and Tracking Moving Objects in Video from an Airborne Observer, In: *DARPA Image Understanding Workshop*.

Collins, R. T., Lipton, A. J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P. and Wixson, L. (2000). A system for Video Surveillance and Monitoring, *Technical Report CMU-RI-TR-00-12*, Carnegie Mellon University.

Cooper, J., Venkatesh, S. and Kitchen, L. (1993). Early jump-out corner detectors, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(8): 823 –828.

Cupillard, F., Bremond, F. and Thonnat, M. (2003). Behaviour Recognition for Individuals, Groups of People and Crowd, In: *IDSS 03 Symposium Digest*, IEE, pp. 7/1–7/5.

Dash, M. and Liu, H. (1997). Feature Selection for Classification, *Intelligent Data Analysis* 1(3): 131–156.

Dellaert, F. and Thorpe, C. (1997). Robust car tracking using Kalman filtering and Bayesian templates, *Proceedings of Conference on Intelligent Transportation Systems*.

Diani, M., Actito, N. and Corsini, G. (2003). New Background Clutter Subspace Basis Selection Criterion for Clutter Cancellation in Infra-Red Naval Surveillance Systems, In: *Proceedings of SPIE, Image and Signal Processing for Remote Sensing VIII*, Vol. 4885, pp. 452–459.

Dick, A. R. and Brooks, M. J. (2003). Issues in Automated Visual Surveillance, In: *Proceedings of International Conference on Digital Image Computing: techniques and Applications (DICTA 2003)*, Sydney, Australia, December.

DNV, D. N. V. (2004). Ship Classification Services.

Doolin, R. (2003). Let's Hear It for The Buoys, http://www.sailnet.com/collections/articles/index.cfm?articleid=doolin0010.

Doretto, G., Chiuso, A., Wu, Y. N. and Soatto, S. (2003). Dynamic Textures, *International Journal of Computer Vision* 2(51): 91–109.

Dowdy, S. and Wearden, S. (1991). *Statistics for Research*, 2nd edn, John Wiley and Sons, Inc.

Dungate, D., Theobald, R. and Nurse, F. (1999). Higher-order Kalman filter to support fast target tracking in a multi-function radar system, In: *Proccedings of IEE Colloquium on Target Tracking: Algorithms and Applications*, number 1999/090, 1999/215, pp. 14/1 –14/3.

Dunteman, G. H. (1989). *Principal Components Analysis*, Vol. 69 of *Quantitative Applications in the Social Sciences*, Sage Publications.

Elgammal, A., Duraiswami, R., Harwood, D. and Davis, L. S. (2002). Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance, In: *Proceedings of The IEEE*, Vol. 90, pp. 1151–1163.

Erturk, S. (2003). Digital Image Stabilization with Sub-Image Phase Correlation Based Global Motion Estimation, *IEEE Transactions on Consumer Electronics* 49(4): 1320–1325.

Fauzi, M. and Lewis, P. (2003). A fully Unsupervised Texture Segmentation Algorithm, In: *Proceeedings of British Machine Vision Conference*, Norwich, UK, pp. 519–528.

Fuentes, L. M. and Velastin, S. A. (2001). People tracking in surveillance applications, In: *Proceedings of 2nd IEEE International Workshop on PETS*, Kauai, Hawaii, USA, 9th December.

Furuno, M. E. (2004). NavNet Information Brochure, 9-52 Ashihara-cho, Nishinomiya City, Japan.

G-SEC-2, O. E. D. (2003). *Aids to Navigation - Technical Manual Change 5*, U.S. Coast Guard's Directive System, April.

Galvin, B., McCane, B. and Novins, K. (1999a). OSCAR: Object Segmentation using Correspondence and Relaxation, In: *Proceedings of the Second International Conference on 3-D Digital Imaging and Modeling (3DIM)*, Ottawa, Canada, October.

Galvin, B., McCane, B. and Novins, K. (1999b). Robust Feature Tracking, In: *Proceedings of the 5th International/National Conference on Digital Image Computing, Techniques and Applications*, Perth, Western Australia, pp. 232–236.

Galvin, B., McCane, B., Novins, K., Mason, D. and Mills, S. (1998a). Recovering Motion Fields: An Evaluation of Eight Optical Flow Algorithms, In: *Britich Machine Vision Conference Proceedings*, Vol. 1, BMVA, University of Southampton, Southapmton, UK, 14-17 September, pp. 195–204.

Galvin, B., Novins, K. and McCane, B. (1998b). On the Evaluation of Optical Flow Algorithms, In: *Proceedings of Fifth International Conference on Control, Automation, Robotics and Vision*, pp. 1563–1567.

Gleason, S. S., A., H. M. and Jatko, B. (1991). Subpixel measurement of image features based on paraboloid surface fit, *Proceedings of SPIE, Machine Vision Systems Integration in Industry* 1386: 135–144.

Haritaoglu, I., Harwood, D. and Davis, L. S. (2000). W4: Real-Time Surveillance of People and Their Activities, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8): 809–830.

Harris, C. and Stephens, M. (1988). A combined Corner And Edge Detector, In: *Proceedings of The 4th Alvey Conference*, pp. 147–151.

Hawkes, K. G. (2001). Don't Give Up the Ship, Security Management Online article.

Heikkila, J. and Silven, O. (1997). A four-step Camera Calibration Procedure with Implicit Image Correction, *Proceedings of Computer Vision and Pattern Recognition Conference* pp. 1106–1112.

Hitachi Denshi, L. (n.d.). *Progressive Scan Type Black and White Camera KP-F1*, Hitachi Denshi (U.K.) Ltd., 14 Garrick Industrial centre, Irwing Way, Hendon, London NW9 6AQ, UK.

Hitchcock, E. M., Warm, J. S., Matthews, G., Dember, W. N., Shear, P. K., Tripp, L. D., Mayleben, D. W. and Parasuraman, R. (2003). Automation cueing modulates cerebral blood flow and vigilance in a simulated air traffic control task, *Theoretical Issues in Ergonomics Science* 4(1-2): 89–112.

International Maritime Organization (2002). Piracy and Armed Robbery Against Ships. Guidance to shipowners and ship operators., Circulatory of International Maritime Organization.

International Maritime Organization (2004). Imo 904e (b) colreg consolidated edition.

International Maritime Organization (n.d.). Monthly Reports On Acts of Piracy and Armed Robbery Against Ships, http://www.imo.org/home.asp.

Irani, M., Rousso, B. and Peleg, S. (1994). Recovery of Ego-Motion Using Image Stabilization, *Proceedings of the Computer Vision and Pattern Recognition* pp. 454–460.

Iversen, G. R. and Norpoth, H. (1987). *Analysis of Variance*, Quantitative Applications in the Social Sciences, 2nd edn, Sage Publications.

Jain, A. K., Duin, R. P. W. and Mao, J. (2000). Statistical Pattern REcognition: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jain, R., Kasturi, R. and Schunk, B. G. (1995). *Machine vision*, MacGraw-Hill.

Jensen, A. and Silber, G. (2003). Large Whale Ship Strike Database, *Technical Report NMFS-OPR-*, U.S. Department of Commerce, NOAA.

Kastrinaki, V., Zervakis, M. and Kalaitzakis, K. (2003). A survey of video processing techniques for traffic applications, *Image and Vision Processing* (21): 359–381.

Khoral, R. I. (2003). Khoros Pro 2001 Version 3, http://www.khoral.com.

Kim, K., Powers, E., Ritz, C., Miksad, R. and Fischer, F. (1987). Modeling of the nonlinear drift oscillations of moored vessels subject to non-Gaussian random sea-wave excitation, *IEEE Journal of Oceanic Engineering* 12(4): 568 –575.

Kingsley, S. and Quegan, S. (1992). *Understanding Radar Systems*, McGraw-Hill.

Kjerstad, N. (2003). On the safety and training of High Speed Craft navigators along the coast of Norway, In: *Proceedings of The World Maritime Technology Conference*, number D12(R04), San Francisco, California, USa, 17-20 October.

Ko, S.-J., Lee, S.-H., Jeon, S.-W. and Kang, E.-S. (1999). Fast Digital Image Stabilizer Based on Gray-Coded Bit-Plane Matching, *IEEE Transactions on Consumer Electronics* 45(3): 598–603.

Lan, Z.-D. and Mohr, R. (1998). Direct linear sub-pixel correlation by incrporation of neighbor pixels' information and robust estimation of window transformation, *Machine Vision and Applications* 10: 256–268.

Laws, K. (1980). Textured Image Segmentation.

Lee, B. and Hedley, M. (2002). Background Estimation for Video Surveillance, In: *Proceedings of Image and Vision Computing Conference, New Zealand*, University of Auckland, Auckland, New Zealand, November.

Lewis, J. (1995). Fast Normalized Cross-Correlation, In: *Proceedings of The Vision Interface Conference*, Hotel Loews le Concorde, Quebec City, Quebec, Canada, 15-19 May, pp. 120–123.

Li, M. (1994). Camera Calibration of the KTH head-Eye System, *Technical report*, CVAP, Computational Vision and Active Perception Laboratory, KTH, Stockholm, Sweden. revised April, 1996.

Li, M. and Lavest, J. M. (1995). Some Aspects of Zoom-Lens Camera Calibration, *Technical report*, CVAP, Computational Vision and Active Perception Laboratory,KTH, Stockholm, Sweden.

Li, P., Zhang, T. and Ma, B. (2004). Unscented Kalman filter for visual curve tracking, *Image Vision Computing* 22: 157–164.

Lipton, A. J. (1999). Local Application of Optic Flow to Analyse Rigid versus Non-Rigid Motion, *Technical Report CMU-RI-TR-99-13*, The Robotics Institute, carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213.

Lipton, A. J., Fujiyoshi, H. and Patil, R. S. (1998). Moving Target Classification and Tracking from Real-Time Video, In: *Proceedings of the 1998 DARPA Image Understanding Workshop*, DARPA.

Lucchese, L. and Mitra, S. (2001). Color Image Segmentation: A state-of-the-Art Survey, *Proc. of the Indian National Science Academy (INSA-A)* 67(2): 207–221.

Mackworth, N. (1950). Reserarches in the Measurement of Human Performance, *Technical Report MRC spec. Report 268*, HMSO.

Magee, D. R. (2004). Tracking multiple vehicles using foreground, background and motion models, *Image and Vision Computing* 22: 143–155.

Marine Accident Investigation Branch (2004). Bridge Watchkeeping Safety Study.

Maritime Accident Investigation Branch (2002). Report on the Analysis of Fishing Vessel Accident Data 1992 to 2000.

Maritime Transport Comitee (2003). Security in Maritime Transport: Risk Factors and Economic Impact, *Technical report*, Directorate for Science, Technology and Industry Organisation for Economic Co-operation and Development.

Marr, D. (1982). *Vision*, W.H. Freeman and Company.

Martinkauppi, B. (2002). Face Colour Under Varying Illumination - Analysis and Applications.

Maybeck, P. S. (1979). *Stochastic Models, estimation, and control*, Vol. 1, Academic Press, New York, San Francisco, A subsidiary of Harcourt Brace Jovanovich, Publishers, chapter 1, pp. 1–16.

McCane, B. (1997). On the evaluation of image segmentation algorithms, In: *Proceedings of Digital Image Computing: Techniques and Applications*, Massey University, Albany Campus, Auckland, New Zealand, pp. 455–460.

McCane, B., Galvin, B. and Novins, K. (2002). Algorithmic Fusion for More Robust Feature Tracking, *International Journal of Computer Vision* 1(49): 79–89.

McLaughlin, M. P. (1999). *Regress+, A Compendium of Common Probability Distributions.*

Messer, K. and Kittler, J. (2000). Data and Decision Level Fusion of Temporal Information for Automatic Target Recognition, In: *British Machine Vision Conference Proceedings*, Vol. 2, University of Bristol, Bristol, UK, 11-14 September, pp. 506–515.

Messer, K., de Ridder, D. and Kittler, J. (1999). Adaptive Texture Representation Methods for Automatic Target Recognition, *British Machine Vision Conference Proceedings* 2: 443–452.

Micheloni, C. and Foresti, G.L. ans Snidaro, L. (2003). A cooperative Multicamera System for Video-Surveillance of Parking Lots, In: *IDSS 03 Symposium Digest*, IEE, pp. 5/1–5/5.

Mirmehdi, M., Palmer, P., Kittler, J. and Dabis, H. (1996). Complex Feedback Strategies for Hypothesis Generation and Verification, In: *British Machine Vision Conference Proceedings*, University of Edingburgh, Edingburgh, UK, September, pp. 123–132.

Mohr, R. and Triggs, B. (1996). Projective geometry for image analysis, A tutorial given at ISPRS, Vienna.

Morimoto, C. and Chellappa, R. (1996). Fast electronic digital image stabilization, In: *Proccedings of the 13th IEEE International Conference on Pattern Recognition*, Vol. 3, pp. 284–288.

Morris, T. (2004). *Computer Vision and Image Processing*, Palgrave Macmillan.

Murray, J. D. M. and van Ryper, W. (1994). *Encyclopedia of Graphics File Formats*, 1st edn, O'Reilly and Associates, Inc., 103 Morris Street, Suite A, Sebastopol, CA 95472, USA.

Narasimhan, S. G., Wang, C. and Nayar, S. K. (2002). All the Images of an Outdoor Scene, In: *Proccedings of The European Conference on Computer Vision '02*, pp. 148–162.

National Oceanic and Atmospheric Administration (n.d.). http://www.noaa.gov, website.

Nera GmbH (2004). Maritime Navigation Products.

Nickels, K. and Hutchinson, S. (2002). Estimating uncertainty in SSD-based feature tracking, *Image and Vision Computing* 20: 47–58.

Nielsen, M. and Petersen, J. (2001). Collision Avoidance at Sea - Practice and Problems, In: *Proceedings of 20th European Annual Conference on Human Decision Making and Manual Control*, Kongens Lyngby, Denmark, June 25-27, pp. 81–90.

Norland, R. and Loberg, A. E. (2001). A comparison of sea waves in open sea and coastal waters, In: *Proceedings of CIE International Conference on Radar*, pp. 423 –426.

Olague, G. and Hernandez, B. (2002). Flexible model-based multi-corner detector for accurate measurements and recognition, In: *Proceedings of The 16th International Conference on Pattern Recognition*, Vol. 2, pp. 578 –583.

Otsu, N. (1979). A threshold Selection Method from Gray-Level Histograms, *IEEE Transactions on Systems, Man, and Cybernetics* SMC-9(1): 69–66.

Pal, N. (1993). A review on Image Segmentation Techniques, *Pattern Recognition* 26(9): 1277–1294.

Pauwels, E. J. and Frederix, G. (1999). Cluster-Based Segmentation of Natural Scenes, *Proceedings of the 1999 7th IEEE International Conference on Computer Vision* 2: 997–1002.

Pauwels, E. J. and Frederix, G. (2000). Image Segmentation by Nonparametric Clustering Based on the Kolmogorov-Smirnov Distance, In: D. Vernon (editor), *Proceedings of the European Conference on Computer Vision*, Vol. 2, Springer Verlag, pp. 85–99.

Pettofrezzo, A. J. (1978). *Matrices and Transformations*, Dover Publications, Inc., New York.

Phinney, D. (1998). Advanced Tools for Waterway Pilotage, In: *OCEANS '98 Conference Proceedings*, Vol. 3, pp. 1839–1843.

Preetham, A., Shirley, P. and Smits, B. (1999). A practical Analytic Model of Daylight, In: *Proccedings of SIGGRAPH: 26th International Conference on Computer Graphics and Interactive Techniques*, Los Angeles, CA, USA, 8-13 August, pp. 91–100.

Premoze, S. and Ashikhmin, M. (2000). Rendering natural waters, In: *Proceeedings of The Eighth Pacific Conference on Computer Graphics and Applications*, pp. 423–434.

Quenot, G. M. (1996). Computation of Optical Flow Using Dynamic Programming, In: *Proceedings of IAPR Workshop on Machine Vision Applications*, Tokyo, Japan, November, pp. 249–252.

Raymarine Limited (2004). Maritime Navigational Products.

Raytheon Marine GmbH (2001). Maritime Navigation Commercial Products.

Reid, I. (2002). Applied Estimation I,II, tutorial.

Reilly, P. J., Klein, T. and Ilves, H. (1999). Design and Demonstration of an Infrared Passive Ranger, *John Hopkins APL Technical Digest* 20(2): 220–235.

Remagnino, P., Baumberg, A., Grove, T., Hogg, D., Tan, T., Worrall, A. and Baker, K. (1997). An Integrated Traffic and Pedestrian Model-Based Vision System, In: *British Machine Vision Conference Proccedings*, Vol. 2, University of Essex, Colchester, UK, 8-11 September, pp. 380–389.

RMA Electronics Inc. (2005). Calculating lens focal length, http://www.rmassa.com/lenses.htm.

Rosin, P. L. and Ellis, T. (1995). Image difference threshold strategies and shadow detection, In: *Proceedings of British Machine Vision Conference*, Vol. 1, University of Birmingham, Birmingham, UK, 11-14 September, pp. 347–356.

Ruzon, M. A. and Tomasi, C. (1999). Color Edge Detection with the Compass Operator, In: *Proceeedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 160–166.

Sanderson, J., Teal, M. and Ellis, T. (1997). Identification and Tracking in Maritime Scenes, In: *Proccedings of IEE Internaltional Conference on Image Processing and its applications*, Vol. 2, pp. 463–467.

Sanderson, J., Teal, M. and Ellis, T. (1999). Characterisation of a Complex Maritime Scene using Fourier Space Analysis to Identify Small Craft, In: *Proceedings of the 7th IEE International Conference on Image Processing and its Applications*, Vol. 2, pp. 803–807.

Sara, R. (1999). The Class of Stable Matchings for Computational Stereo, *Technical Report CTU-CMP-1999-22*, Center for Machine Perception, Czech Technical University, Prague, Czech Republic.

Sato, Y. and Ishii, H. (1998). Study of a collision-avoidance system for ships, *Contronl Engineering Practice* (6): 1141–1149.

Schalkoff, R. (1992). *Pattern Recognition, statistical, structural and neural approaches*, John Wiley and Sons, Inc.

Schmid, C., Mohr, R. and Bauckhage, C. (1998). Evaluation of Interest Point Detectors, *Technical report*, INRIA Rhone-Alpes, 655 av. de l'Europe 38330 Montbonnot, France.

Scott, D. (1979). On optimal and data-based histograms, *Biometrika* 66(3): 605–610.

Sezgin, M. and Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation, *Journal of Electronic Imaging* 13(1): 146–165.

Shafique, K. and Shah, M. (2004). Estimation of the Radiometric Response Functions of a Color Camera from Differently Illuminated Images, In: *Proceedings of The IEEE International Conference on Image Processing (ICIP'04)*, Singapore, October.

Shapiro, L. S. (1995). *Affine analysis of image sequences*, Cambridge University Press.

Sheikh, Y., Zhai, Y., Shafique, K. and Shah, M. (2004). Visual Monitoring of Railroad Grade Crossing, In: *Proceedings of SPIE Conferences*, Orlando, USA, April 13-15.

Shen, F. and Wang, H. (2001). A local edge detector used for finding corners, In: *Proceedings of International Conference on Information, Communications and Signal Processing*, Singapore, December.

Sheng, F. and Wang, H. (2000). Real Time Gray Level Corner Detector, In: *The Sixth International Conference on Control, Automation, Robotics and Vision Proceedings*.

Shi, J. and Tomasi, C. (1994). Good Features To Track, In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, June.

Skarbek, W. and Koschan, A. (1994). Colour Image Segmentation - A survey, *Technical Report 94-32*, Institute of Computer Science, Polish Academy of Sciences and Institute for Technical Informatics, Technical University of Berlin.

Slater, I. (1989). Practical procedures in the operation of vessel traffic systems, In: *IEE Colloquium on Marine Control, Communications and Safety*, IEE, pp. 5/1–5/6.

Smith, A. and Teal, M. (1999). Identification and Tracking of Maritime Objects in Near-Infrared Image Sequences for Collision Avoidance, In: *Proceedings of the IEE 7th International Conference on Image Processing and its applications*, Vol. 1, pp. 250–254.

Smith, A., Teal, M. and Voles, P. (2003). The Statistical Characterization of The SEa For The Segmentation of Maritime Scenes, In: *Proceedings of 4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications 2003*, Vol. 2, pp. 489–494.

Smith, P., Sinclair, B., Cipolla, R. and Wood, K. (1998). Effective Corner Matching, In: *British Machine Vision Conference Proceedings*, University of Southampton, Southampton, UK, pp. 545–556.

Smith, S. (1992). A new class of corner finder, *Proceedings 3rd British Machine Vision Conference* pp. 139–148.

Smith, S. (1998). ASSET-2: Real-time motion segmentation and object tracking, *Journal of Real Time Imaging* 4(1): 21–40.

Smith, S. and Brady, J. (1995). SUSAN - A new Approach to Low Level Image Processing, *Technical Report TR95SMS1c*, Oxford University.

Spencer, L. and Shah, M. (2004). Water Video Analysis, In: *accepted for presentation at IEEE International Conference on Image Processing*, Singapore, 24-27 October.

Stein, G. P. (1993). Internal Camera Calibration using Rotation and Geoemtric Shapes.

Stein, G. P. (1997). Lens Distortion Calibration Using Point Correspondences, In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, PR, USA, 17-19 June, pp. 602–608.

Strat, T. (1993). Employing Contextual Information in Computer Vision, In: *DARPA93*, pp. 217–229.

Sumimoto, T., Kuramoto, K., Kuramoto, Okada, S., Miyauchi, H., Imade, M., Yamamoto, H. and Arvelyna, Y. (2000). Detection of a Particular Obejt from environmental Images under Various Conditions, In: *Proceedings of the 2000 IEEE International Symposium on Industrial Electronics, 2000.*, Vol. 2, pp. 590 – 595.

Tai, J.-C., Tseng, S.-T., Lin, C.-P. and Song, K.-T. (2004). Real-time image tracking for automatic traffic monitoring and enforcement applications, *Image Vision and Computing* 22: 485–501.

Tethys Research Institute (2004). Photographic evidence of fin whale (Balaenoptera physalus) collisions with ships in the Ligurian Sea Cetacean Sanctuary (1990-2000).

Thacker, N. and Cootes, T. (1996). Vision Through Optimization, British Machine Vision Conference Tutorial Notes.

The Current Sales Corp. (2004). Night Navigator 8520, 2933 Murray Street, Port Moody, BC Canada, V3H 1X3.

The Institute of Applied Anthropology (2001). Lifeguard Vigilance - Bibliographic Study, *Technical report*, The Institute of Applied Anthropology, Paris, 45, rue de Saints-Peres 75270 PARIS Cedex 06.

Toet, A. (2002). Detection of dim point targets in cluttered maritime backgrounds through multisensor image fusion, *Proceedings of SPIE, Targets and Backgrounds VIII: Characterisation and Representation* 4718: 118–129.

Tornieri, C., Bremond, F. and Thonnat, M. (2003). Updating of the reference image for visual surveillance systems, In: *IEE Proceedings of the IDSS Symposium - Intelligent Distributed Surveillance Systems*.

Torr, P. (1998). Geometric Motion Segmentation and Model Selection, *Transactions of the Royal Society A* (356(1740)): 1321–1340.

Torr, P. and Murray, D. (1993). Outlier Detection and Motion Segmentation, In: *Proceedings of the SPIE Sensor Fusion Conference VI*, pp. 432–443.

Trajkovic, M. and Hedley, M. (1998a). Fast Corner Detection, *Image and Vision Computing* 16(2): 75–87.

Trajkovic, M. and Hedley, M. (1998b). Fast Corner Detection, *Image and Vision Computing* 16: 75–87.

Tsai, R. Y. (1987). A versatile Camera Calibration Technique for High-Accuracy 3D machine Vision Metrology Using Off-the-Shelf TV cameras and Lenses, *IEEE Journal of Robotics and Automation* RA-3(4): 323–344.

Tsin, Y., Visvanathan, R. and Kanade, T. (2001). Statistical Calibration of CCD imaging Process, In: *Proceedings of The IEEE International Conference on Computer Vision (ICCV'01)*.

Turn Ltd. (2001). Sea Lynx, Marine Night Vision System.

van den Heuvel, F. (1998). Vanishing point detection for architectural photogrammetry, In: *International Archives of Photogrammetry and Remote Sensing*, Vol. XXXII, pp. 652–659.

van den Heuvel, F. (1999). Estimation of interior orientation parameters from constraints on line measurements in a single image, Vol. 32.

Vector Developments Ltd. (2004). Seenite Stabilised Night Vision Systems.

Vistar Night Vision Limited (2004a). IM223 Marine Night Vision System, Vistar Night Vision Limited, 24 Doman Road, Camberley, Surrey GU15 3DF, United Kingdom.

Vistar Night Vision Limited (2004b). Vistar 350 Night Vision System, 24 Doman Road, Camberley, Surrey, GU15 3DF, United Kingdom.

Vistar Night Vision Limited (2004c). Vistar IM 405 Multi-Sensor Surveillance System, 24 Doman Road, Camberley, Surrey, GU15 3DF, United Kingdom.

Walker, R. F., Jackaway, P. and Longstaff, I. (1995). Improving Co-occurrence Matrix Feature Discrimination, In: *DICTA 95*, pp. 643–648.

Wan-Ching, C. and Rockett, P. (1997). Bayesian labelling of corners using a grey-level corner image model, In: *Proceedings of The International Conference on Image Processing*, Vol. 1, pp. 687–690.

Wang, Z., Lu, L. and Bovik, A. C. (2004). Video Quality Assessment Based on Structural Distortion Measurement, *Signal Processing: Image Communication* 19(1): 1–9.

Welch, G. and Bishop, G. (2001). An Introduction to the Kalman Filter.

White, B. and Wydajewski, K. (2002). Commercial Ship Self-Defense Against Piracy and Maritime Terrorism, *Proceedings of Oceans MTS/IEEE Conference.*

Williamson, T. A. (1998). A high-Performance Stereo Vision System for Obstacle Detection.

Withagen, P., Schutte, K., Vossepoel, A. and Breuers, M. (1999). Automatic classification of ships from infrared (FLIR) images, In: *Proceedings of SPIE, Signal Processing, Sensor Fusion and Target Recognition VIII*, Vol. 3720, Orlando, USA, April, pp. 180–187.

Wolberg, G. (1990). *Digital Image Warping*, IEEE Computer Society Press, Los Alamitos, California, USA.

World Health Organization (1999). Electromagnetic fields and public health: radars and human health, Factsheet No. 226.

Worral, A., Ferryman, J., Sullivan, G. and Baker, K. (1995). A generic deformable model for vehicle recognition, In: *British Machine Vision Conference Proceedings*, University of Birmingham, Birmingham, UK, 11-14 September, pp. 127–136.

Worrall, A., Sullivan, G. and Baker, K. (1994). A simple, intuitive camera calibration tool for natural images, *British Machine Vision Conference Proceedings* pp. 75–76.

Yamamoto, K., Yamada, K., Kiriya, N. and Matsukura, H. (1999). Optical Sensing and Image Processing to Detect a Life Raft, In: *IGARSS'99 Proceedings*, Vol. 1, pp. 4467–469.

Ying, C. and Lawrence, P. (1995). Detecting scale-space consistent corners based on corner attributes, In: *Proceedings IEEE International Conference on Systems, Man and Cybernetics*, Vol. 4, pp. 3549 –3554.

Young, A. T. (2003,2004). Distance to the horizon.

Yusoff, Y., Christmas, W. and Kittler, J. (2000). Video Shot Cut Detection Using Adaptive Thresholding, In: *British Machine Vision Proceedings*, University of Bristol, Bristol, UK, 11-14 September.

Zhang, D. and Lu, G. (2001). Segmentation of Moving Objects in Image Sequence: A review, *Circuits, Systems, and Signal Processing* 20(2): 143–183.

Zhang, Z. (1998, 1999). A flexible New Technique for Camera Calibration, *Technical Report MSR-TR-98-71*, Microsoft Research, Microsoft Corporation One Microsoft Way Redmond, WA 98052.

Appendix A

Optical Flows

An evaluation of optical flow estimation algorithms is presented here. The algorithms evaluated are those discussed and implemented by Barron et al. (1994). All methods require user-defined parameters or thresholds. The parameters used for the evaluation are those suggested as defaults by Barron et al. (1994). The images of the optical flow are generated by sub-sampling the resulting motion field by factor of four and all motion vectors are magnified by a factor of two for clarity.

The results for the following methods are presented:

- ANANDAN - Anandan's method based on region matching
- HORN - Horn's and Schunck's method based on first derivatives of the image
- LUCAS - Lucas' and Kanade's method based on first derivatives of the image
- MB_LUCAS - Lucas' and Kanade's modified method
- NAGEL - Nagel's method based on second derivatives of the image
- QUENOT - Quenot's method based on linear programming
- SINGH - Singh's method based on region matching
- URAS - Uras' method based on second derivatives of the image

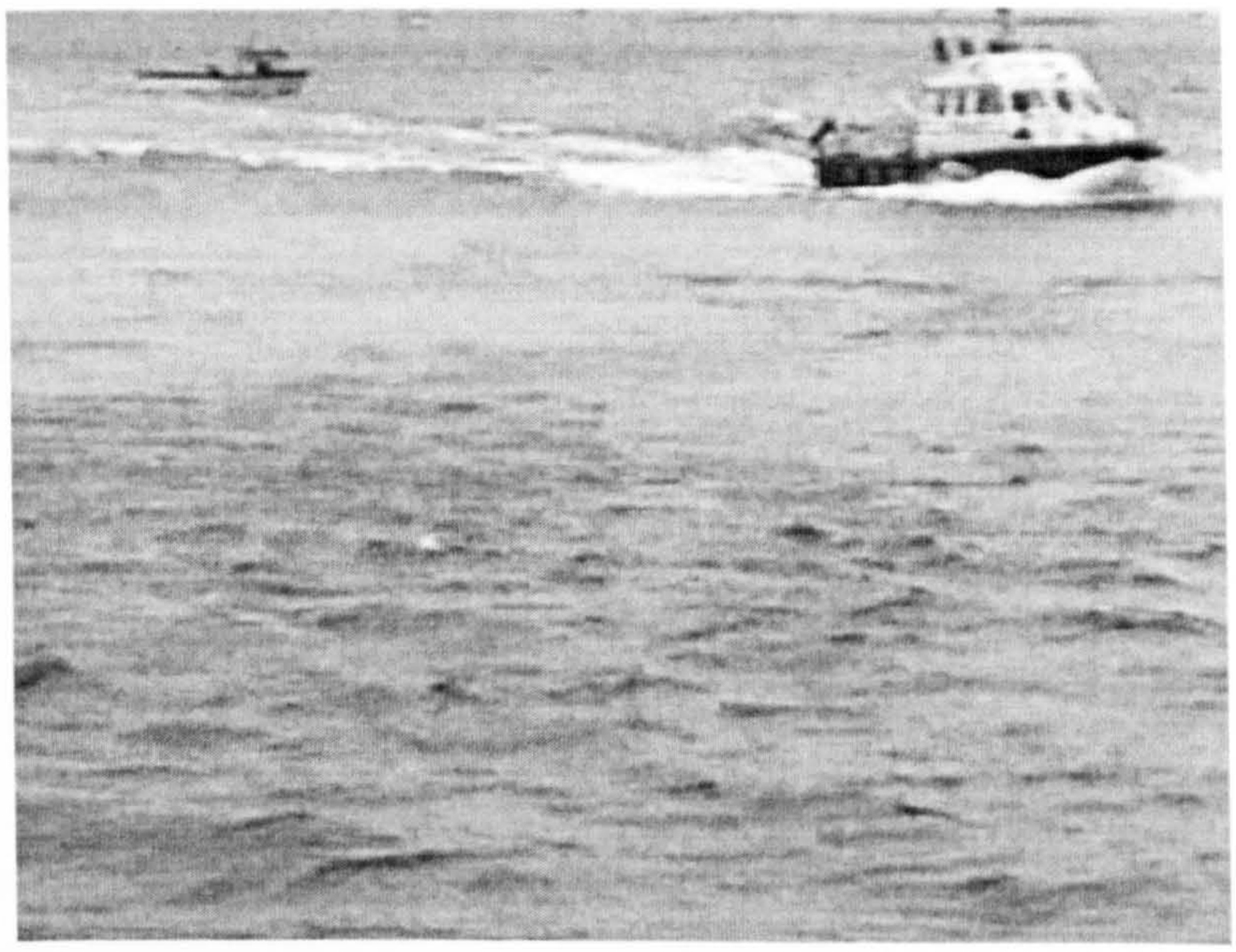
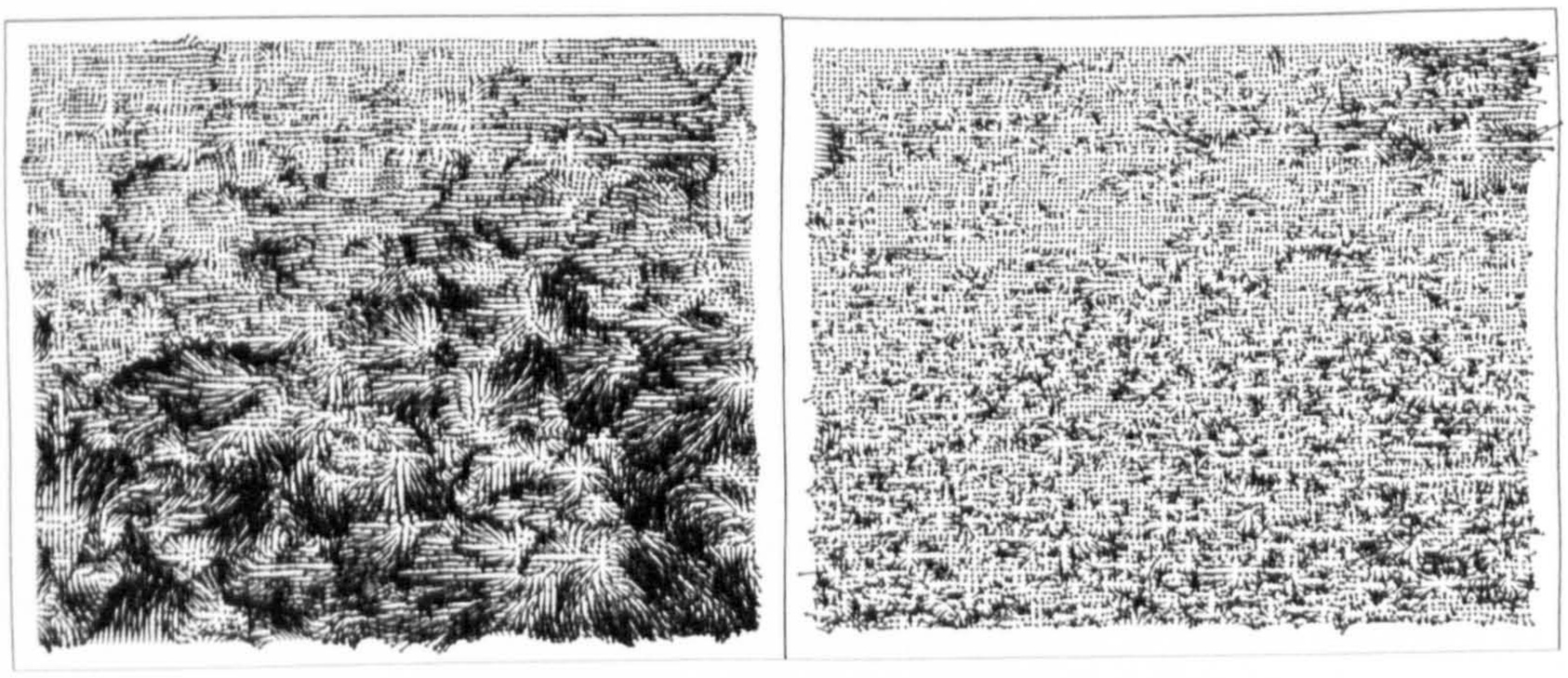


Figure A.1: The original sequence

The testing sequence contains two objects of different appearances moving at different speeds in similar directions. The frame rate of the sequence is 12.5 frames per second.

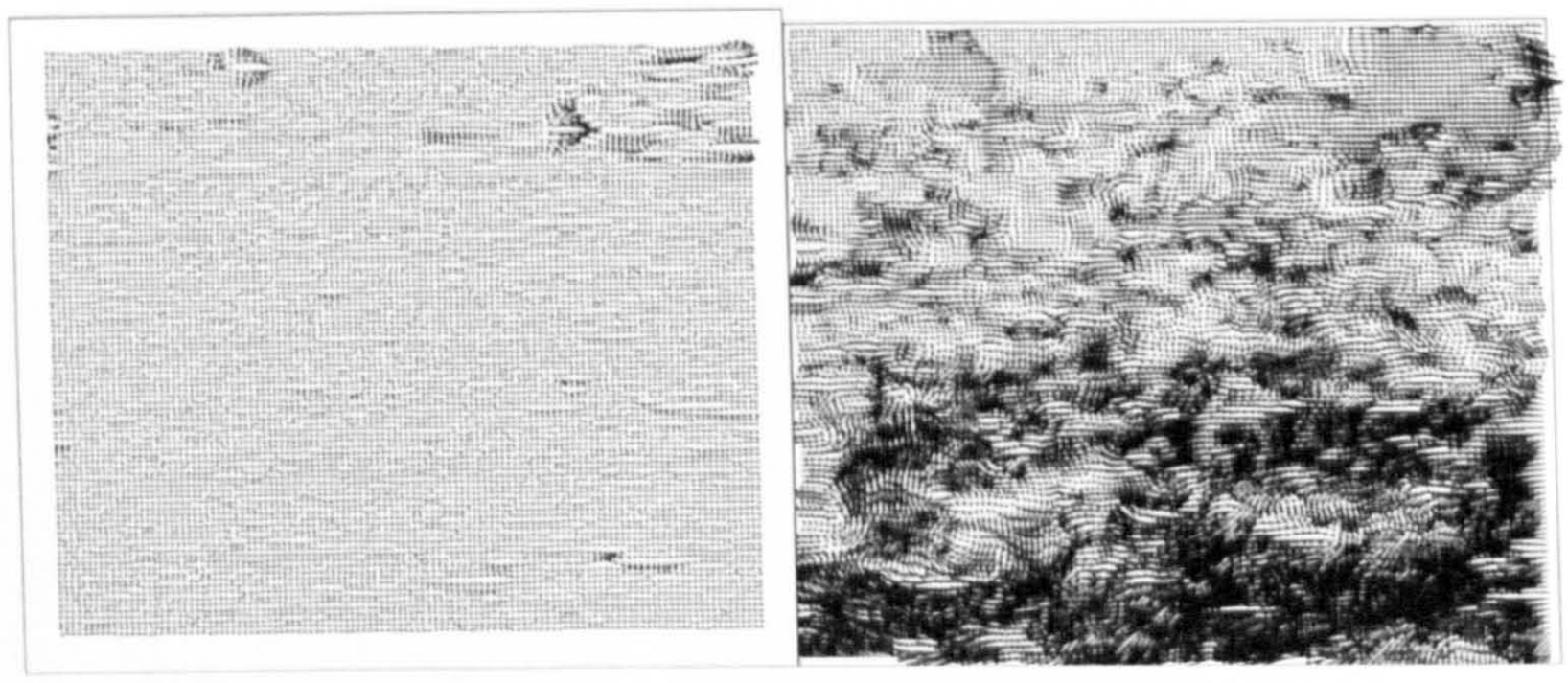


(a) ANANDAN

(b) HORN

(c) MB_LUCAS

(d) LUCAS



(e) NAGEL (f) QUENOT

(g) SINGH (h) URAS

Figure A.2: Optical flow algorithms - the results

Appendix B

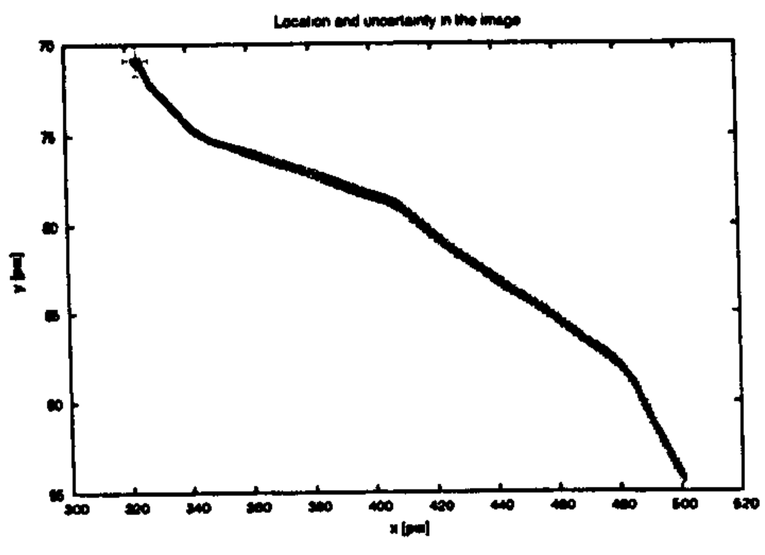
Cross-validation

B.1 Motion Estimation

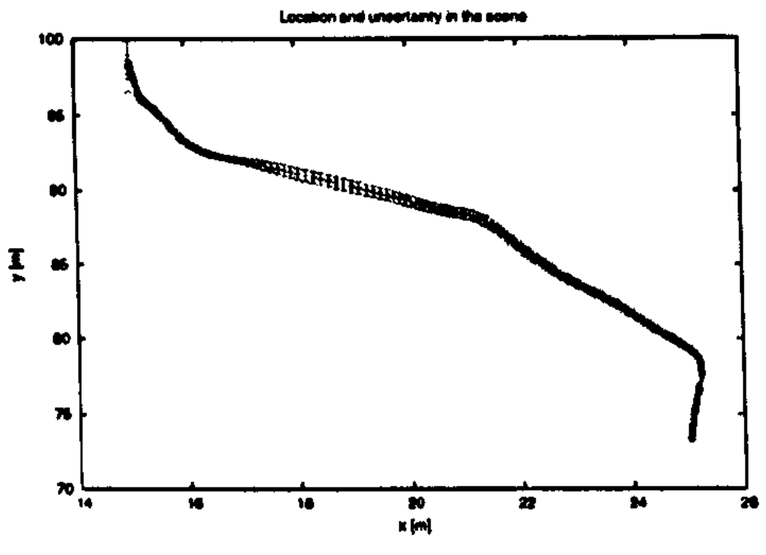
The plots show the results of Kalman filtering applied to objects detected in evaluation sequences. There are four plots for each object. The first two plots show the estimated locations and uncertainties (red crosses) in both the image and scene coordinates. The other two plots show velocity vectors placed at corresponding locations. Every tenth vector is displayed for clarity.

B.2 Inverse Mapping

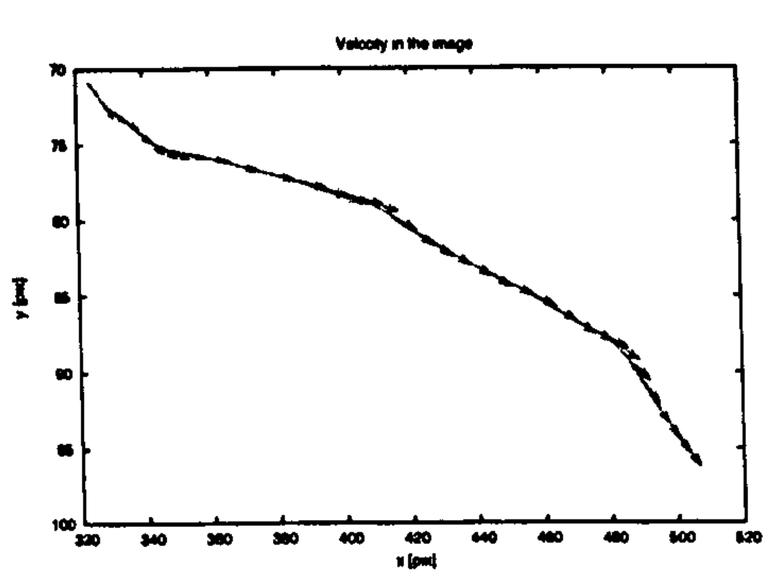
The lists of binarised segments containing buoys used in the evaluation of inverse mapping are presented here.



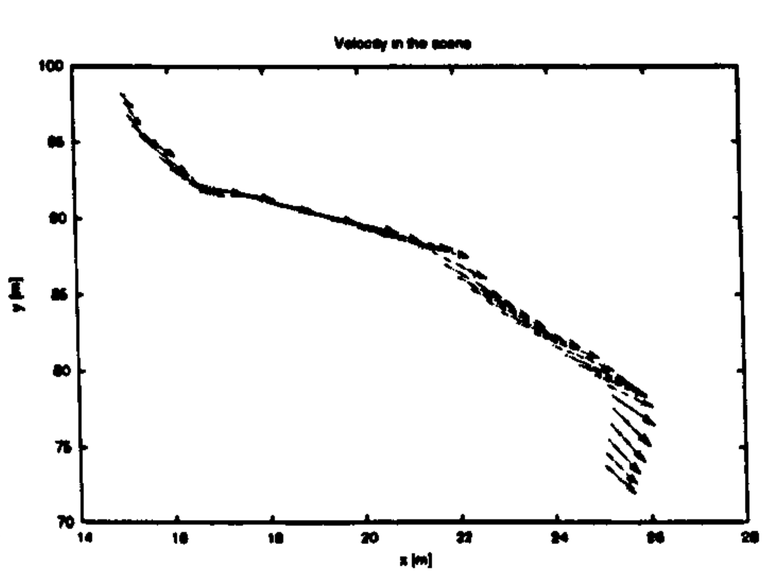
(a) Estimates of location in image



(b) Estimates of location in scene

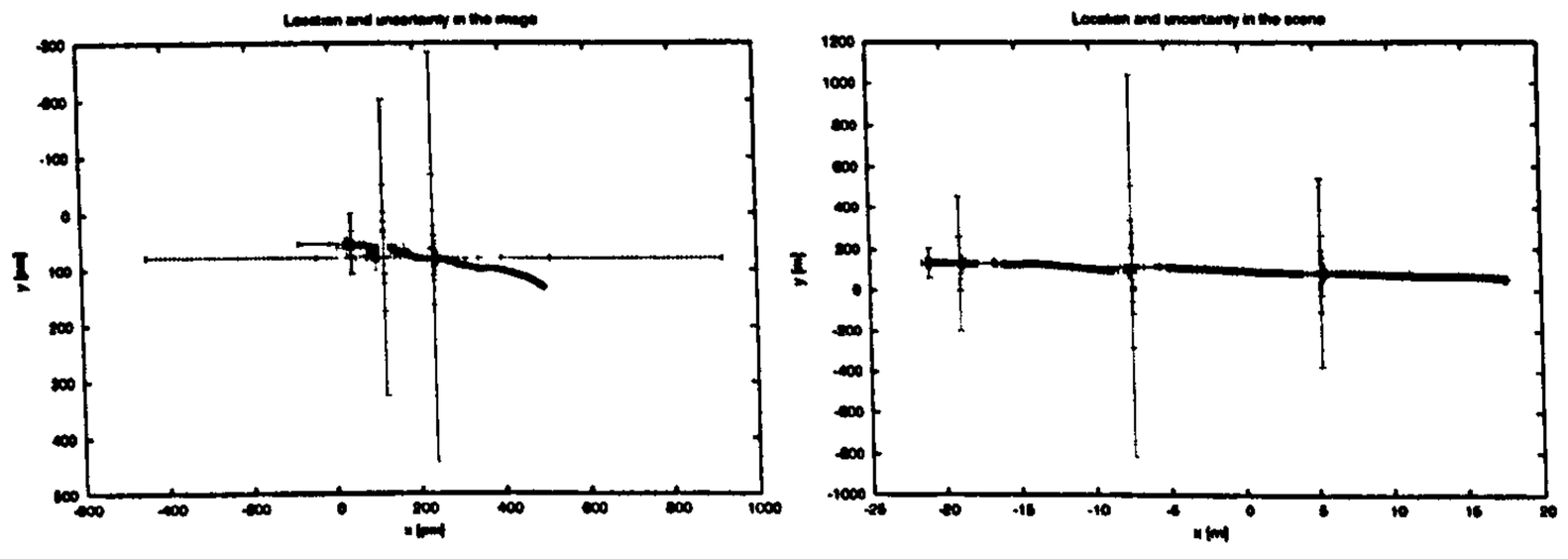


(c) Estimates of velocity in the image



(d) Estimates of velocity in scene

Figure B.1: Estimates of location and velocity in the image and the scene for the LARGE BUOY object in the sequence A.



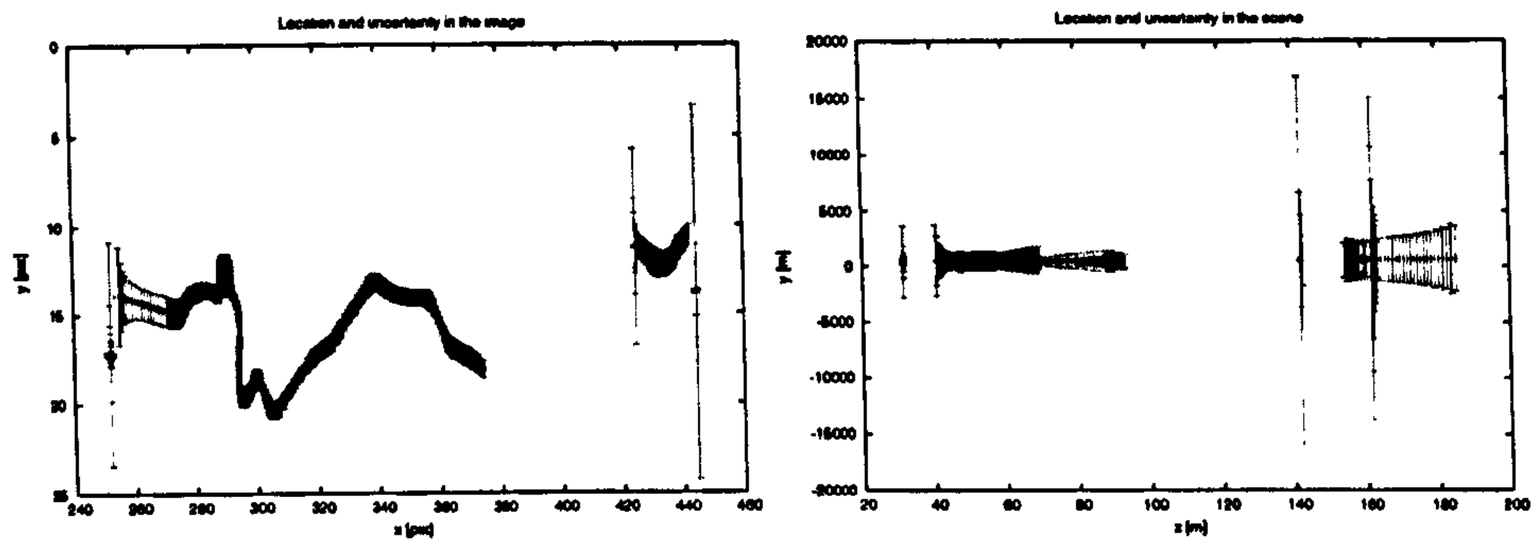
(a) Estimates of location in image

(b) Estimates of location in scene

(c) Estimates of velocity in the image

(d) Estimates of velocity in scene

Figure B.2: Estimates of location and velocity in the image and the scene for the YACHT object in the sequence A.



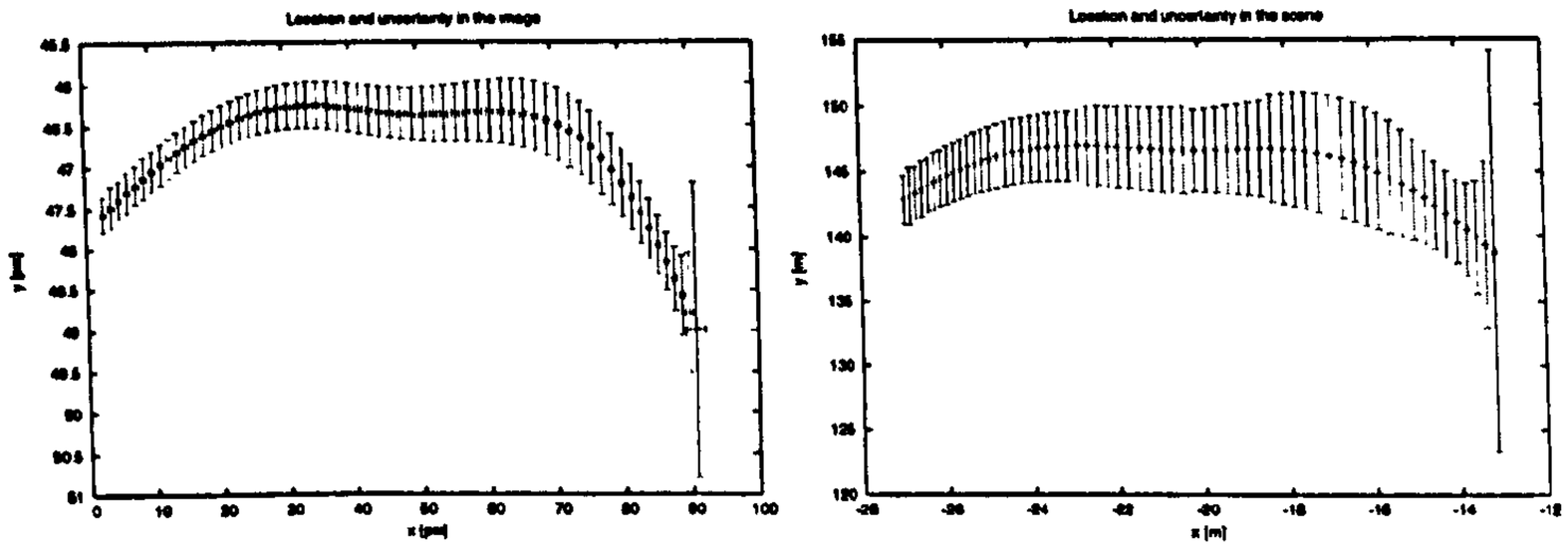
(a) Estimates of location in image

(b) Estimates of location in scene

(c) Estimates of velocity in the image

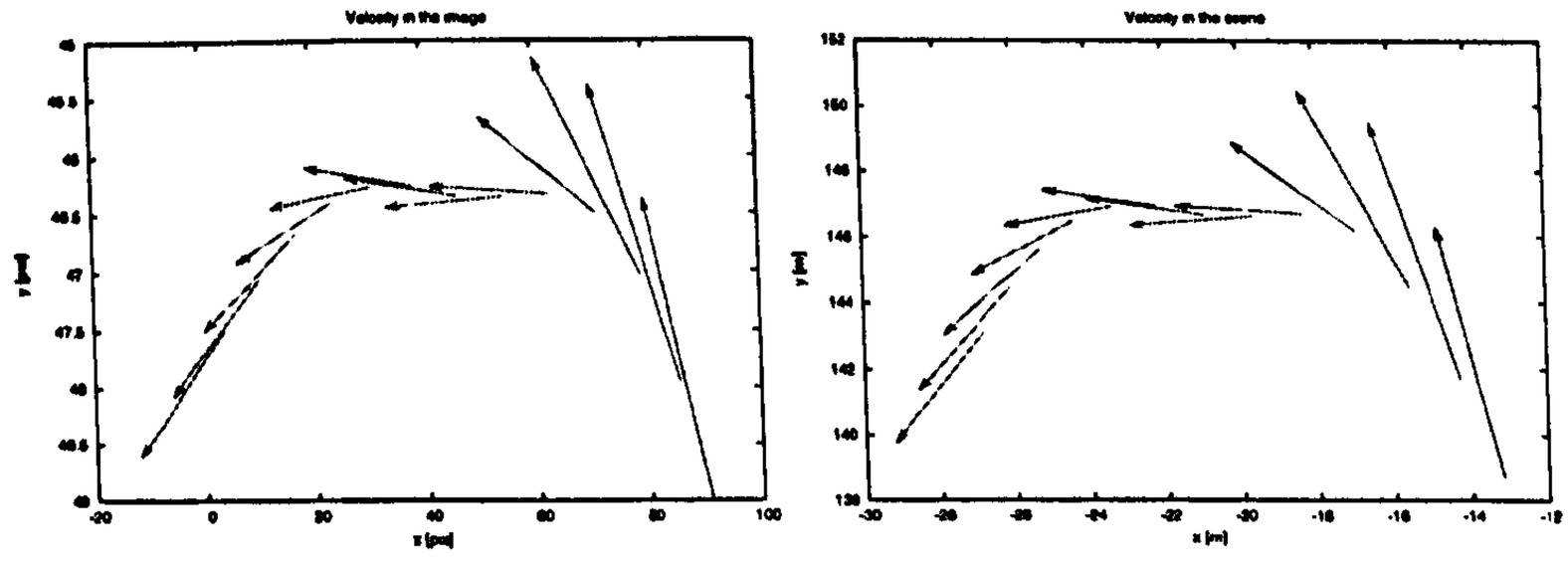
(d) Estimates of velocity in scene

Figure B.3: Estimates of location and velocity in the image and the scene for the SMALL BUOY object in the sequence A.



(a) Estimates of location in image

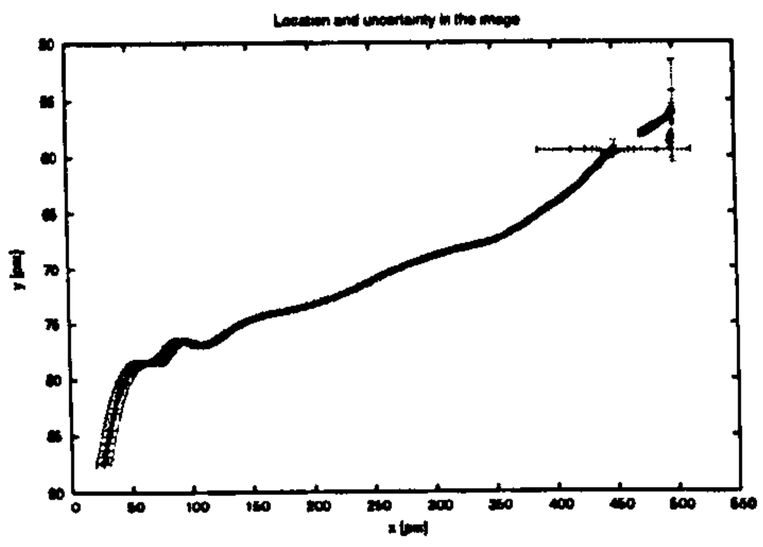
(b) Estimates of location in scene



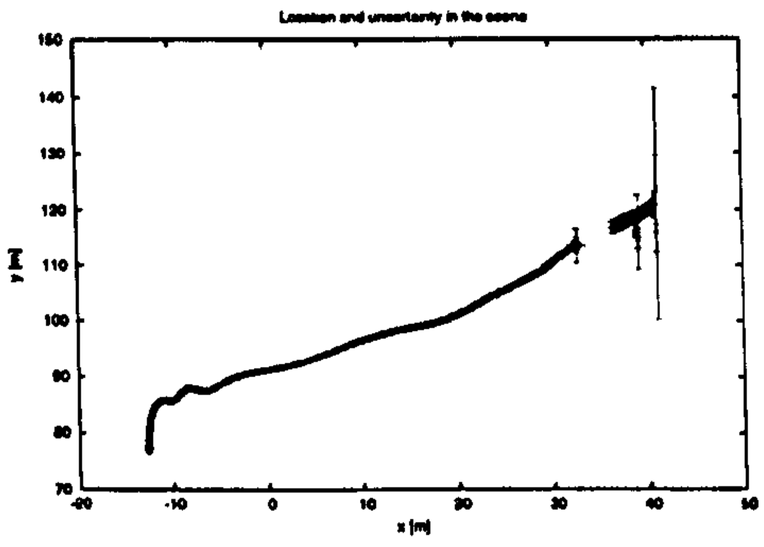
(c) Estimates of velocity in the image

(d) Estimates of velocity in scene

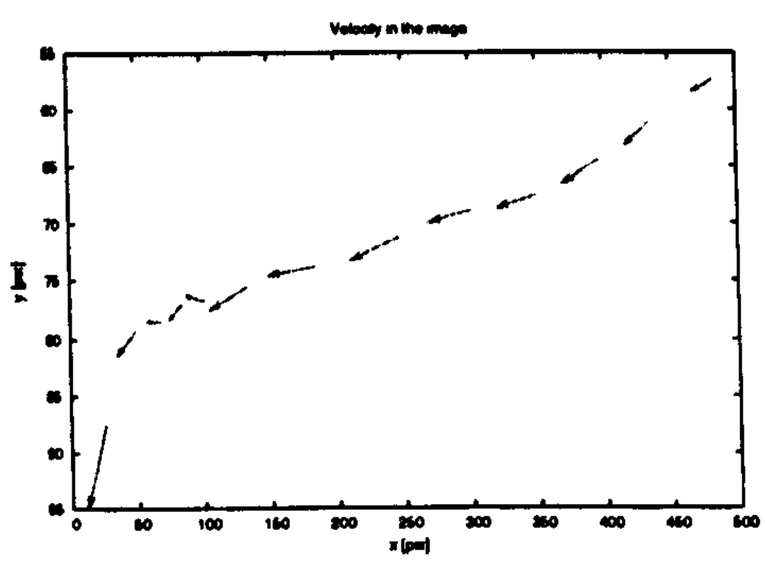
Figure B.4: Estimates of location and velocity in the image and the scene for the LARGE BUOY object in the sequence B.



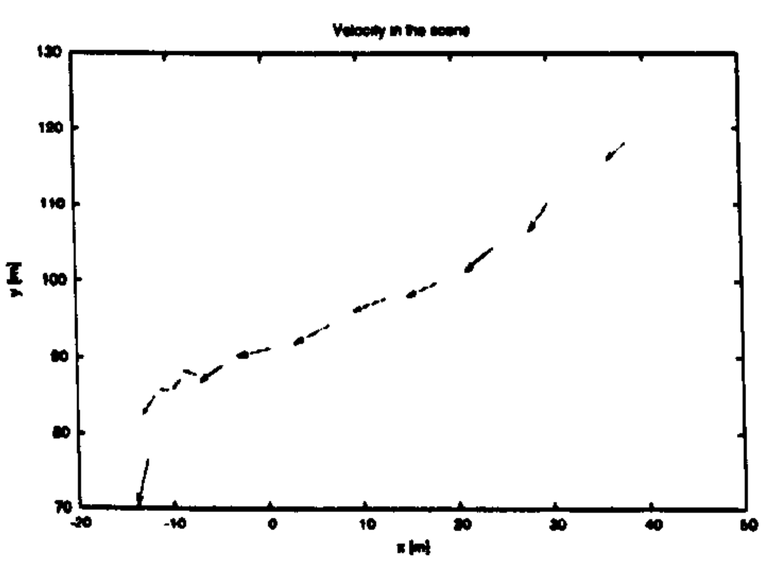
(a) Estimates of location in image



(b) Estimates of location in scene

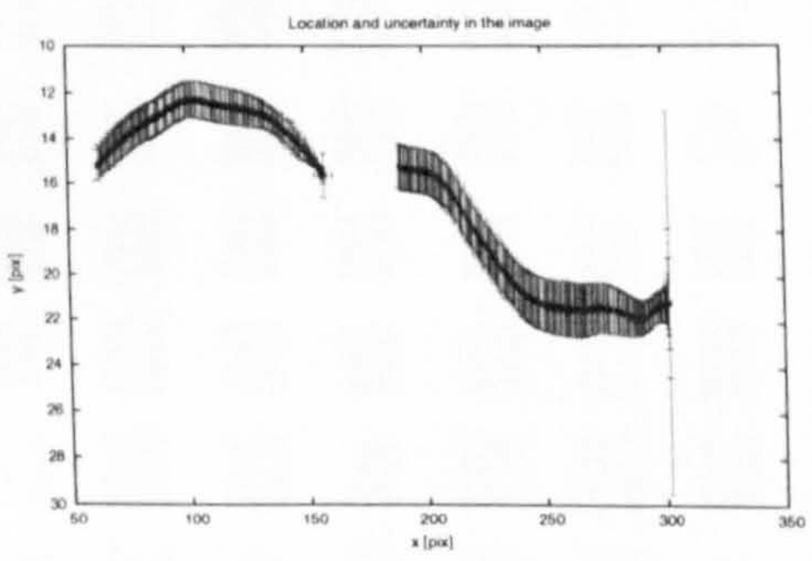


(c) Estimates of velocity in the image

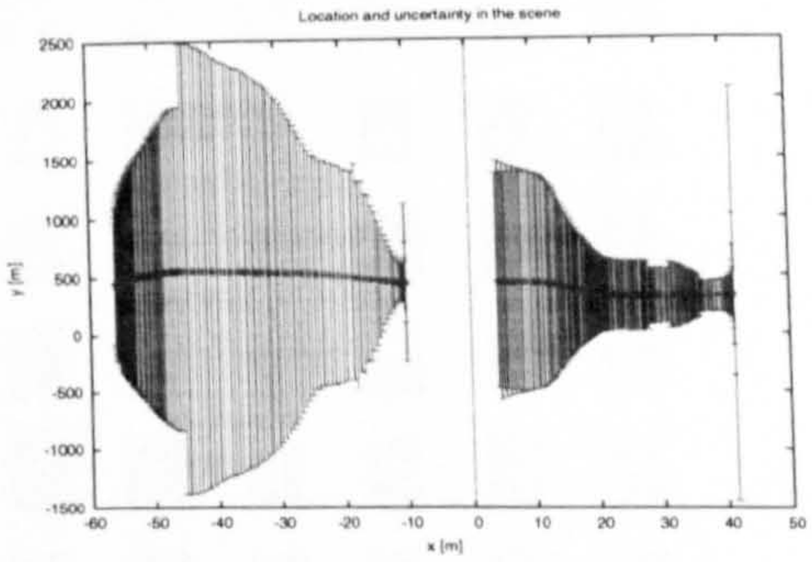


(d) Estimates of velocity in scene

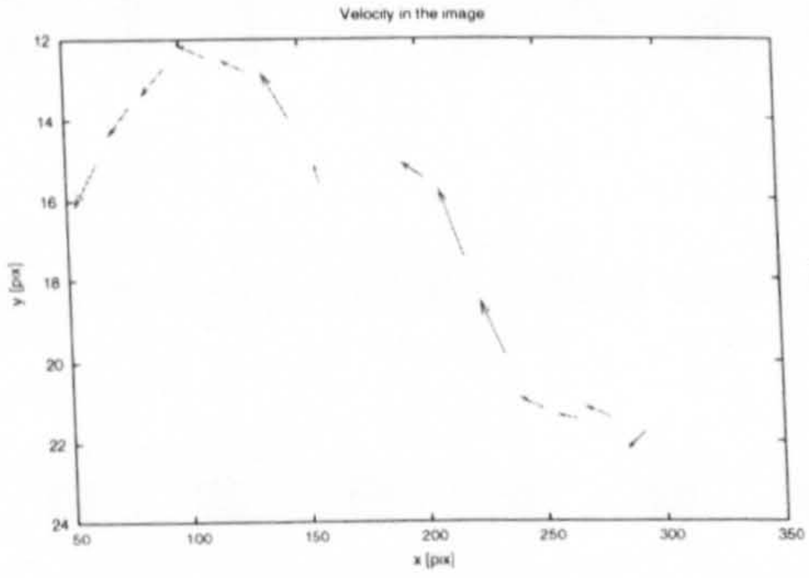
Figure B.5: Estimates of location and velocity in the image and the scene for the BOAT object in the sequence B.



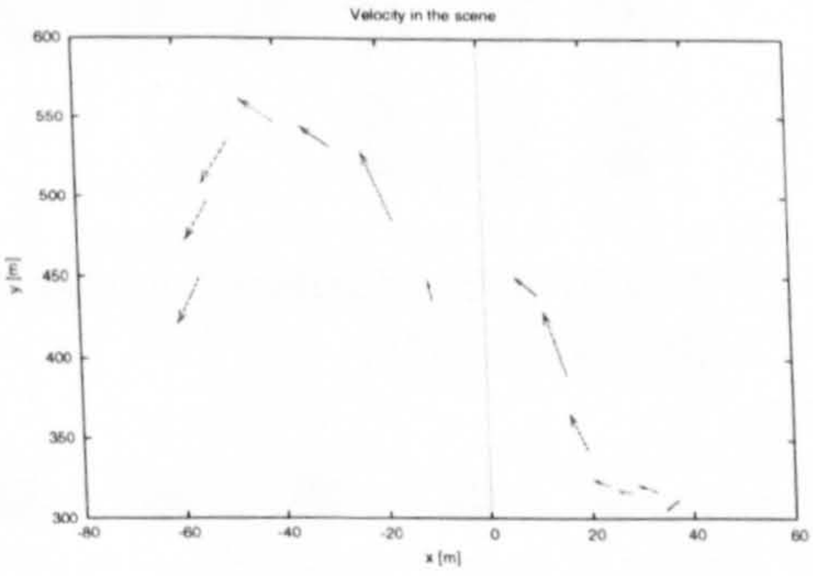
(a) Estimates of location in image



(b) Estimates of location in scene



(c) Estimates of velocity in the image



(d) Estimates of velocity in scene

Figure B.6: Estimates of location and velocity in the image and the scene for the YACHT object in the sequence B.

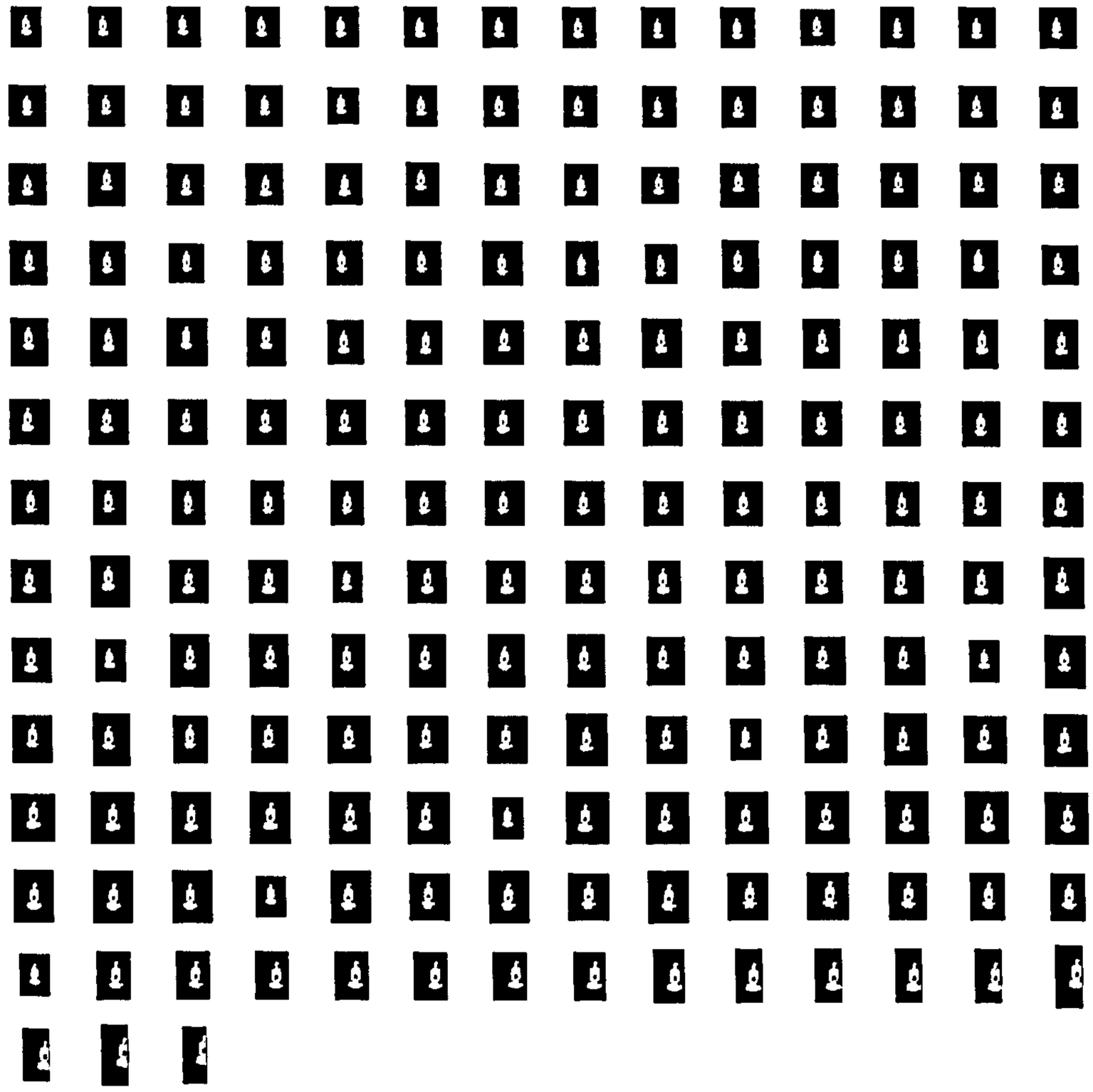


Figure B.7: Binarised segments with LARGE BUOY object in sequence A (frames 15-200).

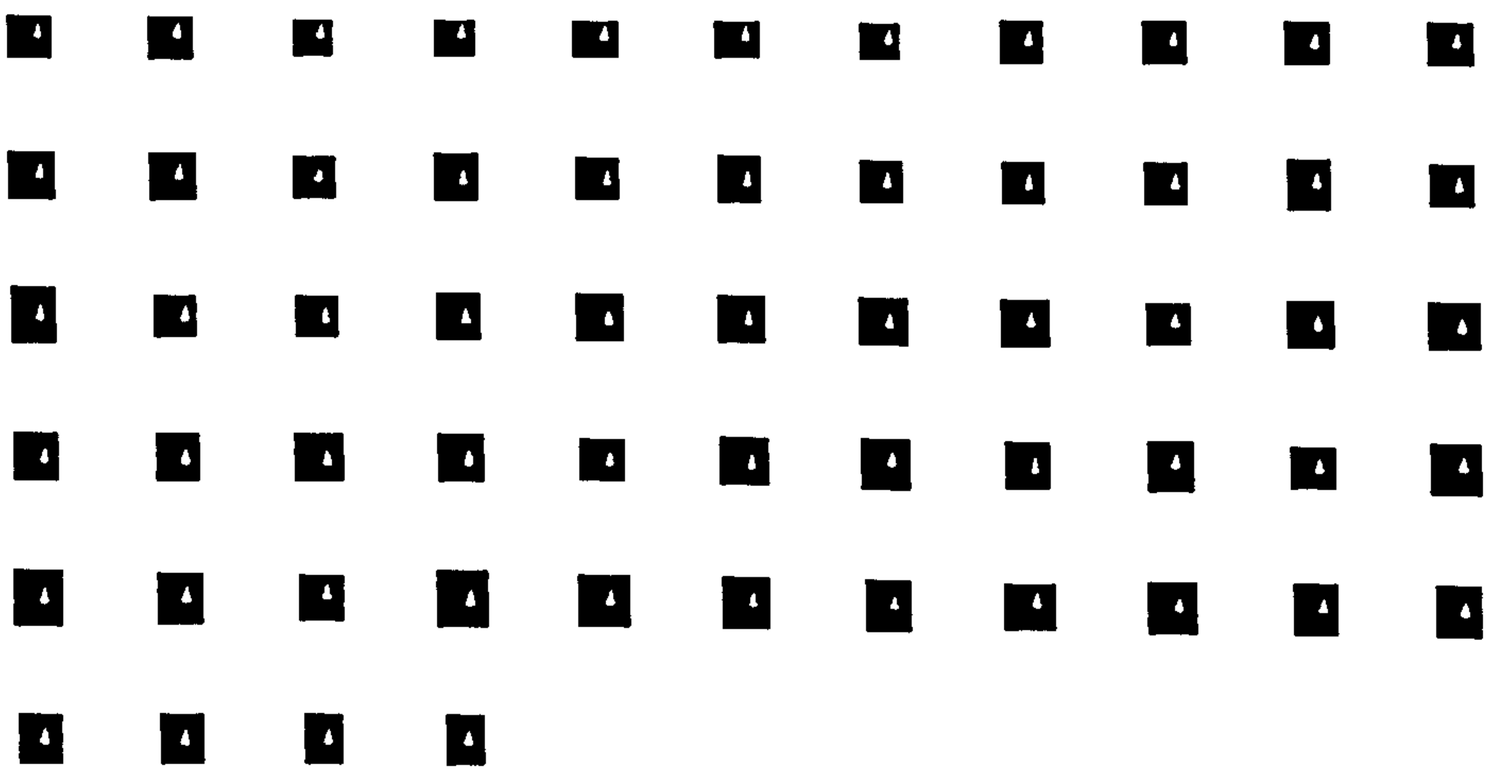


Figure B.8: Binarised segments with LARGE BUOY object in sequence B (frames 15-74).