Machine Learning of Projected 3D Shape

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Medical and Human Sciences

2009

Simon Coupe

School of Medicine

Abstract

This thesis primarily investigates the potential of the Pairwise Geometric Histogram (PGH) representation as the basis of a machine learning edge and view-based 3D object recognition computer vision system. The work extends 20 years' worth of associated research within the TINA computer vision research group [1]. PGHs have formerly been engineered as a solution to the presented problem, directly addressing all of the invariance characteristics required by such a representation. Previous research has proven the power of the proposed techniques for 2D object recognition through difficult, real-world viewing conditions including scene clutter and occlusion. This project extends the associated methodologies into the third dimension, exploring methods for representing scaled 3D objects' continuous appearances around their view-spheres.

The research agenda has also included a comparative analysis of the pre-existing TINA [1] stereo vision-based 3D Model Matching (3DMM) system, which is able to localise specified 3D objects in 3D scenes. In support of both mono and stereo methodologies, a quantitative scheme for accurately localising and verifying the presence of hypothesised image-projected 3D edge-feature models has been implemented. Full view-sphere sampled 3D model matching tests have been conducted for the competing methodologies, identifying significant shortfalls with the stereo-based approach to 3D model matching. The more powerful and reliable view-based techniques are subsequently analysed with regard to the more demanding task of comparative 3D object recognition.

Institution	The University of Manchester
Candidate	Simon Coupe
Degree Title	Doctor of Philosophy
Thesis Title	Machine Learning of Projected 3D Shape
Date	30th September 2009

Contents

List of Figures

Glossary

AI	Artificial Intelligence
AAM	Active Appearance Model
ASM	Active Shape Model
CAD	Computer-Aided Design
CCD	Charged Coupled Device
3DMM	3D Model Matcher
PCA	Principal Components Analysis
PGH	Pairwise Geometric Histogram
PHT	Probabilistic Hough Transform
ROC	Receiver Operating Characteristic
SIFT	Scale Invariant Feature Transform
TINA	There Is No Alternative
VB3DMM	View-Based 3D Model Matcher

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

Special thanks go to Dr. Neil Thacker for supervising and conceiving this project.

Gratitude is extended to Manchester University and their Imaging Science and Biomedical Engineering (ISBE) department for providing a great place to study and to the Engineering and Physical Sciences Research Council (EPSRC) and Siemens Corporate Research (USA) for project funding.

Chapter 1

Introduction

1.1 Background

In little over 60 years, by 2009, computing technology has evolved to revolutionise our lives. As associated technologies continue to evolve, applications including interactive robotics and intelligent environments will become a reality. Essential to many such technologies is the sense of sight.

Sight is typically our primary sense; its primary purpose being to work out what is where by looking. Whilst appearing an effortless task which we have evolved to be particularly adept at, the processes involved are far from fully understood. Evidence indicating that as much as half of the human brain is devoted to the sense of vision [101] puts the problem in perspective. Without a definitive theoretical model of natural vision, the problem of artificially emulating such systems has symbiotically challenged researchers for several decades. Despite a significant research effort, in 2009, computer vision is popularly characterised as being an immature and diverse subject [102].

The interpretation of projected 3D shape, or object recognition, lies at the heart of the problem of computer vision. Often inspired by insights into the workings of our own visual system, various techniques have been devised to perform this task, with, however, typically little relative success. While the concept of directly interpreting three-dimensional scene structure in order to subsequently infer objects' identities is held by many as the most plausible theoretical model of object recognition, a number of studies have provided evidence suggesting that objects are instead predominantly recognised by humans via reference to collections of topologically distinct two-dimensional view-based representations. Such an approach bypasses the problems associated with reliably inferring three dimensional structure in a bottom-up fashion and provides a viable means by which to recognise objects in images. The remaining problems involve determining which image features are most significant for the task, how best to limit the number of 2D representations required and how to efficiently match such image features to associated internal representations for recognition.

The problem of computer vision is further constrained by the sheer volumes of data involved. The size of typical detailed images and the rates at which visual events can occur require the processing of vast amounts of data in real time. Compounded by the complexity of the problem and associated memory and indexing considerations, only just now (2009) are mainstream computers becoming powerful enough to be able to perform such tasks.

It is common knowledge that much of a typical scene's visible structure can be inferred from a corresponding set of image edges, i.e. discontinuities in the intensity values across the image's pixels. The ease with which we can recognise objects from simple edge-descriptive line drawings confirms this. By definition, information is concentrated at edge regions, with adjacent, smooth image regions conveying little or no information. Edges therefore represent a much more compact and manageable form of input data, which preserve much of a typical scene's information. Another key reason for using edge-based representations for recognition is that edge features (i.e. ones defining structural discontinuities and projective boundaries) offer a high degree of invariance to environmental illumination conditions. Since an object recognition system should reasonably be expected to be able to recognise objects in novelly illuminated scenes, such a representation is essential. Studies have also provided evidence suggesting that the extraction of such edge-related information is a primary function of the human visual system [32] in the scene interpretation task. The reduction of over one hundred million light receptive retinal cells to around just one million optic nerves indicates the scale of this information filtering process. In line with mainstream computer vision theory, this project is accordingly concerned with the detection, modelling and recognition of view-based edge features.

Pairwise Geometric Histograms (PGHs) [42] have previously been devised and proposed as a solution for the recognition of 2D edge-based shape, possessing all the characteristics required by such a view-based learning recognition system. As the literature review will discuss, there are very few, if any, examples of other techniques with such characteristics described in the computer vision literature. In essence, PGHs encode the pairwise geometric relationships between an image's edge segments, allowing any such image features to be indexed to any object representations in memory. This project is primarily concerned with extending these view-based techniques for the recognition of 3D shape. Upon prospective development of the proposed shape recognition system, other auxiliary visual cues such as colour, texture and shading are envisaged to be accounted for.

Although a vast number of different viewpoints may be required to completely describe an object's projected appearance, aspect graph theory [9] has introduced the notion that objects' appearances can be described by a limited subset of such views. These 'aspects' represent local regions of view-space for which shape changes in a smooth, predictable pattern, preserving topological relations between incorporated features. PGHs are appropriate to encode such a representation because they too typically change in a smooth manner with small changes in model orientation or viewpoint. The main problem posed by this study is therefore how to most appropriately encode 3D objects' shape variability via collections of prototypical PGHs.

The inherent complexity of the problem of computer vision necessitates a modular system framework. The proposed shape recognition algorithms are therefore intended to contribute towards the foundations of a generically applicable computer vision system. Consistency and interoperability between any vision modules (and sub-modules) is to be maintained by employing systems' engineering principles and by relating the system's design and operation to quantitative statistics. Probability theory is proposed as being the only self consistent framework for data analysis and decision making.

Accompanying the existing 2D object recognition system, the TINA [1] computer vision development environment's current vision capabilities are complemented by a stereo vision-based 3D Model Matching system (3DMM). Another aim of the research is to evaluate the relative merits of view and stereo-based approaches to 3D model matching and object recognition.

1.2 Summary of the Aims of the Research

- Investigate the potential of the PGH representation as the basis of a machine learning computer vision system for the recognition and (3D) localisation of edge-defined projected 3D shape
- Assess the relative merits of competing methodologies described in the computer vision literature
- Establish a generic quantitative framework for the optimised localisation and verification of projected edge features
- Acquire a test set of 3D objects and corresponding view-based wireframe models
    - develop a user interface for view-based wireframe model construction
- Assess the relative merits of the pre-existing TINA stereo vision-based computer vision system (3DMM)

1.3 Overview of the Thesis

Following an introduction to the workings of the human visual recognition system, Chapter 2 goes on to describe the history of research in object recognition-oriented computer vision. Observing that the task of object recognition is highly diverse and challenging, research is focussed towards the problem of learning to recognise specific rigid 3D objects with well-defined shapes. The associated problems of deformable, more abstract and class-based object recognition are deferred for future research. The chapter ultimately proposes the Pairwise Geometric Histogram (PGH) representation as an optimal solution to the outlined problem.

Chapter 3 reviews the history of the Pairwise Geometric Histogram (PGH) representation, justifying the validity of associated methodologies for the task of edge and view-based 3D object recognition and localisation. The challenges facing development of the proposed 3D object recognition system are also outlined.

In the shared context of the pre-existing TINA stereo vision-based 3D Model Matching system (3DMM), Chapter 4 introduces a quantitative mechanism for the optimised alignment, camera parameterisation and subsequent verification of hypothesised 3D model matches in images.

Chapter 5 essentially describes the problem of interpolating PGHs around 3D objects' scaled view-spheres so as to provide a continuous representation of projected appearance. The bearing of the projective effects of perspective are also discussed.

With a methodology for representing continuous, scaled PGH deformation in place, Chapter 6 introduces associated mechanisms for the autonomous learning of such representations. Methods for subsequently recognising and localising objects in 3D space are also presented.

Chapter 7 investigates the utility of the proposed mechanisms for the individual 3D-model matching and localisation task. This functionality is a pre-requisite of any such system for the more demanding task of object recognition. The utility of the proposed view-based system is assessed relative to that of the aforementioned stereo vision-based model matching system (3DMM).

Chapter 8 serves as the pinnacle of this project's research, analysing the applicability of the proposed techniques for the task of 3D object recognition in arbitrary scenes.

Finally, Chapter 9 concludes the thesis, reviewing the research undertaken, identifying any significant findings and suggesting avenues for future associated research.

1.4 Contribution of the Research

The major contribution of this work is the development of a machine learning computer vision system for the recognition of edge-defined projected 3D shape, i.e. 3D object recognition.

The research is focussed on the particular problem of learning to recognise specific plain-surfaced rigid 3D objects, with the associated problems of class-based, textured and deformable object recognition being deferred for future research.

Unlike many other computer vision recognition systems, the proposed approach is also concerned with inferring the precise location and pose of any recognised objects in 3D space relative to the camera's frame of reference, i.e. 3D model matching.

The research proposes the Pairwise Geometric Histogram (PGH) representation as an optimal representation (i.e. adhering to the required invariances (see Chapter 2)) for encoding oriented spatial frequency edge distributions for the 3D object recognition and model matching task. Other prominent techniques for view-based 3D object recognition described in the computer vision literature are shown to fundamentally lack the representational scope and fidelity required for application to real-world recognition scenarios.

Observing the complexities associated with encoding high-dimensional non-linear shape continuously around objects' view-spheres, a locally linear manifold modelling technique mediated by piecewise triangulation has been established (Chapter 5). A view-based method for automatically learning 3D objects' continuous projected appearances has also been implemented along with a gradient descent type optimisation scheme for the efficient matching of complex triangulated view manifolds (Chapter 6).

Contrary to previous PGH-based research, where reference PGHs were stored at a range of consecutive scales for each object, the research presented in this thesis uses a single learned representation and instead scales the image geometry in intervals to match. Although a small number of learned representations may still be required to account for changes in an object's projected edge-based shape through scale extrema, the proposed method significantly reduces the memory overheads required by the system in accordance with the suspected associated processing of the human visual recognition system [98]. Treating local scale interval-based interpolation as an independent 1D model, a quadratic global linear model accounting for scale and view-point variance has been formulated (Chapter 5).

With a framework for learning and recognising scaled PGHs in place, a method for determining the 3D pose and location of a recognised object relative to the camera from a single PGH has been implemented (Chapter 6).

A fully quantitative scheme for evaluating the quality of match of a projected edge feature in an image has been developed (Chapter 4). A joint probability accounting for both edge location and orientation has been constructed and validated across multiple views of a set of test objects. Sampling of such terms across projected features allows for the corresponding models to be optimally aligned with the image data and for the presence of any projected features to be probabilistically verified. Fixed feature object models are shown to be suboptimal for the object localisation and verification task, with an adaptive modelling technique being introduced to accommodate a range of generic illumination and modelling dependencies.

As well as enhancing the functionality of the proposed view-based 3D model matching and object recognition system, the presented quantitative scheme for model match verification provides essential (previously missing) functionality for the pre-existing TINA stereo-vision based model matching system (3DMM). With a standardised localisation and verification metric in place, a comparative analysis of the competing mono and stereo-based approaches to model matching has been undertaken (Chapter 7), identifying significant failings with stereo-based model matching methodologies and, conversely, great potential with the proposed view-based methodologies. Stereo-based methodologies are accordingly discounted as the basis of a model matching and prospective object recognition computer vision system with view-based methodologies proving to be far more reliable.

With the integrity of the proposed view-based 3D model matching system validated, subsequent research analysed the applicability of the representation for the task of comparative 3D object recognition (Chapter 8). Initial experiments proved that the representation was sufficient to differentiate 3D objects as presented on their own against plain backgrounds. As with the model matching experiments, results were realised using limited sets of reference model and scene features in order to limit associated temporal processing costs. In transferring the proposed object recognition techniques to the task of recognising objects in cluttered and occluded scenes, significant problems arise in terms of the amount of data that requires processing. The sequential nature of standard computing hardware (2009) was shown to be insufficient to support object recognition in complex scenes across large object databases (including the test set of 15 objects) in a reasonable amount of time. The proposed model matching techniques are otherwise shown to be robust to partial occlusion and mild scene clutter, allowing recognition to be based on the proposed projected model verification metric with limited focus feature sets.

Chapter 2

Background to Computer Vision-Based Object Recognition

2.1 Introduction

Following swiftly on from the inception of the programmable computer in the late 1940s, the emergence of the fields of artificial intelligence and robotics brought with it interest in developing an artificial sense of vision. The machine-based emulation of natural visual processing has challenged researchers ever since, with the topic of computer vision now being relatively mainstream throughout the world's academic and industrial research arenas.

At the heart of the problem of computer vision is the task of object recognition. An 'object' can perhaps best be defined as a tangible and visible entity. Clearly, this definition encompasses a very broad range of things, existing at a very broad range of scales, opening up the definition of the recognition task. A wide range of literature has been put forward suggesting solutions to the problems associated with machine-based object recognition. However, it is evident that a working system, able to emulate the associated capabilities of the human visual system, remains far beyond our current capabilities. Instead, most applications tend only to very restricted aspects of the problem of object recognition, such as, the recognition of specific sets of objects under heavily constrained conditions, whilst typically contributing little towards a generic solution. In relation to the evolution of natural vision, man-made vision is still in its primary infancy.

The multifarious nature of vision and its seemingly inextricable association with a more general sense of intelligence, suggests that a generic solution may still be a long way off. It is evident, for instance, that we are able to utilise many separate channels of information to aid our interpretations of viewed-object identity. For instance, at a basic level, shape, texture and colour may all be used to aid recognition, while material and functional properties and even contextual scene information are evidently further utilised to support our decision making. Relative to the limits of human visual perception, the problem can thus be regarded as being AI complete, although there are many visual competences that are readily addressable. Alongside the intellectual challenges and philosophical musings associated with solving machine vision, the subject is very much application driven. To date, the majority of research funding has been associated with surveillance, robotics and web-search type applications. There is very little funding available for research addressing the more deep-rooted scientific issues related to solving artificial vision, especially in the UK. Computer vision applications are also key to advanced warfare technologies such as missile guidance systems and autonomous vehicles. One can only speculate as to the exact nature of computer vision technology beneath the bounds of the published literature.

Within the realm of object recognition, there are 2 key areas of research. These are specific object recognition and more generalised object class recognition. This research is specifically concerned with developing an infrastructure for learning to recognise the 3D shapes of specific objects. It is however envisaged that exposure to and knowledge of a number of entities from the same class will result in strong visual associations being drawn to a novel object of the same class, thus supporting class-based inference.

Object recognition potentially involves the need for recognition of anything visible; from natural forms, such as leaves, plants, fields, clouds, animals or people, to man-made objects, in all their guises throughout domestic and industrial domains. Driven primarily by commercial interest in remote biometric sensing technologies, the human face has perhaps been the most studied type of object in the history of computer vision. The recognition of human faces is one of the most challenging aspects of object recognition because faces' structures are typically very similar to one another and are relatively free to deform through expression changes, occlusion (hair, apparel, makeup) and a number of other temporal considerations (illness, surgery, weight-change). As humans, we are able to effortlessly distinguish vast numbers of different faces, highlighting the acuity of the human visual recognition system. It should however be noted that separate, so called ‘expert object recognition’ brain centres are thought to be responsible for such specialised recognition processes [87]. Interestingly, human face recognition is sensitive to image-plane-orientation, with subjects showing a clear preference for upright faces [81], whereas general object recognition tends to be image-plane-rotation invariant. This is attributed to the fact that we typically only ever view faces in an upright pose, so our brains have learned to focus their processing around such observances. We can otherwise learn to dissociate such invariances through repeated exposure, highlighting the plasticity of the human visual system. This project is primarily concerned with the more generic task of recognising rigid 3D shape, in this case; man-made 3D objects, with a view to subsequently extending the techniques for deformable objects. Applications pertaining to interactive robotics are primarily envisaged.

A review of the history of face recognition would take up the bounds of such a chapter in its own right, so will only be discussed loosely during the remainder of the chapter. The subject of face analysis introduces a fundamental consideration regarding the analysis of imaged shape, one that has been a source of contention for computer vision researchers and enthusiasts throughout the history of computer vision. This involves the distinction between appearance-based representations, in terms of the raw values of the pixels encountered across the image, and feature-based representations that utilise higher-level derivatives of the pixel values as the basis of recognition. These issues will be discussed in detail, in the context of the existing literature, throughout the remainder of the chapter.

Early work in object recognition concentrated on the sub-problem of 2D shape recognition, i.e. being able to recognise head-on planar objects, or 3D objects from fixed viewpoints. Since we are essentially dealing with the recognition of 2D projections of 3D objects, i.e. retinocentric inputs, this was a natural starting point, even a prerequisite, to the problem of 3D object recognition. The main problem with transferring the techniques to 3D object recognition is that vast numbers of views may be required to make up many objects’ appearances. This raises the final key issue regarding object recognition research; that of whether internal representations are based around 3D models, or upon collections of 2D view-based appearance features sampled from around objects’ view-spheres.

Accompanying developments in computer vision, studies into the workings of the human visual system have offered valuable insights into to how exactly we interpret visual information and see. Such physiological knowledge, pertaining to the best known operational scheme for object recognition, has significantly influenced the nature of computer vision and object recognition research. This chapter goes on to review some of the most influential theories of object recognition, before detailing the state of the art in computer vision-based object recognition and justifying the research herein undertaken.

2.2 Natural Vision

No investigation into the development of an artificial vision system should proceed without reference to biological vision. Nearly all super-primitive creatures on earth have some form of visual system. As humans, we are very much visually oriented beings, with our sense of vision serving as a direct, multifaceted and generally invaluable interface to our environment. The human brain and visual system are thought to be the most advanced of their kind, although many other creatures have extremely acute specialised vision systems. It is natural that vision be harnessed by any prospective artificial sensing systems.

At the base-level, human sight involves the sensing of projected distributions of electromagnetic radiation in the visible range of the electromagnetic spectrum. Sight is then an intelligent interpretation of these distributed patterns of relative intensities in terms of a virtual immersion of the viewer's conscious frame of reference and a binding between virtual 3D object representations of recognised objects, with all their metaphysical associations, and any temporally observed scene features.

In more pragmatic, albeit greatly simplified terms, such photon energies are focused through our eyes' lenses onto the retinas at the backs of our eyes. In all but the very central regions of our retinas (the 2.5 to 3 mm diameter foveae) a grid of receptor cells (the rods) samples the incoming light in terms of grey-level intensity, with cell density increasing toward the central foveae. Because these rod cells are so sparsely situated, especially at the perimeters of our visual field, their acuity is very low. They are, however, very sensitive to low-level light and are essentially used for peripheral and night vision and motion detection. There are thought to be around 120 million rod cells in a typical human retina, each measuring approximately 0.002 mm in diameter. The foveal regions are instead populated with a much more concentrated distribution of exclusively colour sensitive cells called cones, the density of which falls off logarithmically toward the peripheral vision. There are around 6.5 million cone cells in a typical human eye, each about triple the diameter of a rod cell. These cone cells are less sensitive to low-level light than their counterparts, but are useful for very fine scale and colour vision. There are then 3 distinct types of cone cell, corresponding purely to the red, green and blue wavelengths of the visible spectrum. Nearly two thirds of the cones are estimated to be red responsive, one third to be green responsive and then just 2% to be blue responsive. In contrast, the peripheral rod cells respond inverse-proportionally to the mono-sampled intensities of the red, green and blue channels of light. The whole spectrum of visible colour is inferred from the relative intensities of these 3 types of cone cell.

Upon detection, the conversion of light energy into virtual object representations, consciously projected from within the human brain, becomes an extremely complex affair. Decades of research in neuroscience, electrophysiology, psychology, psychophysics and functional neuro-imaging have however offered invaluable insights into the workings of the human visual system. At the outset, it is established that the retina performs a form of edge-feature detection. Although there are around 120 million rods and cones in a typical human eye, there are only around 1.2 million retinal ganglion cells that go on to connect to the back of the brain via the lateral geniculate nucleus. Approximately half way to the back of the brain, the optic nerves merge at the optic chiasm, from where respective left-right halves of each eye pass through the optic tracts to the respective halves of the brain, so that the left half of the image, as viewed from both eyes, is received by the left half of the brain and vice-versa. These ganglion cells act as intensity gradient edge feature detectors, of varying complexity, individually branching out across the retina into sets of centre-surround-type receptive fields, selectively tuned to detect specific combinations of edge features at various relative orientations, positions, polarities, wavelengths, intensities and scales. The cells associated with high-definition foveal vision tend to have very simple, focused and localised receptive fields, while more-general receptors in the peripheral vision may connect to many thousands of widely distributed photoreceptors. It is thought that the brain subconsciously sees based purely on such edge feature detection. Such an understanding has motivated much computer vision related research towards finding object representative feature sets that are relatively invariant to viewpoint transformations. For the uninitiated reader, the key qualities of edge features are that they are extremely compact, descriptive and robust to deformation across wide-field changes in environmental illumination. There is otherwise simply no information conveyed by a featureless surface and it is a waste of resources to concentrate on processing such redundant information on a pixel by pixel basis. Image edges are typically all that the human vision system requires for shape or object recognition and localisation. It is otherwise a moot point as to whether the absence of surface information can itself be treated as information in support of vision related tasks.

There are thought to be around 9 specialised brain centres corresponding to distinct visual tasks such as recognition, localisation and colour and motion inference. These are distributed through the dorsal and ventral streams of the visual cortex. The dorsal visual stream is fast, independent of colour and thought to be more concerned with spatial aspects of vision, such as motion, localisation and navigation. Object recognition and detail perception are instead thought to be conducted via the slower, but more detailed and colour-responsive ventral stream. Object recognition is suggested to correspond to a massively parallel information filtering process, in which constellations of various image edge-type feature detectors are indexed to specific views of specific objects, as stored in long-term associative visual memory. Visual occurrences of known objects invoke a significant number of feature detectors to elicit a discernible combined response to prompt recognition of the viewed object.

Although our conscious experiences of sight are continuous, surface-filled, coloured, detailed and relatively wide-field, we only see detail, e.g. computer monitor-based text, through a view-cone of just about 1 or 2 degrees in diameter for each eye at any one time. We also only see colour through the central, cone populated region of our vision and even have blind-spots where the retina's ganglion cell axons bunch together to pass back through the optic nerve. To compensate, our eyes rapidly and subconsciously dart around interest points in a scene, collecting and stitching together detail in a process known as saccading. We are effectively blind during saccadic motion because the eyes move more quickly than they can process any detected light. Our sense of vision essentially constructs a virtual, continuous, coloured, wide-field, surface-filled, overlaid 3D representation of our relatively situated environment. Psychophysical research has shown [98] that humans' gaze patterns are remarkably consistent in terms of saccading to specific regions of imaged objects. There is however no consensus as to exactly which object elements draw the human visual systems' attention or how exactly these regions are processed for recognition. This knowledge may in itself provide a replicable solution for computer vision-based object recognition. We can however assume that any such features are of relatively low resolution, since that is all that exists in the peripheral vision. This may be beneficial to object class recognition, since objects from the same class will tend to merge in similarity as their global shape appearances are blurred.

The standard uniform rectangular nature of CCD camera sensors clearly differs from that of their biological analogues. Given enough pixels, we can however replicate any distribution of sensors, such as that of a retinotopic map. Thacker [98] has suggested that there are many advantages to be gained in designing vision sensors in accordance with the invariance characteristics exhibited by biological vision sensors. The varying spatial resolution of the retina is shown to support scale invariance through concurrent multiple-scale sampling (as proposed for recognition in this project; see Chapter 5) and the spherical nature of the retina is shown to provide invariance to perspective distortion in the peripheral vision as would otherwise be experienced with a wide-field flat sensor. Furthermore, the retina's logarithmic sensitivity to illumination intensity is shown to directly support homogenous error sampling in regard to any illumination invariant sampling (gradients) across the photoreceptors. In terms of higher-level visual processing, the brain's parallel processing routines can be replicated on our traditional sequential computing hardware, although standard computing power is only just now approaching the levels required to support high-level temporal scene analysis tasks. Typically, computer vision-based object recognition is treated as high-resolution foveal vision in a static and uniformly sampled context, or the concept of a logarithmic radial sensor is adapted to a uniformly concentric rectangular one. Grey-scale data is commonly used for object recognition tasks, bypassing the complexities associated with combining the three separate colour channels. Grey-level data carries all the information required for shape recognition, although colour is exceptionally useful for coloured target detection and enhanced discrimination in certain circumstances.

Given the expense of processing high definition, temporal visual information, it may be that a multi-scale, saccadic processing paradigm, akin to our own, may be required by any intelligent, autonomous sensing machines. Despite the added value that vision brings to our lives, it is of course completely useless when the lights go out and in relation to anything happening above or behind us. To this end, artificial vision systems should consider application of complementary, super-human viewing technologies. A simple extension of artificial vision sensors into the infrared band of the electromagnetic spectrum, for example, would allow the vision system to detect environmental heat traces independently from a more traditional sense of sight.

Despite the many astounding competences exhibited by the human visual system, its potential limitations are further evidenced by phenomena such as change blindness [58]. Change blindness is a very interesting phenomenon that occurs when our vision is temporarily cut-off through saccadic motion, blinking or distraction of attention. Such effects can be artificially induced by inserting blank image frames into viewing screens' projection streams. It can then be demonstrated that a wide range of otherwise extremely obvious sequentially-projected changes can be made to the image presented after the flash without the average viewer being able to notice (as long as they are not paying attention to that element at the time). Given an appropriate visual distraction, it has been shown [62] that people can even be interchanged in social interactions without the subject being able to notice. This provides evidence supporting the dichotomy of the key visual competences credited across the dorsal and ventral streams, i.e. generalised spatio-temporal consciousness and more subconscious high-fidelity specific-object recognition. The medical condition of agnosia further supports this specialised brain centre hypothesis, providing evidence that patients with damaged ventral streams are able to see and copy line drawings of objects, but are unable to recognise them.

Just as mankind has leveraged the power of natural competences such as locomotion and flight through the deployment of advanced technologies, it may be that machines ultimately pose the power to operate a sense of vision far more advanced than our own. The neural processing of the human visual system may have little bearing on that required by any prospective omni-directional, multi-focus and high resolution sensing machines. In the mean-time, it provides an invaluable example of how light energy may be systematically processed for remote scene analysis. It is interesting to speculate how earth-bound natural vision may itself evolve throughout coming millennia with and without the bearing of assistive and interactive technologies.

2.3 Marr

Perhaps the most influential work in the history of computer vision has been that associated with David Marr (1945-80). Credited with co-founding the discipline of computational neuroscience, Marr went on to revolutionise the science of computer vision, by taking inspiration from the sciences of psychology, AI and neurophysiology, to produce a mathematical account of human vision. His efforts culminated in the book: ‘Vision: A Computational Investigation into the Human Representation and Processing of Visual Information’ [12]. Sadly, however, Marr was to die of leukemia whilst completing this work, Vision being published posthumously, nearly 2 years after his death.

Marr describes vision as an information processing task with the purpose to produce, from images of the world, descriptions which are useful to the viewer and which are not cluttered with irrelevant information. Marr came to the conclusion that the shapes of objects could be determined in a bottom-up fashion by sight alone (i.e. without using prior knowledge of the viewed objects or preconceptions of expected scene content). Marr concluded that vision’s primary function was to build descriptions of the shapes and positions of things from images. Noting that it would be impossible to achieve this feat in one step, Marr proposed a sequence of representations that would facilitate the progressive recovery of 3D geometric data from 2D images.

The first step of Marr’s proposed approach was the formation of the ‘primal sketch’, which can itself be decomposed into three stages:

(1) The detection of zero-crossings.
This stage of processing is concerned with detecting significant pixel value intensity changes in an image. Marr argued that such features could best be found by detecting zero-crossings, produced by the Laplacian of Gaussian second derivative gradient operator, operating at multiple scales, around an octave apart. Marr further argued that physical edges in an image would tend to be indicated by correspondences between these zero-crossings at adjacently sized scales. This was termed the ‘spatial coincidence assumption’.

(2) The formation of the raw primal sketch.
Given the locations of an image’s zero-crossings, this information is classified by assigning corresponding representative tokens, such as edges, bars, blobs and terminations, with attributes such as orientation, contrast, length, width and position.

(3) The creation of the full primal sketch.
The final stage of the process is to group together the primitive tokens formed by the raw primal sketch, to recursively form higher level constructs at different scales, such as; further tokens, virtual lines and boundaries. This process would culminate with

a multifaceted description of the spatial organisation of any intensity changes within a given image.

Following formation of the primal sketch, Marr's second stage of representation was the '2.5D sketch'. This would form a viewer centred representation of the orientations and relative distances of the visible surfaces within a scene, complemented with contours of discontinuity. This would represent the limits of what pure visual perception could achieve. Whilst Marr discussed various representational possibilities for the 2.5D sketch, he ultimately left such an interpretation open, suggesting that while surface orientation required accurate explicit representation, depth only needed representing relatively roughly.

The formation of the 2.5D sketch would however be a very complicated affair, drawing on information from a variety of processes which the central portion of Marr's book describes. Such processes concerned; stereopsis, directional selectivity, structure from motion, optical flow, occluding contours, surface contours, texture and shading. Each associated process would utilise information from various stages of the primal sketch.

The final stage of processing was the conversion of the viewer-centred 2.5D sketch to an object-centred '3D model representation', which Marr proposed would be required for recognition. Marr's reasoning was that since objects can appear different from a vast number of different perspectives (and under a wide range of illumination conditions), an inordinate number of descriptions would be required in memory by a view-based system. Whilst Marr reflected on Minsky's [7] idea of minimising the number of descriptions needed by using appropriate shape primitives and selected views, he foresaw limitations with such an approach, perceiving the best solution to be to use single 3D object representations. The proceeding problem was that of how to represent such volumetric shapes.

Marr listed the criteria for 3D shape representations as being; accessibility, scope and uniqueness, and stability and sensitivity. He concluded that a suitable representation should involve an object-centred coordinate system, including, although not exclusively, volumetric shape primitives, along with some form of modular, hierarchical organisation. For the sake of simplicity, Marr limited his discussion to shapes with axes based on elongation or symmetry. Generalised cones [6] were proposed as being such a suitable representation. (Generalised cones refer to the volume formed by moving a fixed, possibly variably sized 2D shape along a smooth axis.)

Given such a scheme for representing objects in memory, the ultimate stage of Marr's proposed approach was to infer the coordinate systems and associated relative component axes of any models represented by the 2.5D sketch. These descriptions could then be matched to memory in order to recognise known objects. Whilst pointing out that not all shapes could be represented in this way anyway, Marr states that he has no complete solution to this problem. Instead, only solutions to examples under restricted conditions are discussed. Presuming that such information could however be reliably inferred, Marr goes on to discuss various strategies for matching such descriptions to any in memory. In essence, a coarse to fine resolution model component search strategy is suggested. At this stage, the bottom-up strategy is complemented by a top-down one, where correspondences with specific parts of models can provide information about the likely appearance of whole associated objects.

Throughout 'Vision', Marr makes it clear that there are many gaps in his account and that his book in no way solves the subject of human visual perception, instead focusing on the early stages of vision. Marr's illness and related untimely death presumably also constrained the content of his work. While, undoubtedly, Marr's work has brought inspiration and progression to the science of computer vision, it is evident that the validity of some of his methodologies is questionable. In context, over a quarter of a century later, it appears that no one has been able to replicate such a working object recognition system.

2.4 Recognition by Components

Following on from Marr's 'Vision', Irving Biederman was to propose an alternative account of object recognition, described in his 1987 paper: 'Recognition-by-Components: A Theory of Human Image Understanding' [17]. Biederman provided evidence that objects could be recognised just as readily from line drawings as they could from colour photographs. This would support Marr's reasoning, in line with overwhelming physiological evidence [32] that our early visual pathways are primarily concerned with detecting edge related information. In correspondence with Marr's approach, Biederman's is concerned with detecting 3D models of objects, from 2D images, which can be compared to any 3D models in memory to indicate recognition. As with Marr, Biederman proposed to do this by detecting connected groups of 3D shape primitives. Biederman's theory would however be a lot simpler than Marr's, offering a potentially more viable means by which to perform 3D model extraction and recognition.

Central to Biederman's approach is the use of geons (geometric icons), which are a set of 36 primitive shapes (e.g. blocks, cylinders, wedges, and cones (all describable by generalised cones)), combinations of which were proposed to represent a wide range of 3D objects. Biederman likened this approach to the way in which words and sentences are recognised from combinations of the 55 different phonemes in the English language. Avoiding the complications of constructing 3D scene representations (such as the 2.5D sketch), Biederman proposed to be able to detect these geons in 2D images by detecting related 'non-accidental', generally invariant edge properties, such as; curvature, collinearity, symmetry, parallelism and co-termination. By detecting such properties in association with any regions of 'deep concavity', groups of connected geons could be detected that could be matched to any models in memory to indicate the occurrence and location of any known objects. These geon properties were exactly those suggested by the Gestalt community in the 1930s and 40s to be required for the perceptual grouping of scene and object structure [76].

Biederman suggested that the detection of just 2 or 3 connected geons would facilitate object recognition. Furthermore, such a recognition system would offer a high degree of invariance to factors such as occlusion, rotation and image degradation. For simpler objects, consisting, for example, of only single geons, Biederman suggested that representations could be supplemented with other information, such as; colour, texture, small details or context. Biederman's later work also suggested the possibility of recognising clusters of geon-based models for general scene recognition tasks [73]. It must however be noted that Biederman's account avoids the specifics of edge detection, instead presuming that significant edges could be reliably extracted from images. Such a presumption, including the use of manually composed line drawings, is likely to be a severe restriction on the practical scope of Biederman's theories. Biederman goes on to suggest using neural networks to help in implementing his theories, although, he admits to his model being nothing more than a working hypothesis, being admittedly incomplete. Whilst a neural network implementation was further researched in association with John Hummel, the associated paper [34], published in 1992, reported limited progress in this endeavour, concluding with a list of shortcomings. It is evident that no further progress has been reported to this day.

Biederman's originally cited paper [17] also discusses the research of Palmer et al.

(1981) [11] who investigated the perceptibility of objects viewed at different orientations. This work indicated that a three-quarters front view (a canonical view) was generally most effective for object recognition, with subjects showing a clear preference for such views. In relating this to his recognition by components theory, Biederman suggested that such canonical views would maximise the information with which to match the components in images to the representations of the objects. With this reasoning, Biederman discusses that some, albeit rare views of geons, would not allow for any non-accidental properties to be recovered. For example, in the extreme, looking head on at a rectangular box would only indicate the presence of a rectangular planar region. In the same discussion, the research of Jolicoeur (1985) [13] was mentioned, in which, reaction times for naming objects were found to be proportional to the degree of misorientation from the object's normally upright, or, only experienced positions. Whilst Jolicoeur concluded that this was due to some form of mental rotation, Biederman seems to discount such theories as relating only to rare conditions where relations among models' components are rearranged. Jolicoeur's findings seem however to apply to far more general conditions than these, in contrast with Biederman's theory that geons and associated models should be recognised just as easily from most viewpoints. Despite further evidence supporting mental rotation of objects for recognition (Shepard and Metzler 1971) [5], Biederman concluded that evidential mental rotation rates appear to be too slow and require too much effort to account for the ease and speed with which we can recognise objects at varying orientations.

Inspired by the 'Recognition by Components' approach for generic object class recognition, Bergevin and Levine introduced the PARVO computer vision system in 1993 [40], concerning themselves with the recognition of classes of objects, such as cups and stools, based purely upon representative line-drawings. These line drawings were similar to basic (hard-drawn) CAD models with hidden lines removed. The key to their work was in segmenting any closed regions, e.g. any planar or cylindrical surface regions, and detecting what geons these could relate to; for instance, the top of a lamp shade or the base of a pan. Any distinctive sets of prototypical connected geons were then inferred via a discrimination tree in order to recognise a range of 3D objects from a wide-range of viewpoints. Although shining light on the processing requirements of such generalised visual tasks, the authors observe that their system is critically affected by missing or occluded line data, although it is shown to work under certain classes of occlusion. Although a system can be designed to recognise an object based upon only fragmentary global evidence, PARVO bases its whole operation on the availability of data that just is not typically available in images, i.e. complete line-drawings with detailed conjunctions. Even the most popular techniques for edge detection, such as Canny [14], produce edge maps of objects that are severely disrupted through everyday factors such as specularity, lack of contrast and occlusion. Recognition systems must be designed in consideration of the limits of the available data. These considerations suggest that it may be better to base such difficult recognition problems directly on the 2D projected data, rather than abstracting the basis of representation into the very noisy and error-prone third dimension. Geons are also likely to be very ambiguous for many objects' qualitative feature projections and the view-based processes involved in individual geon detection could perhaps just as easily be extended to detect objects as viewed without having to then search for correspondence in 3D.

Some progress was made towards geon-based 3D object recognition during the 1990s, which can be broadly summarised by an academic-expert-panel review in 1997 [60]. The authors concluded that although workable in a number of artificial (CAD-based) schemes, the matching of image detail to qualitative descriptions remained an unsolved problem. In contrast to recognition by components theory and supported by some of the findings just discussed, a growing body of research would come to support an alternative approach to object recognition, involving not 3D models, but collections of 2D views.

2.5 Recognition by Views

Marr discredited view-based object recognition methodologies because too many different views would be required in memory to be able to match all the possible appearances an object might pose. Marr's reasoning was further supported by the prevailing need for us to be able to infer depth and 3D structure from vision. However, a number of researchers have noted that objects' projected shapes often change in a smooth predictable fashion across collections of adjacent viewpoints. Notably, Koenderink and van Doorn's 1976 and 1979 research papers [8] [9] introduced the concept of the visual potential graph, in which a viewing sphere, surrounding a specific object, would be partitioned into 'aspects', each accounting for a cluster of topologically stable viewpoints. The extremities of each aspect region would mark the viewing positions at which unpredictable visual events would occur, such as features coming into or going out of view. Such a representation would mean that a 3D object could be roughly fully represented by a limited set of views, thus providing a simple and viable method for recognising objects in images directly from their appearance.

The motivation for developing view-based object recognition systems was further supported by observations articulated by Rosenfeld in his 1987 paper [15]. Rosenfeld estimated that humans could recognise unexpected objects in around 100 neuron-firing times, indicating that very few computational steps are involved. A recognition scheme such as Marr's, involving object centred model representations, is however inherently sequential, casting doubt on the validity of such processes. In further support of such a view-based recognition scheme, a number of other studies [23][33][35][49] would come to light, indicating that humans store only specific views of objects for the purpose of recognition. Related experiments showed that the recognition of objects at previously unseen viewpoints was progressively worse as a function of the distance from the nearest familiar viewpoint. This seemed to support the theory that we need only memorise certain generalisable views of objects for recognition, in a similar manner to the formulation of an aspect graph. For situations when we observe an object at a novel viewpoint, mental interpolation or extrapolation schemes would effectively adjust the object's projected appearance to correspond to the nearest stored view (or vice-versa). If no views correspond, a new representative reference node is stored. For familiar objects, experience of the objects at all viewpoints results in a complete aspect graph type representation, offering viewpoint invariance for recognition.

The remainder of this chapter goes on to discuss the history of developments in view and computer vision-based object recognition. It should be noted that certain aspects of 3D object matching still play a significant role in human sensory perception and should not be excluded from general consideration. For instance, stereo visual perception gives us strong cues relating to dynamic feature-based 3D shape in our immediate vicinities, which can be used for reasoning about object identity and, especially, relative position. This will be discussed in more detail in Chapter 7, which describes the TINA stereo model matching system in direct contrast to its view-based counterpart. Interestingly, it has been observed that many herbivores have opposing eyes and no sense of stereo. This gives them the advantage of being able to see all around them so as to be able detect any impending environmental threats, while stereo vision in carnivores is thought to have evolved to aid hunting of such prey in dynamic, close-quarter scenes. The domains of any potential machine-vision applications will presumably have similar bearings on the evolution of machine-vision.

The field of range data analysis has also experienced a significant amount of research attention over recent years. Range data analysis is a form of 'active vision', in which objects' 3D surfaces are remotely sensed (instead of detecting their projected 2D edges) using technologies such as radar or laser scanners. Object recognition is then performed by comparing 3D surface-based models. Although humans do not have such sensing faculties, such recognition paradigms draw parallels with human tactile object recognition. The bottom line is that natural vision is particularly adept at single-view-based object recognition and this is the most straightforward, convenient and potentially powerful means by which to remotely sense our environment.

Given this evidence indicating that our own object recognition systems deal with view-based information, rather than 3D models, and considering the practical benefits to be gained from such representations, a number of related research papers transpired in the early 1990s. Poggio and Edelman (1990) [26] and Edelman and Weisenhall (1991) [29] reported neural network schemes that were able to recognise wiry 3D objects from any viewpoint, with only a small number of stored views. The latter paper reported performance comparable to that of human subjects. Since the system had no provisions to rotate 3D objects, this would support a theory of the use of blurred template matching or non-linear interpolation in human recognition. Ullman and Basri's 1991 paper [30] presented a scheme in which objects' continuous appearances were represented by linear combinations of 2D views of line drawings, although restricted to orthographic projection under strong idealised assumptions of model to scene feature correspondence. One concern with the aforementioned types of scheme is that they are very artificial in terms of the wireframe objects they use and the information that is assumed available to the system for recognition. Point to point correspondence is generally only realisable once an object has been recognised. More recently, Leek (2005) [91] presents psychophysical evidence confirming the hypothesis that human visual recognition is not mediated by volumetric image segmentation or volumetric shape primitives, but by 2D shape components.

2.6 Appearance-Based Object Recognition

Many objects exhibit very consistent appearances throughout changes in illumination. This is especially true for man-made objects displaying 2D printed texture, e.g. cereal boxes, especially when viewed through largely ambient illumination. Although there are obvious complications with any appearance-based approach to recognition from the outset, regarding directional illumination and the variability of relative shading across 3D surfaces, this has inspired research into 3D object recognition based directly upon projected appearance; i.e. shape matching based upon the raw pixel values sampled across an object. Such an approach potentially bypasses any complexities associated with having to select specific types of image features for recognition. In industrial quality control scenarios this may be exactly what is required to automatically detect any blemishes or faults on objects in a uniform production line, although, as will be discussed, shape recognition is essentially mediated by appearance gradient information.

The widely cited research of Murase and Nayar (1995) [48] describes an appearance-based approach to object recognition, purporting very high levels of recognition competence across a database of 100 3D objects in real-time. Their approach revolves around reducing an object's projected appearance to a low-dimensional (more tractable) subspace, spanned by the principal vectors of an eigen-decomposition of a set of such images This process essentially identifies the main independent features of an object's global appearance across its sampled view-sphere. Although noting that many eigenvectors may be required for very accurate object modelling, the authors observed that only a few are required to capture the significant appearance characteristics of many objects. Their work proceeded by parameterising connected selections of normalised training images of objects in this low-dimensional space and by then forming a continuous manifold passing through each associated sample for each object. Their work simplified this process by only considering objects viewed in a 1D geodesic trajectory around their view-spheres (as captured upon a turntable), so creating a continuous path through parameter-space for each object's manifold. To recognise the object from a novel viewpoint, the object's appearance would be projected into this eigen-space and the closest point on the learned continuous manifold would be taken to correspond to the orientation of the observed object. It was suggested that a separate, global representation space be used for initial inter-object recognition, although it was observed that this raises complications when adding objects and having to recalculate the interrelated bounds of this global representation space. Around 20 eigenvectors were suggested to be required to represent each object for their experiments.

Despite the apparent qualities purported by Murase and Nayar's approach to 3D object recognition, closer inspection of the limitations of such an appearance-based approach to object recognition raises some significant concerns. In only considering the problem of recognising objects from single geodesic paths around the view-sphere, under only a handful of separate light sources, the authors are trivialising the problem of object recognition. Perhaps the most critical observation is that any 3D object's general appearance may deform substantially under changes in directional lighting. Although we can assume that environmental illumination will typically be sourced from above, either from the sun or conventional man-made light fixtures, this can never be assured, particularly if an object is turned upside down. It is also desirable, as with the human visual system, to be able to recognise an object independently from whatever illumination sources may come its way. The authors go on to quote that it would be impractical to parameterise a library of objects in this way if arbitrary rotations and illumination directions are considered. The number of eigenvectors required for unambiguous shape recognition across a large database would be too large to make the scheme tractable.

There is another fundamental problem with such appearance-based schemes for object recognition, in that the representation space is constructed from samples registered in the same relative coordinate frame. This means that an object will only be recognised if it is the same way up and of the same size as when learned. Although the size of the extracted object can be normalised, to be able to recognise the object at any image plane orientation, an additional search would be required through the 360 degrees potentially posed by the object in the image frame. As if these issues are not enough of a problem, there remain complications associated with clutter and occlusion.

It is reasonable to insist that any object recognition scheme be robust to scene-bound interference from clutter and occlusion. The remaining concern with global appearance-based representations is that they require objects to be viewed independently from any other scene structure. This invalidates the approach in terms of recognising objects that are only partially visible and means that any object requires automatic segmentation (and orientation) from the image before it can be processed. This is tantamount to saying that, unless the object is presented on its own against a plain background, we need to recognise the object before we can match its appearance. A chicken and egg type problem then ensues. A key part of the problem of object recognition is to segment objects' appearances in the first place and without knowing what the objects are this can be a very difficult, if not impossible, task.

'Object Recognition Using Appearance-Based Parts and Relations' (1997) [56] attempted to address the short-sights of previous appearance-based object recognition work by breaking objects' representations down into their key components, such as any planar faces, so that each part could be recognised regardless of whether any others were occluded. The relative positions of any recognised object parts could then be considered to support object recognition. The authors define 'parts' as being polynomial surfaces approximating closed, non-overlapping image regions that optimally partition the image in a Minimum Description Length (MDL) sense. Although the authors report success in their endeavour of recognising objects in cluttered and occluded scenes, an immediate observation with their approach is that almost all the information they therefore actually use is edge information. As discussed, largely uniform regions of an image carry little or no information. Sight is exclusively concerned with detecting relative changes across the values of its sensors and almost all of this information is conveniently conveyed by edge features (connected peaks in gradient space).

Another concern with appearance-based approaches to object recognition is that without high-level intelligent learning processes, an object will have to be examined under every possible set of illuminations that the concerned environment may pose, in order to be reliably recognised. Although these illumination conditions can be generalised from a limited number of sample reference conditions, this would still involve a lot of work in training. Even if all these possible view templates were to be imagined by the brain for learning, this would still require a lot of unnecessary processing if the same end conclusions could much more straightforwardly be drawn from an image feature set such as an oriented, spatial-frequency edge distribution.

One area of research where the aforementioned concerns with appearance-based representations have not presented so much of a problem is face recognition. Typically, whenever we view a face it is in an upright position, otherwise, the human face has a relatively distinct oriented appearance (throughout scale space) that can be used to unambiguously orient it to a standard upright fame. Our social encounters usually entail eye to eye contact and a full frontal facial shot is typically most characteristic of a person's appearance. Furthermore, there are many industrial scenarios that require biometric identification of a person's face, as viewed head-on. For the purposes of face recognition, in many scenarios, it could therefore be assumed that the object to be recognised would be presented from approximately the same viewpoint and under approximately uniform (top-down) illumination. Given that faces, in all their subtle guises, are a class of object that are particularly amenable to global appearance analysis, the appearance-based recognition of faces has stood at the forefront of object recognition research throughout the history of computer vision.

2.7 Human Face Recognition and Appearance Modelling

Primitive publications relating to machine-vision-oriented face recognition can be traced back as far as the 1960s [4][3]. The idea of parameterising appearance data in a low-dimensional sub-space predates the work of Murase and Nayar [48] for object recognition. Associated work is perhaps best characterised by Turk and Pentland's eigen-faces (1991) [28]. Eigen-faces are the principal linear components of a set of normalised, full-frontal face images. By treating each training image as a vector (with dimension equivalent to the number of pixels in the image), Principal Components Analysis (PCA) can be performed on the covariance matrix associated with the training set, to identify the main independent (linear) global modes of variation in the data. The number of independent (orthogonal) dimensions being that of the dimension of the image vector. The highest ranked principal components (eigenvectors) can then be exclusively used to encode the data set in terms of a limited number of eigenvalues. The main idea is that relatively few of these dimensions will be required to encode virtually all the information in the training set; the data essentially being compressed with a moderated loss of information. By projecting a new query image into this reduced-dimensional space, we are able to assess its similarity to any of the training examples in simple terms of 1D correspondence in each of these dimensions. This could make the difference between having to compare many thousands of pixels for each query face sample against just 50 reference vectors representing the learned dataset's principal modes of global variation.

Although approaches such as eigen-faces made some ground in automating facial recognition by computer, there were many problems associated with the process. Certain applications may allow for faces to be sampled in a strictly head-on manner, but this is generally an unrealistic condition upon which to base face recognition. Similarly, such approaches were too sensitive to illumination conditions and were not able to cope with subjects changing their facial expressions. A number of variants on the theme accordingly transpired throughout the remainder of the century. One such approach was Fisher-faces [59], a merger of eigen-faces and a derivative of Fisher's

Linear Discriminant (FLD), which aimed to maximise the ratio of inter-class to intra-class scatter in representation space. A number of training images for each subject across different viewing conditions were therefore required, with favourable results being obtained. However, it should be noted that the performance figures have been found difficult to repeat independently [88].

Active Appearance Models (AAMs) appeared in 1998 [69], proposed as a more robust framework for synthesising and recognising faces based upon their appearance. Although in keeping with the idea of assessing facial data in terms of linearised low-dimensional grey-level appearance, the key idea here was to allow a generalisable face model to adaptively align itself with a query face image, so that any assessments of similarity could be mediated more effectively in corresponding reference frames. Reference points are anchored at prominent features of the face, such as the corners of the eyes and mouth. This would loosen any opening constraints on the position and size of a face in an image, so that faces could be recognised in a more realistic range of scenarios. Beyond sampling such a detector at every possible image position, scale and orientation, this introduces the semi-independent subject of 'face detection'.

The process of generalised human face detection in images has served as a prominent area of computer vision research in its own right [82]. Many digital cameras, for instance, use generic face detection software to enhance photographs of faces. There exist efficient schemes for detecting faces in images. The Viola and Jones face detector [89], by prime example, automatically learns a minimal set of simple rectangular features acting as weak classifiers for face detection at various image positions, orientations and scales. Such simple, tailor-learned contrast features offers a high degree of invariance to general illumination conditions. There are suggested to be 4 such base-level detectors of main significance to low-resolution face detection in images. The down-sampling of any more complex images allows for any imaged faces to be detected very efficiently. The most obvious of these learned facial feature detectors are the ones around the eyes. The combination of a horizontal bar (brow) detector and a lowered dual- (darkened) eye-patch detector is typical of a human face viewed approximately head-on. Very simple rectangular contrast sensors can therefore be learned and used to quickly identify face type regions in arbitrary images. Integral images were notably used by Viola and Jones for efficiently scanning images with these base rectangular feature detectors at a range of scales in a single pass.

Procrustes Analysis can be used to align 2 face images into a common coordinate frame by minimising a sum-squared pixel difference optimisation function over rotation, scale and origin [77]. On the assumption that it is possible to effectively find faces through a range of appearances in images with minimal effort, Active Appearance Models (AAMs) have given rise to an abundance of associated research addressing various aspects of the representation and its implementation. Although faces have been the deformable object of primary interest to appearance model-oriented computer vision researchers, the techniques have also proved effective for recognising objects such as human hands [90] and for medical image analysis [69]. As far as AAMs have come in the last two decades, such appearance-based approaches to object recognition have fundamental limitations.

Appearance-based representations of this type are inherently limited by the apparent bounds of the exemplars in the training set. The methodology does not generalise well to unseen data. The calculation of the low-dimensional representation space, so called eigen-space, is based upon the shared set of characteristics observed across the learned examples. The more information rich dimensions of the training set are exclusively used to differentiate the given learning set. If we try to match an appearance model consisting purely of these dimensions to a novel face, with a substantially different appearance from everything in the training set, the fixed appearance model may be unable to align and deform with any useful effect and may be rendered useless for recognition and validation.

The training of AAMs presents another potential problem with the technique. Such learning schemes are typically based upon manually and precisely annotated reference points in training images. It is otherwise very difficult to program a computer to do this automatically with enough reliability. This makes it very difficult to get the system to autonomously learn, which can be regarded as a required competence of any prospective intelligent recognition systems. Low-dimensional global-feature sampling techniques also have the problem that the whole representation space requires recalculation when new objects are learned. Otherwise, an extra dimension may be required for each new distinct view of an object, so that the number of parameters required to represent a large library of 3D objects soon becomes unfeasibly high. Any mistakes made in the autonomous learning of an object's representation space could otherwise be catastrophic to recognition.

Although appearance models capture all the appearance information across an object in an image, they have no direct means to assess any outer-boundary detail.

Because it is impossible to predict what may appear directly behind an object, it is impossible to model the edge of a template in terms of appearance. Although some objects' more global appearances may clearly correspond to the same underlying shapes, appearance models cannot represent outer edge regions, which may be critical to inferring accurate alignment and recognition for many objects. In more up to date research, Cootes and Cristinacce [93] have shown that more accurate and robust facial matching and recognition can be achieved using a Constrained Local Model (CLM) search over individual face-feature appearance models. The local appearance models correspond to information rich features such as the eyes, nose and mouth, allowing for face detection to be more reliably performed though occlusion. The combined response is considered in an overarching projected 2D framework. The evidence that shape-constrained local feature template matching proves more effective for many recognition tasks than more global appearance matching highlights the advantage to be gained in basing the representation around the information rich areas of an object's appearance, i.e. its edge features. Cootes and Cristinacce also suggest that gradient, over appearance, could potentially be used as the basis for AAM-type recognition.

Active Appearance Models (AAMs) are an extension of Active Shape Models (ASMs) [52], which were originally used to model and recognise faces and other objects by virtue of their projected edge features. AAM-related research evolved to aid recognition of deformable objects such as faces whose appearance characteristics have been found to be particularly amenable to more comprehensive analysis. Although ASMs are less-useful than AAMs for face recognition, they can be used to cheaply align 2 samples before the more intensive appearance-based processes are actuated. Similar face feature reference points can be sampled as connected sets of associated edges instead of bounds of appearance. Although AAMs have proved exceptionally useful for face modelling, the inherent limitations of such linear approximation schemes have meant that other more specialised recognition methodologies have superseded AAMs in terms of wide-field practical face recognition prowess.

This introductory review of face recognition only scratches the surface of the associated academic literature. By 2006, computer vision-based face recognition technology was shown to be able to outperform human face recognition in a number of competences [99]. Artificial recognition systems were even shown to be proficient at distinguishing identical twins and have been deployed in a number of commercial applications. The ultimate performance of any face recognition system is governed by the levels of detail attended to. Any potentially visible features should be represented and learned as required. The problem can thus be engineered away in practical terms. People's infrared profiles, iris patterns and skin textures can be used in support of more traditional appearance-bound features for highly informed and accurate facial recognition.

2.8 Shape from Shading

It would be convenient if we could obtain a set of functions that related any projected surface shading to any corresponding 3D surface models so that surfaces could be identified and objects recognised. This would constitute a shape from shading and object-centred approach to 3D shape recognition, in line with Marr's original research. Although in restricted circumstances we can infer cues relating to relative surface depth from shading, the potential for arbitrary illumination and the apparent 'washing out' of any plain surface regions makes the process intractable in the general case.

The ability to infer surface dimensions from projected distributions of otherwise uniformly textured surface markings also supports reasoning about relative surface depth and orientation that may be useful for surface-model-based 3D object recognition. Active vision systems that project grids of structured light onto objects to infer the underlying 3D surface structure operate in similar regards.

The bottom line is that vision is potentially able to make use of all available data in performing 3D shape inspection and recognition tasks, but many tasks may require recognition of plain surfaced 3D objects from single 2D images, for which the only information visible may be a subset of their projected edge contours. Generic 3D shape recognition must facilitate projected edge feature detection and this forms the basis of the research herein undertaken.

2.9 Interest Point-Based Object Recognition

Given the advantages to be gained, in terms of object appearance invariance, by using intensity gradient features, the issue remains as to how best define, model, learn and recognise any such sample of image gradient information. One popular approach to 3D object recognition is to sample objects' projected appearances in terms of a set of 'point' reference features. Since we are aiming to take full advantage of any invariances to the model-imaging process to make the task of object detection as efficient and robust as possible, it is natural that fixed reference points such as corners be considered. Corners can be regarded as the minimal amount of information required to specify many objects' projected appearances, being potentially robust to factors such as location, rotation, illumination, scale and noise. In the simplest of terms, a corner can be recognised as the conjunction of 2 or more edges in an image, although point features can more generally be defined as any distinctive, well-localised features such as edge terminations or points of inflexion. Although a number of variants exist, the Harris corner [20] detector has proved popular in practice, essentially operating by detecting any points with a relatively high surrounding gradient in 2 orthogonal directions. 1 direction would otherwise indicate a continuous edge feature and no gradient a planar region.

A number of early attempts at object recognition were based purely around simple generic corner features [22] [24]. Any distributions of corners found in an image could be matched to any learned sets to identify any possible model projection matches. Huttenlocher and Ullman [24] showed that 3 corresponding points were all that was required required to align a 3D model with its projection in an image. Hash tables based around 3 sub-sampled reference points were typically used as look up tables to indicate set correspondence and recognition. As observed by Grimson [27], such processing paradigms were however very ineffectual in cluttered or noisy scenes, suffering from combinatorial problems and often producing very many false positive match hypotheses. Although many objects may have very pronounced features with definite corners, many do not and corner features alone proved to be too unreliable a means of identifying 3D objects in images. The natural extension to basic corner matching was to sample more specific information relating to the interest point and its local image region. An interest point could therefore be defined as being a well localised image feature offering a high degree of invariance to the imaging process.

One of the first significant approaches to recognition based upon key-point descriptors was Schmid and Mohr's (1996) [57]. Their research indicated that their techniques were fast and over 99% reliable in recognising a wide range of objects through scene clutter, occlusion, viewing transformations and limited scale across a database of over 1000 test images. This represented a significant step forward in object recognition research, serving as the first major attempt to recognise objects by virtue of their local appearance characteristics with invariance to any similarity transformations in the image (e.g. orientation). The basis of their approach was to sample a range of differential invariants from a set of key-points that were then used with a robust voting algorithm with semi-local constraints. An implementation of the Harris corner detector was used for key-point sampling and the point-based differential invariants were inspired by Koenderink's preceding theoretical research [18]. A hash table was used as a voting algorithm, with the constraint that matching sets of key-points must correspond in terms of local gradient direction. Evidence suggested that only a limited number of representative views of selected 3D objects were required to support full-view-sphere 3D object recognition. Beyond an apparently astounding display of competence in object recognition through a wide range of real-world viewing conditions, the authors concluded that their system was limited by robustness to scale change. As will be discussed, the nature of their system lends itself more to texture recognition than 3D shape recognition, suggesting inadequacies with their approach to the learning and recognition of plain-surfaced 3D shape.

Inspired by the success of Schmid and Mohr's approach to object recognition, David Lowe went on to introduce what has become the most prominent and popular technique for 3D object recognition; the Scale Invariant Feature Transform (SIFT) (1999) [72]. The key advantage of SIFT is that it operates throughout scale-space, providing an effective and robust means of matching 3D objects' 2D sampled appearances throughout a wide-range of scene and imaging conditions. The SIFT function operates by blurring an image with a Gaussian filter at a range of scales and by then calculating the differences between these images at adjacently sized scales. Maxima and minima points throughout this 3D difference-of-Gaussian image pyramid are then extracted as 'key-points' which are taken to highlight areas of the original image that contain significant information for object recognition. For each such key-point, a grid of gradient values and orientations is sampled from a surrounding Gaussian weighted image region at a corresponding scale. These grids are then down-sampled by a factor of four or eight into histogram bins, where each bin indicates the summed gradients of any pixels in that local region, at a range of (typically eight) orientations. A canonical orientation is then selected for each feature point in order to provide key-point orientation invariance. The resultant SIFT features thus loosely encode the expected local appearance of various purportedly significant image points at a range of scales. 128 histogram bins are traditionally used for each SIFT key-point. To recognise objects in an image, key-points are extracted from the image and are compared to those registered in memory using a nearest neighbour search strategy. Any clusters of at least three points corresponding to the same object view in memory are then taken as being object recognition hypotheses, which can then be verified by fitting associated geometric models to the image data.

Although the SIFT literature purports success in the task of recognising 3D objects through scale, clutter, occlusion and partial in-depth rotation, the validity of the processes involved is questionable in a number of respects. Whilst evidently an effective technique for the purpose, the SIFT algorithm involves recognising objects from a limited subset of features. Other information that might be critical to describing an object's shape and structure may be missed. This raises the consideration that perhaps a more powerful and practicable system would be able to recognise an object by directly utilising all available defining edge feature information. Furthermore, by assigning each key-point a local area gradient descriptor, this raises problems when identifying any boundary step edges adjacent to arbitrary backgrounds, where such local area information cannot be known a priori. With this in mind, the system's purported (image plane) orientation invariance would then be invalidated at step edge features, where a key-point's canonical orientation could not be predicted due to interference from arbitrary adjacent background features. Although evidently very useful for recognising textured objects, or matching textured image features, SIFT is of little use in recognising textureless objects by virtue of their defining boundary edge features.

The general success of SIFT for image indexing-based applications has inspired much associated computer vision research, resulting in a number of publications describing variants on the general theme. Any number of different parameterisations, feature types and associated processes are potentially viable. SURF (Speeded Up Robust Features) [100], for instance, speeds up the process by using integral images to detect interest points in a manner similar to that proposed by Viola and Jones for their face finding algorithm [89]. Furthermore, SURF uses sums of oriented wavelet responses across each point's local region instead of gradient orientation histograms, although apparently without any prominent advantage. GLOH (Gradient Location-Orientation Histogram) [92], by further example, aimed to improve the robustness and distinctiveness of SIFT by using more bins spread out in a log-polar distribution, i.e. with more weight for more central regions of the sample zone. PCA is used in GLOH to approximately halve the dimension of the descriptor to 128. In these regards, GLOH has been shown to be the most effective interest point descriptor of its type in terms of performance reliability [92]. Other approaches to 3D object recognition have utilised similar techniques to identify any local affine invariant 2D patches of objects' surfaces that can be assembled in a 3D framework for object recognition [94]. Observance that such techniques are only truly pertinent to recognising objects with textured surfaces, reintroduces the idea of using more global edge contour information for 3D shape recognition.

2.10 Edge Contour-Based Object Recognition

'Shape contexts' (2002) [96] have been proposed as a representation suitable for recognising objects by virtue of their extended edge distributions. Shape contexts are radially symmetric feature histograms that encode the frequency of edges passing through each bin surrounding a reference point. Although shape contexts can be oriented relative to the main reference line upon which they are sampled, each histogram bin only details how many edge feature pixels pass though it without reference to their orientation. The shape context can be regarded as a simplified, radially symmetric SIFT type detector that samples extended gradient peaks (edges) without regard to local orientation. The reference grid is typically more highly sampled in the centre than the periphery so as to be less affected by surrounding scene interference. Global coherence between a sparsely sampled set of matching contexts can be used to localise any corresponding reference models or views for match verification purposes.

Although an apparently useful shape descriptor, shape contexts are point descriptors with fixed reference grids. Many objects' projected appearances are composed of extended line segments, however, for which shape contexts suffer from the limitation that they can only be matched at approximately the same point on a line. Although shape contexts could be sampled uniquely at each end of any line, as discussed, real world imaging and the potential for arbitrary backgrounds very often breaks any line continuity. The drawback of shape contexts for edge recognition is that they need to be sampled at many points along a line, whereas other techniques, including the one presented in this work, may bypass this constraint. Although orientation is not sampled locally across the shape context's reference grid, this simplified representation requires less processing, which might be advantageous in certain circumstances.

'Shape Recognition with Edge-Based Features' [84] appeared in 2003 in another attempt to base object recognition on local edge information. The idea here was to have a square sample grid with local edge orientation sampled across each bin. A simple 4 bin grid with 4 orientation cells per bin was proposed as a coarse local area histogram that could be used to quickly identify any candidate matches in an image. Given any weak match hypotheses returned from this process, a higher-fidelity histogram (with all the dimensions doubled) could then be used to verify any potential matches. Although the authors focused their research around flat tubular objects such as bicycles and sports rackets with many extended edge segments, their descriptor again operates over a fixed reference grid so that to recognise an extended line segment any reference histogram may need to be compared at almost every point along the image feature. The inherent coarseness of the encoding supports recognition across large affine distortions and even across loosely similar objects from the same class, but at the cost that many false positives may arise. An expensive Hough transform type process is therefore required to accumulate any corresponding votes in parameter space. The authors concluded that their edge-oriented representation was a suitable complement for SIFT. The sparse nature of the structures of the tubular objects examined otherwise invalidated the SIFT (texture-based) approach to recognition, because of the potential for critically disruptive background interference.

Carmichael introduced another histogram-type point operator for the recognition of image edges in 2004 [83]. The idea here was to use a circular reference window with around 40 uniformly sampled edge 'probes'. Again for a given oriented point of reference, each probe serves to sample the edge density in its Gaussian weighted local neighbourhood. One of the main contributions of this work was the use of a cascade of apertures around the reference point, which involved progressively sampling more outward information given enough support in the more local vicinity. This was shown to speed up matching because any areas of the image that obviously did not match could be identified early and excluded from further unnecessary processing. The idea was to get the system to learn a set of descriptors from a set of labelled images and then to recognise any scene query edge points as being clutter or as belonging to the learned object. The presented work was however very restricted in its scope, limiting its application to loosely recognising chairs and step-ladders in cluttered images. Again, the point-based nature of the representation does not lend itself well to the task it sets out to attend to; edge contour analysis. The sparse and un-oriented nature of the representation also undermines its suitability for generic, large object-database oriented recognition. The authors point out that performance is dependent on manual tuning of any image edge parameters.

One approach to recognising 3D objects based upon sampling of their actual extended edge contours is that adopted by Chen and Stockman (1997) [65]. This work involved the higher-level detection and recognition of freeform objects' defining curves, as extracted from a Canny [14] edge map. The Canny step edge detector is the most popular edge detector in the computer vision community because of its relative reliability in extracting continuous, well-defined edge contours. To learn an object's edge-bound representation throughout 3D view space, the authors partitioned the view-sphere into local regions of interest (aspects) and, for each, sampled a well-defined edge contour map of the 3D object from 5 viewpoints clustered in a centre-cross formation. The 5 view samples' corresponding features were then aligned in the 2D view-space so that a 3D (2.5D) model could be inferred to use for projected pose estimation continuously across that region of the view-sphere. The view sampled at the centre of the cross formation was then used as the basis of 2D recognition by splitting its silhouette into a set of curved parts representing 'codons'. Codons are semi-closed curves delineated by concave points of inflexion along an object's silhouette; shapes ranging from domes, Gaussian profiles and mushrooms (or skewed versions thereof) are typically observed. An object's representative codons are then quantised in terms of a number of distinguishing characteristics; compactness, roundness, skewness, convexity and the sum of their normalised two second moments. Invariant part characteristics are then indexed to corresponding model aspects via a hash table, so that any similar curves found in an image can be used as part-based model match hypotheses. Coherence across any distributed features from the same object model aspect can then be used to infer a corresponding pose for the 2.5D aspect model so that it can be projected over the scene data to support global optimisation, verification and any prospective interactivity.

Although presenting seemingly robust recognition results across a range of 20 freeform 3D objects, Chen and Stockman's approach to 3D object recognition is too unrealistic to warrant it as the basis of a real-world-oriented vision system. The main problem is that the complex features they use cannot be reliably extracted in noisy cluttered scenes. Although noting that recognition can be inferred from a sub-set of feature matches, the authors go on to observe that there is too much confusion across different objects' aspects' features and they conclude by stating that their approach will not scale up to general object recognition.

A number of publications by Selinger and Nelson came to light by the end of the 20th century [67][68][75], detailing another curved edge reference feature-based scheme for the view-based learning and recognition of freeform 3D objects. Large-scale tests of their recognition system across a set of 24 complex objects through scale, clutter and occlusion indicated that recognition was robust across large complex scenes [68]. The authors suggested that these results were the best in the literature to their knowledge in 1997 [75]. The approach taken here was to base their representation about keyed context features. Observing that lines are too simple, full contours too unreliable and templates too pose sensitive, the authors suggested that curved contour fragments broken at points of inflexion were the most suitable base components of their recognition system. Any such curve fragment is assigned a fixed coordinate frame about its endpoints. By again taking advantage of a fixed reference frame, the system becomes largely invariant to image-plane scale, position and rotation. Once again, the system is only able to operate if its sampled features are intact, although the simple nature of their base descriptor makes the representation potentially more viable than that proposed by Chen and Stockman.

Given keyed edge contour segments as local projected shape descriptors, Selinger and Nelson proposed that each feature be associated with each other feature passing through a 2D box positioned about the start and end points of the curve segment. Any edge features sampled about the reference curve are then loosely associated by virtue of their relative projections. Two 2D key-curve invariants were used for initial match hypothesis generation, before correspondence between all the features in the local context was sought. These local area edge descriptors are then themselves loosely associated with each other in terms of their relative positions, scale, and orientations in accordance with any learned model views. Although detail specifying the exact nature of feature correspondence is sparse, the loose relational constraints are proposed by the authors as being very useful for recognising objects through significant natural or artificial distortion. The authors conclude with the observation that their approach is very similar to the abstract artistic movement of cubism (as popularised by artists such as Picasso and Braque), in that objects' representative elements are often omitted and distorted. Surely, such a level of abstraction can only hinder the process of object recognition when discriminating large numbers of objects in diverse real-world conditions.

Selinger and Nelson's system was presented in terms of a forced-choice decision regarding the presence of a known object in a given image. Across 24 objects with 1802 aspects and through a diverse range of imaging conditions the system achieved over 97% reliability in recognition. In cases where the system made mistakes, the correct match hypothesis was shown to typically be in the top few votes. This was taken in support of the distinctiveness and suitability of the representation for general object recognition tasks. Cases where object recognition failed were however generally identified as instances when an object's defining curves could not be extracted in full because of interference from background clutter leading to curve breakage. Although higher-level processes may ultimately be used to more reliably perceive and extract full continuous contours, there is no established solution and this cannot be performed with sufficient reliability for recognition. It was finally concluded that their representation was useful for class-based generalised 3D object recognition but did not scale well to tasks involving fine scale perception, as might be required to differentiate objects within the same class. The authors did not go so far as to implement any 3D model building or projection alignment and verification routines.

Prior to SIFT, Lowe also investigated the use of more general edge features for 3D object recognition tasks. In 1987 Lowe introduced the SCERPO 3D object recognition system [16], which was designed to model and match 3D objects in 2D images by virtue of any 'non-accidental' image features, limiting his proposal (for simplicity) to instances of collinearity, connectivity and parallelism between line segments. The SCERPO system operated under the 'viewpoint consistency constraint', which verified whether any detected invariant features were consistent with the distributions learned for objects from sampled viewpoints. Given sufficient support for a learned model view through these so called 'trigger features', Lowe proposed a top-down verification stage, in which any representative 3D models, at appropriate positions, orientations and scales, could be projected onto the original image, to confirm object occurrence and pose through global feature correspondence. Whilst highlighting various shortcomings with his account, Lowe comments that SCERPO is only a research and demonstration project, requiring much more work and specification for scaling up to a workable system.

In later work in association with Pope (1996) [54], Lowe extended the basic SCERPO architecture for a more complete and functional view-based object recognition system. To learn an object, their system required a series of segmented training images, displaying the object at a range of viewpoints. Following edge extraction, each image would be partitioned into a graph which described various image features and their spatial arrangement. The features which were proposed for use included intensity edge segments, certain groupings of such segments and corners. Groups of similar graphs were then automatically clustered together to represent characteristic views. The graphs for the images within each view set were then merged to form a single model graph incorporating probabilistic information regarding the expected distribution and significance of the enclosed features. Each object would therefore be assigned a probability distribution describing its range of appearance. Essentially, any groups of specified features detected in an image could be compared to the learned objects' view graphs in memory to indicate the likelihood of a known object appearing in the image. Given enough training images, the system was shown to be able to learn to recognise complex, real world objects in cluttered scenes. The methodologies and results outlined would be largely mirrored by the same authors' OLIVER object recognition system described in their 2000 publication [78]. The authors went on to indicate that their approach was however susceptible to problems involving reliable feature detection and grouping and suggested that a broader repertoire of features would be required by a more general system. Furthermore, the proposed system was still more of a demonstration project, lacking an indexing component and only being tested across a small number of objects.

Lowe went on to address issues of object indexing in work associated with Beis (1999) [74]. In this work, recognition was again based upon perceptual groupings of straight line edges, concentrating on chains of co-terminating segments and parallel segment groupings. Multi-dimensional feature vectors were derived from any feature groups corresponding to angle and length ratios. The authors suggested that any perceptual feature groupings could potentially be used within their framework and that each view of an object should contain a number of such feature groupings. Observing that hash tables were not robust enough for high-dimensional feature indexing, especially because a set of nearest neighbours was sought for each sampled image feature, a best-bin-first modification of a k-d tree nearest-neighbour search algorithm was suggested as the basis for feature recognition. To recognise an object, any specified feature groupings are extracted from an image and a set of nearest neighbour matching features are returned from the search tree. Any mutually consistent part feature hypotheses could then be ranked as potential model-view match hypotheses for verification. Verification was based upon an iterative least-squares algorithm for pose estimation, with each modelled line segment being verified as present if over half its length was matched to a corresponding image feature. Although the performance of the proposed system was deemed good for a handful of objects in images with a moderate level of scene clutter, the system was shown to be ineffectual in highly cluttered scenes. The authors suggested that beyond the limits of their primitive alignment and verification processes, the main problems with their system related to feature detection and grouping. A richer set of feature groupings were deemed to be required.

2.11 Alternative Recognition Methodologies

Another approach to recognise 3D objects in images is to use local area histogram descriptors without regard to feature correspondence. The idea here is that a set of specific types of feature detector can be pooled together, so that objects' appearances, either in whole or part, can be loosely characterised by defining combinations of such features. By sampling any query images or image regions with such feature sets, correspondence with any known objects or object views can be taken as recognition hypotheses. Although direct correspondence between any image and model features is bypassed, such techniques offer a potentially simple, fast and powerful means of recognising objects in sampled image regions. This means that no internal model is required and arbitrarily complex deformable objects may be recognised. Although the potential for ambiguous model matching can be moderated by having complex enough feature sets, such an approach would be ineffective for differentiating similar objects with shared defining features. Aside from any issues relating to the associated tasks of 3D model pose determination and verification, correspondence free histogram approaches to recognition are hindered by complications associated with window-frame correspondence and contamination from scene clutter. Mel's SEEMORE object recognition system [71], by way of example, uses 102 viewpoint-invariant non-linear filters to detect specific combinations and parameterisations of colour, corner, contour segment, blob or Gabor functions at several scales and orientations. The system was tested across a 100 object database with 12 to 36 training views for each object. Although the system's recognition performance was shown to be 97% reliable for objects presented against a plain background, recognition degraded significantly under noise and clutter. Furthermore, recognition performance was shown to be highly reliant on the colour channels of the histogram. As discussed, shape recognition should operate independently from colour.

2.12 Conclusions

This review has broadly outlined the history of the subject of computer vision-based object recognition. Many thousands of research papers have been published throughout the last 40 years, attending to various aspects of the problem. Although impossible to cover such a far reaching range of literature in full in such a short chapter, the most prominent, pertinent and up to date theories regarding 3D object recognition have been reviewed.

Insights into human visual processing have had great bearing on the evolution of the field of computer vision. To complete the discussion, it should be noted that recent studies [86] have suggested complementary roles for view and model-based object recognition processes in humans. Our ability to recognise objects via touch or from other communicated descriptions indicates the generic, far-reaching capabilities of our brains and perceptual systems. The research presented in this chapter does however make the case that single view-based object recognition is fundamental to visual perception because, in many circumstances, there is simply no information available upon which to base any bottom-up inferences of 3D structure.

It has long been established that human vision is heavily dependent on the analysis of projected intensity edge contours. Our ability to effortlessly recognise objects from fragmentary line drawings supports the hypothesis that object recognition is directly mediated by 2D projected contour-based pattern association. Furthermore, edge features have been shown to be very compact and powerful shape descriptors, offering a high degree of invariance to illumination and background clutter. Accordingly, image intensity edge information is commonly utilised throughout the computer vision community as the basis of object recognition.

Although established that machine vision-based object recognition should exploit edge information, there is no consensus regarding the most appropriate framework for object modelling and recognition. There are many competing schemes essentially attending to the same underlying problem. The work to be described in this thesis stands as an attempt to define and establish an appropriate basis for generic edge feature-based 3D object recognition. Pairwise Geometric Histograms (PGHs) are accordingly presented as a solution for recognising arbitrary projected oriented-edge configurations in images. PGHs are thoroughly reviewed in Chapter 3 in the context of any more directly associated literature.

The approach generally adopted for object recognition, as outlined throughout this chapter, is to identify a subset of features and feature groupings that are accumulated as object indexes. These methods lack any explicit consideration of information, invariance or statistics and issues of occlusion, clutter and background contamination are simply ignored in the hope that they will subside. PGHs stand out as a viable solution because they address all of these issues by design.

Chapter 3

Pairwise Geometric Histograms

Introduction

PGHs were conceived in the late 1980s as the representational basis of a semi-(biologically)-realistic neural network-based object recognition system [25]. The concept was further developed throughout the next decade [38][37][41][42][51][50][46][44][61][64]; ultimately offering a solution to the problem of recognising contour defined projected shapes in images. The concept was constrained by an assessment of the ways in which invariances can be used to restrict the number of distinct patterns required in order to learn and recognise a shape. This assessment concluded that an important factor in pattern recognition is the stability of the representation and that requiring this stability to be encoded in the representation (i.e. as probability distributions (see below)) precluded the construction of systems which were invariant to scale and out-of-plane rotation. PGHs were accordingly designed to fulfil the representational requirements of a real-world, view-based shape recognition system, thus offering a workable degree of invariance to the following factors:

- Illumination (in so far as using defining edge information in illuminated scenes provides)
- In-plane image orientation
- Image position
- Occlusion

- Background clutter
- Edge degradation (e.g. sensor noise, lack of contrast)
- Scale (although not directly, as described below)
- Scalability (able to support discrimination across very large shape data sets)

The PGH representation is proven to provide a complete and statistically optimal representation of 2D shape, suitable for use in a learning view-based 3D object recognition system [50]. The research described in the previous chapter suggests that no other published representation of projected shape possesses such characteristics. Indeed, PGHs can be seen as a tailored solution to the representation of projected, edge-defined shape.

This document goes on to give a detailed description of the PGH representation and previous associated research. Notably, there have been 3 previous Ph.D. theses published regarding PGH development. The first, submitted by Evans in 1994 [41], essentially introduces the concept of the PGH, analysing the stability and utility of the representation, before quantifying its use for 2D recognition and specifying a learning framework for neural network-based 3D object recognition. The second, submitted by Ashbrook in 1998 [61], extends the previous work, introducing a probabilistic Hough transform for object pose determination, methodologies for the recognition of scaled shapes and analysis of the capacity of the representation. Finally, Ashbrook deviates from the problem of view-based recognition, introducing related methodologies for encoding surface shape from (range found) polygonised depth images. The third thesis, submitted by Aherne in 1998 [64], investigated use of a multi-objective genetic algorithm to optimise PGH parameterisation.

3.1 PGH Format

PGHs are defined for linear edge segments, encoding the relative (perpendicular) positions and orientations of any other linear edge segments in a specified local image region (Figure 3.1). Since any curved edge can be approximated, to a desired degree of precision, by line segments [50], PGHs can encode arbitrary local image shape in a form amenable to computerised analysis. Shape recognition is performed by searching for correspondence between sets of learned representative object histograms and image sampled ones.

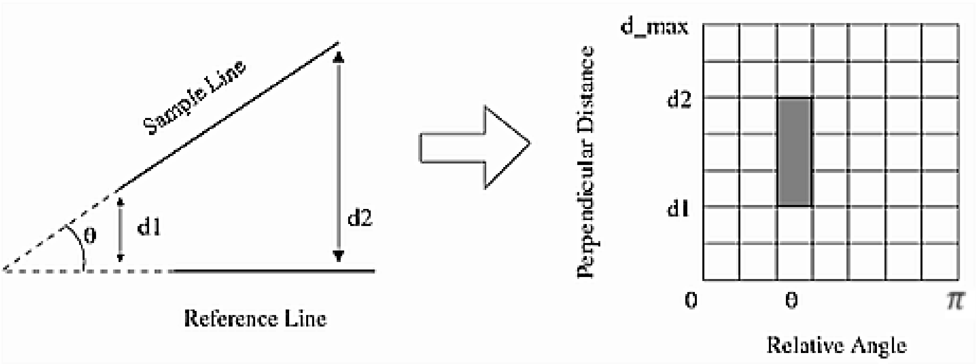


Figure 3.1: *The relationship between a pair of image lines can be detailed by the angle θ (defined at their point of intersection) and two perpendicular distances.*

As indicated in Figure 3.2, the use of perpendicular distance and orientation means that the sample line is effectively free to shift back and forth along the reference line within the bounds of the sample zone without changing the histogram entries. Though a second orthogonal spatial ordinate could be encoded to unambiguously represent any edge configurations (essentially amounting to a fixed edge template), this would require a fixed point of reference, which simply cannot be assumed under typical viewing conditions for edge features (in accordance with the 'aperture problem'). Moreover, the use of relative perpendicular distances is of key importance in accounting for fragmented portions of model lines. Edge fragmentation is commonplace across imaged objects and this potential for line breakage otherwise invalidates any approach based on distances from fixed reference points. By self-referencing the reference line in each PGH, the relative contribution of any fragmented line segments can be reliably encoded. A 2D shape's set of defining PGHs allows for full, unambiguous shape to be recovered [50] from a number of distinct geometric co-occurrences. This is because the data encoded is sufficient to solve for the spatial edge density distribution from a corresponding set of projection constraints. This completeness property validates the chosen representation.

Rather than being a fixed solution to edge-pattern encoding, PGHs represent a family of related metrics, essentially offering a tradeoff between specificity and complexity. At one extreme, one-dimensional histograms can be used to measure the spread of angles without reference to relative distances. Although providing representational

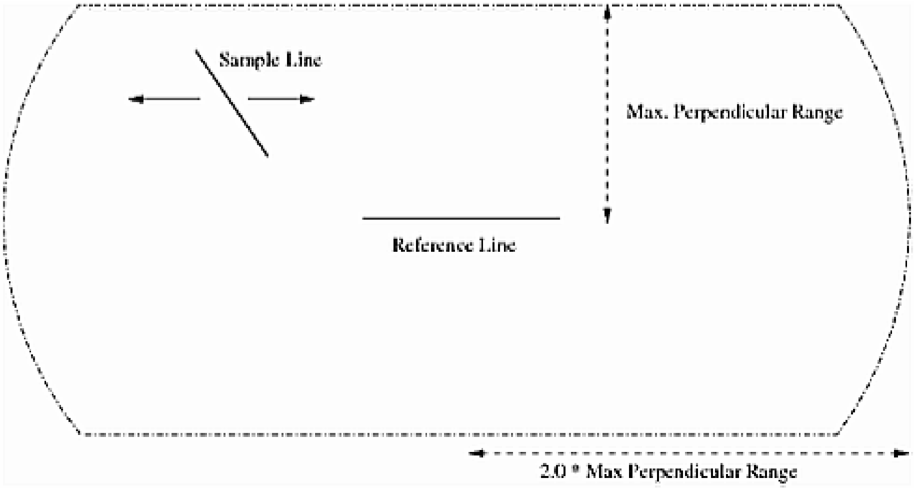


Figure 3.2: *According to the pairwise relationships detailed in Figure 3.1, sample lines are free to shift along an axis running parallel to the reference line. This is of key importance in encoding the relative contribution of fragmented portions of extended lines and extracting disjoint line segments across imaged scenes. The dotted border indicates the range through which lines are sampled for inclusion to the reference line's PGH.*

invariance to scale, excluding any relative distance information makes the representation too ambiguous for practical use.

Further consideration is required regarding exactly how the relative angles and distances are encoded between image lines. A full analysis is provided in [50]. One factor is that lines may be directed according to the contrast polarity of each referenced pair, so that relative angles may be specified from 0.0 to 2π, or equivalently, $-\pi$ to π. Although contrast information may be available and pertinent to some recognition applications, this project is concerned with generic 3D shape recognition, for which contrast polarity is typically unreliable or unavailable. This is especially true for the recognition of plain surfaced 3D objects under arbitrary illumination and against arbitrary backgrounds. Fully directed line-based representations are therefore discounted from present consideration. Relative angles can instead be defined relative to the point of intersection of the 2 (extended) lines, as indicated in Figure 3.1. If the (extended) sample line intersects the reference line, the reference line is split at that point and the two parts are treated independently, with the resulting information being integrated in the final histogram.

The simplest form of distance-based histogram is a mirror symmetric one, which simply encodes relative angle (0-π) against unsigned perpendicular distance, thus offering invariance to any mirror symmetries across the reference line. While useful in certain constrained circumstances and offering a very compact representation, this form of histogram limits the specificity (sparseness) of the representation and generally introduces too much representational ambiguity. Alternatively, the angle between line segments (vectors directed from their extended intersection) can be extended full circle to 2π. This doubles the size of the histogram, making it more discriminatory. Equivalently, the relative angle can be fixed from 0 to π and the perpendicular distances can instead be signed according to which side of the reference line (directed away from the point of intersection) they lie.

The PGH representation can be made fully (image-plane-) rotation invariant, by essentially adding rotationally equivalent contributions from each side of the reference line to the same histogram bins. The reference line is assigned a direction pointing away from the point of intersection with each sample line, allowing a signed perpendicular axis to be defined and relative angle to be inferred, as indicated in Figure 3.3. Any intersecting lines are split at the point of intersection and are treated independently. While sample lines are otherwise free to displace laterally without affecting the representation (see Figure 3.2), this new constraint limits such displacements up to the point at which the intersection meets the reference line. This is because the polarity of part of the reference line will swap as the intersection point crosses it, so that histogram entries will switch into the opposite distance axis. Such rotational invariance cannot however be achieved without some loss of recognition specificity. To overcome this, if required, the reference line can be assigned an arbitrary direction, so that PGH entries are assigned relative to which side of the directed reference line they emanate from. The trade off here is that each image line will require analysis in each direction independently.

Although using a number of local non-collinear reference lines to represent a shape already enforces a global coherence constraint, the potential for low-level ambiguity can be lessened by categorising which side of the reference line each sample line is directed towards. Although this division is built into the rotationally invariant PGH, directed reference line-based PGHs can instead be duplicated, with each half corresponding to sample lines pointing to one side of the reference line as indicated in Figure 3.4. A continuous representational flow is maintained across the 2 representation spaces because the reference line is again split at any points of intersection, with each part being treated independently. This form of directed histogram represents the limit of information that may be conveyed by a PGH in the general case (i.e. ignoring the polarities of each PGH's sample lines).

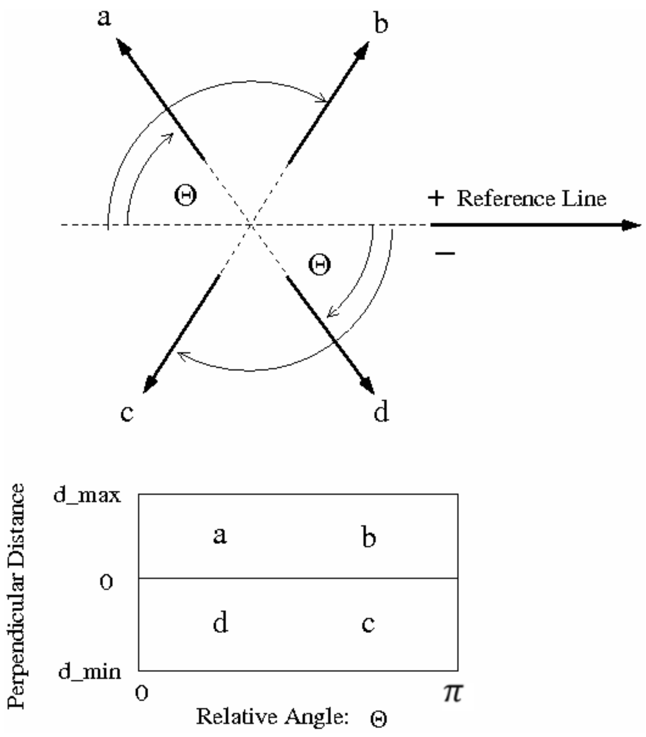


Figure 3.3: *For the image-plane-rotation invariant PGH format, the sample and reference lines are directed away from their point of intersection, with relative angles being entered into the histogram as indicated. If any extended sample lines intersect the reference line, the reference line is split at that point and each part is processed independently. The diagram can simply be rotated by 180 degrees to indicate the assignment of angles for reference lines pointing in the opposite direction.*

Although previous research [41][61] describes use of circular reference regions centred on the reference line's midpoint, it is now deemed more appropriate that these regions be extended lengthways to account for extended reference lines and possible fragmentation. A circular region of twice the diameter can therefore be defined, with a perpendicular cut-off at the prescribed pixel limit (see Figure 3.2).

Because PGHs are particular to single straight line segments, 2D shapes are represented redundantly by sets of overlapping localised PGHs. Such a local part-based representation is useful in terms of recognising occluded objects, as any competent recognition system should be able. Otherwise, if only full models are used, any very localised correspondences will likely be obscured against coincidental feature matches across other complete object models and scene clutter. Although single histograms will be susceptible to the ambiguity issues just discussed (lateral displacement), fragmented portions of objects will still be represented by local clusters of PGHs, albeit reduced numbers, so that these ambiguities can be disregarded due to the recognition constraints imposed by the other reference lines.

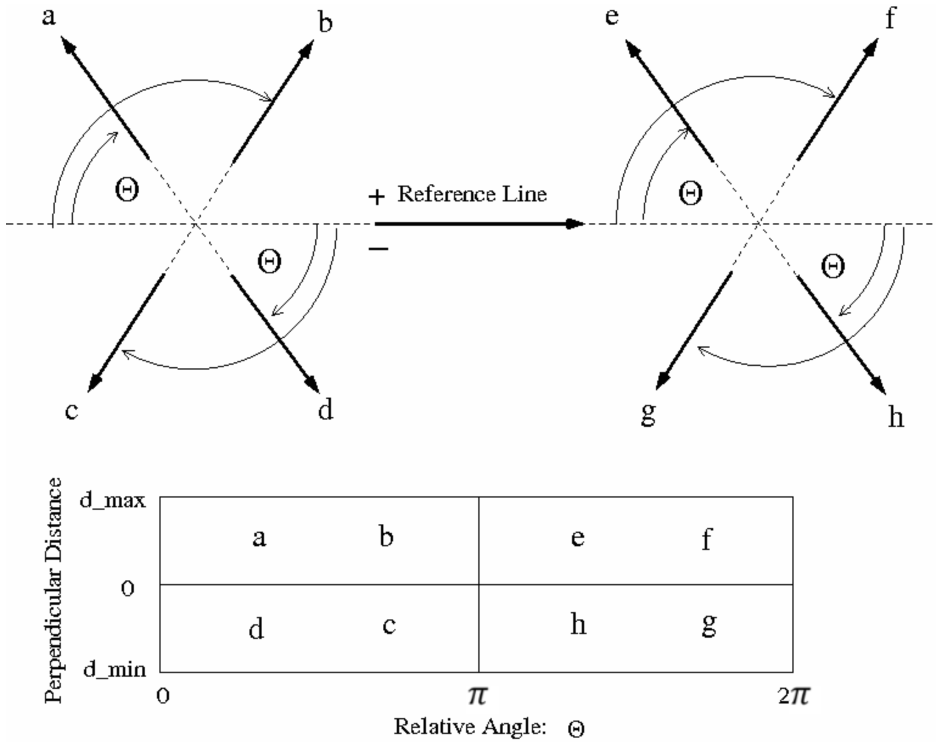


Figure 3.4: *For the directed reference line-based PGH format, the reference line is assigned a fixed direction so that each image line requires matching in each direction. In the simplest case, only sample lines a-d and the PGH from* 0-π *are required. To enhance the discriminability of the representation, the adjacent PGH ranging from* π *to* 2π *can be used to separate sample lines for which the reference line points towards their point of intersection (e-h). The reference frame can again be rotated by 180 degrees to indicate the same PGH bin assignments for reference lines pointing in the opposite direction.*

Considering that many PGHs may be required to represent an object, the issue of feature saliency was discussed in [41] with regard to streamlining the recognition process. Instead of registering a PGH for each and every edge feature, a subset of key features could instead be used, according to some saliency criterion. Unfortunately, it is very difficult to define a measure of saliency in the general case and there can never be a guarantee that any significantly reduced subset of features will be sufficient for recognition under difficult viewing conditions. Line length, for example, was considered in [41], but it was observed that for many shapes, the spatial distribution of the longest features would be very uneven.

Previous research [41] has further discussed the possibility of recognising 2D shapes via single global histograms that accumulate the histogram evidence from each line segment. Although these histograms prove capable of recognition of fixed segmented 2D shapes, they are very sensitive to factors such as scene clutter, are devoid of any completeness properties and therefore offer very little practical utility in terms of general shape recognition.

Although the mirror and rotationally invariant PGHs offer advantages in terms of recognition processing costs, the enhanced discriminative powers of the bigger directed histograms have proved advantageous in previous research regarding 2D object recognition. The purported completeness property of the representation is also only strictly applicable to directed PGHs.

3.2 PGH Bin Entry Assignment

Beyond whichever PGH format is adopted for recognition tasks, consideration is required regarding the number of histogram bins used to sample the angle and distance axes. As explained below, this is related to the measured information available in the data, which is encoded via the degree of cross-bin blur applied to entries. At a more practical level, PGH parameter selection essentially becomes a trade-off between specificity and stability. A key consideration regarding histogram quantisation is that of application. If the computer vision task concerns very fine scale discrimination between similar shapes, then a high level of detail will require encoding. From a computational perspective, it is however desirable to limit the number of histogram bins, so that an excessive amount of memory and bin comparisons are avoided [*].

Since we are dealing with multi-view-based vision, it is also desirable to allow some form of generalisation to limit the number of views required and to account for any measurement errors such as those introduced from line quantisation or sensor error. It is also critical that the representation is stable, so that small changes in input result in similar smooth variation in histogram form. This is especially important when attempting to map out continuous shape manifolds for view interpolation. Although histogram binning inherently provides some degree of blurring, a further stage of cross-bin blurring is required to avoid discontinuous shifts in the representation

[*] Typically, with images of approximately 0.25 million pixels, PGHs are sampled with a 50 pixel perpendicular bound in each direction, with sufficient discriminatory resolution and stability being provided by splitting each distance axis into 10 bins and the angle axis (from 0 to π) into 32 bins.

due to bin quantisation. This strategy is justified by a probabilistic interpretation of histogram structure as representation of the level of uncertainty in local geometric structure (see below). Higher levels of bin blurring therefore allow the representation and recognition of deformable objects, although at the expense of decreased discrimination. Though the visual change induced by an out-of-plane rotation is a form of deformation, the recognition of more general forms of deformable object is beyond the scope of the current research.

The process of PGH bin blurring is described in [41]. First considering the angle axis, it was observed that the required degree of blurring is related via error propagation to the degree of curve linearisation imposed by the system. If the curve linearisation factor is low, so that relatively few lines are used to represent curves, then a higher degree of blurring will be required to account for shifts in relative angles due to arbitrary curve-line partitioning. A Gaussian distribution approximation was temporarily implemented. A wrap around function was also introduced, so that any blurring at extremal bins (around 0 or π) was wrapped around to the bins at the opposite side of the angle axis. Similarly, it was noted that determining the distribution of perpendicular distance bins was problematic. In this case, a simple rectangular blurring function was implemented for each entry.

In the current state of operation, bin entries, in each regard, are blurred in the form of isosceles trapeziums, i.e. rectangular central regions with sloped sides, with associated widths being specified in each case. This is evidently a simple yet effective solution to the problem. Although inherently related to histogram bin width, factors affecting choice of width of blur are described in section 3.4, which details the stability of the representation.

The overriding interpretation for PGH construction is that the histogram is a quantitative representation of the frequency of geometric co-occurrence of edge features, with the constraint that the sum of multiple entries from a fragmented line must be directly proportional to the total entry for a corresponding un-fragmented line. Fundamentally, PGHs can be considered as the integration of equally weighted entries from individual edges (as might be supported in biological systems), but their invariance to line fragmentation allows us to construct them more rapidly (on a serial computer) from linear approximations to edge boundaries. As a final point, a system which were to construct the histograms from individual edge samples would automatically generate blurring, as described above, as a consequence of measurement noise.

These histograms would also possess Poisson sampling characteristics, which forms the basis of the matching scheme now to be described.

3.3 The Bhattacharyya Match Metric

As histogram formation is defined as a counting process, normalised PGHs can be regarded as sampled conditional probability distributions (of geometric co-occurrence); with the idealised view being that each bin is equivalent to a Poisson distributed random variable [50][45]. Standard statistical tests can therefore be drawn upon as the basis of histogram similarity for recognition purposes. The standard method for histogram comparison is the χ^2 (chi-squared) test, which is defined as a sum of squared differences operator, normalised by the expected measurement error. As discussed in detail in related work [45], the χ^2 operator is an approximation to 'Fisher's exact' test and is only typically valid across small differences in pattern space. Alternatively, the Bhattacharyya metric is proposed as being a more appropriate means by which to compare two histogram distributions because it embodies a square root transform [45]. Only with such a variance normalising transform can a measure of similarity be reliably and uniformly assessed across large distances in pattern space as a Euclidean distance. Indeed, the Bhattacharyya metric can be regarded as an exact form for the comparison of probability densities (Appendix A.1).

The Bhattacharyya metric performs a dot product of the square rooted PGHs (a suitable construction for computation in neuronal tissue). We can intuitively consider this as returning the cosine of the angles between the two sampled vector spaces, i.e. 1.0 for identical (unit-normalised) histograms. Notably, maximising the Bhattacharyya metric is essentially equivalent to minimising the Matusita distance measure and offers a computationally efficient and stable means of assessing PGH similarity.

$$D_{Bhattacharyya} = \sum_{i}^{n} \sqrt{a_i}\sqrt{b_i} \tag{3.1}$$

where n is the number of histogram bins.

However, the Bhattacharyya metric is specifically appropriate for assessing bin-to-bin similarities, for which sampling noise is the primary source of interference (cases of scene clutter and occlusion are discussed below). Since visualised PGHs will typically distort laterally with in-depth model rotation, this means that although two PGHs may be topologically similar, despite blurring and bin quantisation, their corresponding Bhattacharyya overlap scores may be relatively very low. I.e. any zero valued entries in the histograms will eliminate any non-zero values in the other PGH's corresponding bins. Thus, no distinction is made between systematically spatially distorted and completely disparate shapes. In order to get the best out of this matching process it is therefore necessary to model view induced (correlated) changes in histograms appropriately. This emphasises the significance of an interpolation procedure for PGH reconstruction throughout view-space, which stands as a key issue in current research.

3.4 Stability of the PGH Representation

One immediate potential concern with the PGH representation relates to the use of line segments to approximate curved edges. Curve linearisation is performed by iteratively subdividing the curve into linear segments, so that the ratio of perpendicular separation from the midpoint of each line to the curve to the length of the line is kept below some preset threshold. It is therefore critical that the PGH-encoded shape representation is largely invariant to shifts in the placement of linear sections across curves, as would be the case when linearising arbitrary fragments of curves in images. Such shifting would alter the orientation of the reference line and any related bin assignments. As discussed, the process of bin widening and entry blurring can be used to compensate for such factors, as proven experimentally in [41]. Essentially, a trial and error process is used to set a curve linearisation factor appropriate for the specific form of PGH used in the specific recognition application. There will otherwise be a limit of linearisation at which PGHs become negligibly indistinguishable, according to Bhattacharyya-based similarity. By constructing any reference models at this scale of linearisation, we are able to limit any effects of relative difference due to line segmentation, although at the potential cost of having to generate a lot of PGHs for curved regions. The bearing of these factors on determination of angle-bin blurring is further discussed in [50].

The other issues that may affect the stability of the PGH representation for recognition are shape fragmentation (e.g. due to a lack of contrast or occlusion) and sensor error. Given the proposed use of the Bhattacharyya metric for PGH similarity evaluation, we are able to directly evaluate any such effects in terms of their effect on the match score.

Edge fragmentation is very common in typical images of objects. Aside from occlusion factors, this is typically due to a lack of contrast across certain edge regions as a result of conspiring illumination or background conditions. Evans' thesis [41] quantitatively illustrates that the Bhattacharyya similarity metric (D) degrades linearly and thus very stably, relative to the proportion of remaining shape

$$D \rightarrow D(1-k) \tag{3.2}$$

where k is the proportion of missing data, i.e. a fixed quantity. We therefore expect that occlusions and missing regions of an object will not invalidate the use of this similarity measure in competitive matching strategies - particularly if an effort is made to robustly integrate the information from multiple histogram matches in an inclusive (i.e. non categorical) manner. The linear nature of this response can be seen once again as being related to simpler template-based counting strategies, albeit with additional edge orientation sensitivity.

Scene clutter is another commonly occurring factor that the representation must be able to cope with. Such clutter can manifest itself as interference from other objects in the scene or other lighting induced artefacts such as shadows. Evans' thesis [41] indicates the resilience of PGHs to these factors by measuring the effect on recognition of adding increasing levels of randomly oriented spurious line segments across reference images. Experiments showed that matching of geometric feature distributions is theoretically robust to the presence of arbitrary spurious image lines. It was however observed that such artificial *line noise* was unrealistic and that actual line interference may be much more correlated with image structure, e.g. shadows. Because of the ambiguity of the PGH representation with regard to parallel localisation of features relative to the reference line, any spurious local features that are parallel to template lines should however present more of an adverse effect on match scores by overly weighting any specific histogram entries. The robustness of the representation should however be enough to still support reliable recognition. In later work [44], match scores (D) were shown to degrade as:

$$D \rightarrow \frac{D}{1+a} \tag{3.3}$$

where a is the proportion of uncorrelated scene clutter. Again, for a fixed pattern of background clutter, this process is found to introduce approximately fixed reductions in match score, which do not invalidate the representation's use during competitive matching strategies. The use of localised reference regions helps to further reduce the effects of any such interference [44].

Finally, Evans' thesis [41] discusses any adverse effects that sensor error may have on match reliability. Having noted that actual sensor error, i.e. image noise, would typically only affect detected edge locations by very small amounts, up to approximately 0.5 pixels, the analysis considers the cumulative effects on the detected positions and orientations of edge segments from additional factors such as line fragmentation and curve approximation (as just discussed). It is observed that these effects are proportional to line separation and inversely proportional to histogram resolution and the level of blurring used. Without such quantitative encoding of the level of uncertainty in relational geometry, it is reiterated that match scores will fall off sharply and irregularly.

3.5 Recognition Across Scale

As should now be clear, because of the incorporation of distance measurements, PGHs are directly suited to the recognition of (2D) projected shape at predefined image scales. The scale invariant, one-dimensional, orientation-only encoding PGH has otherwise been shown to be too ambiguous for practical use [50]. A recognition system must however be able to recognise objects as they move closer to or further from the camera, at a range of projected scales. PGHs are partially invariant to shifts in scale and will accordingly degrade relatively smoothly and consistently, so that collections of adjacently scaled histograms can be used to offer a degree of invariance across extensions of scale space (see Figure 3.5).

In the extreme, the relative scale of an imaged object may occur from a pixel to the full size of the image, ignoring partial object visibility at larger scales. For a range of such smaller scales, an object's representation will be composed of clusters of pixels, e.g. 5*5 or 15*15, which may offer little or no discernible information regarding unambiguous object identity. At lower resolutions, especially for intricately structured objects, any edge-based descriptors may be vastly different from those at full resolution, as sets of edges are effectively blurred together in pixel bins. Furthermore, any errors on edge-based measurements will become more pronounced as scale reduces. For the moment, research is focused on identifying objects at a range of scales for which objects' edge-based appearances are relatively stable, e.g. for projected object diameters of 50 through to 200 pixels for relatively simple object structures. The use of discrete models through scale space does however offer the potential for adaptation of the models at each scale according to any feature visibility constraints.

Changes in the scale of a PGH-encoded edge pattern are represented as uniform compression or stretching of distance entries. Ashbrook's thesis [61] and [46] provide a thorough analysis of these issues, illustrating that PGHs form smooth although globally non-linear trajectories throughout pattern space relative to image scaling. It is further shown that these shape-specific trajectories can be approximately modelled by a sample of consecutively scaled PGHs. This is performed by taking a PGH at the lowest required scale, then iteratively adding new reference nodes at the maximum inferred scale difference, so that a tolerable error is continuously maintained. Any fall offs in recognition accuracy at mid-sample points can be accounted for by the noise models and associated blurring functions discussed in previous sections. Although it was initially proposed that scale be inferred directly from the scale of the nearest corresponding scaled PGH [61], it was noted that a uniform scale error would be introduced. Instead, the scale-oriented methodologies were proposed as input to a probabilistic Hough transform procedure, to support assessment of observed object scale.

3.6 The Probabilistic Hough Transform (PHT)

Typically, an object recognition application will return a list of hypotheses of possible model to scene correspondences. Given determination of which edge lines correspond to which objects', it is relatively trivial to infer the corresponding image position, orientation and scale of any so discovered object from line correspondences and optimisation of any transformation parameters. This allows for the position of the object

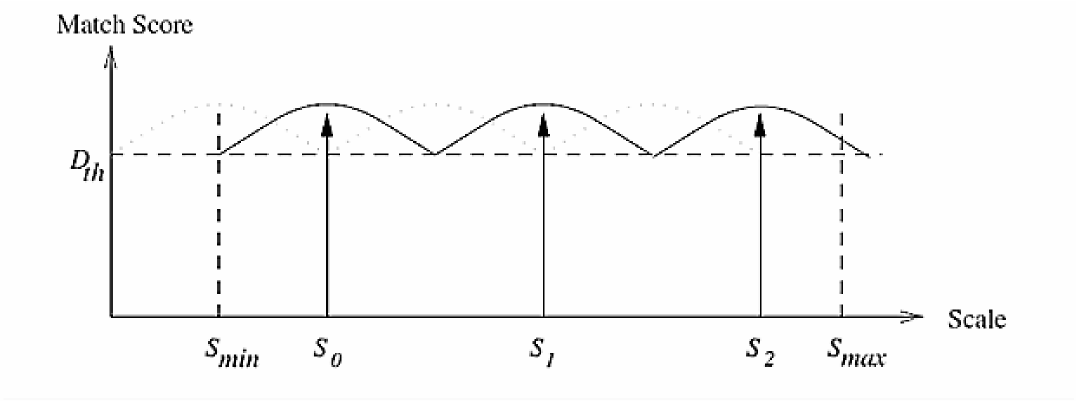


Figure 3.5: *The above diagram indicates how individual histograms may be combined to offer invariance across a predefined range of scales up to a tolerable degree of error. The diagram is reproduced from [46].*

to be estimated relative to the camera and, potentially, for any recognised edges to be removed from subsequent scene analysis. However, as discussed in [41], the problem is rarely this simple and there will typically be much ambiguity regarding which scene edges relate to which objects at which parameterisations. This is typically due to interference from noise, occlusion and background clutter. For these reasons, the generalised Hough transform was proposed in Evans' thesis as a robust means to identify and parameterise well-supported model match hypotheses in noisy conditions, albeit without any initial appreciation of scaling.

The generalised Hough transform was introduced to the computer vision community in 1981 [10]. The procedure essentially operates by amassing discrete votes for each model match hypothesis, according to the hypothesised location and orientation of the considered model (at a fixed scale for now). Consistent sets of votes from related edge segments should therefore produce readily detectable peaks in representation space that can be used to indicate the most likely parameterisations of any model matches. The relative intensity of the peak will be indicative of the amount of support for that hypothesis. Spurious edge matches should otherwise distribute their votes randomly into a noise field that can be disregarded from further consideration.

In the initially considered case of 2D recognition [61], 3 parameters are required for object parameterisation; 2 for location and 1 for relative image plane orientation. Votes are cast from pairs of edge segments for particular models. An object's centroid is typically used to detail the location of the object. Although such a point can only be represented along a line running parallel to the reference line in PGH form

(Figure 3.2), since we are using pairs of lines, supposedly from the same model at the same scale, the centroid's position can be determined at their point of intersection. The relative orientation parameter can be subsequently inferred from the associated orientation of the hypothesised model. These methods were proven able to support reliable determination of model match parameterisations without need for subsequent global model optimisation in noisy, cluttered images, although only with regard to fixed scale object recognition.

Ashbrook went on to propose the use of a Probabilistic Hough Transform (PHT) [61][46], serving to account for any inferred errors on the estimated location and orientation parameters. Relating the techniques to those of maximum likelihood statistics, this would come to introduce a far more robust and accurate means of inferring the parameterisations of any valid model match hypotheses. Indeed, the process is essentially equivalent to performing a robust least squares fit of the projected model data. Crucially, the PHT also incorporated scale, returning the most likely inferred scales of any parameterised model matches detected in the scene.

The Probabilistic Hough Transform (PHT) process is discussed in detail in [61]. In brief, a PHT is performed for a specific model, given enough support from the PGH matching process. A (2D) location PHT is initially performed. A single entry (a conditional probability) is initially made in the PHT for each pair of scene lines that are in reasonable agreement regarding the position and scale of the model with regard to any associated measurement errors. As previously discussed, each line will constrain the model's centroid to lie along a straight line in the image, so that each pair of lines hypothesises the position of the centroid at the point of intersection of these constraint lines. The errors on the positions of any line endpoints are assumed to be Gaussian distributed, so that error propagation can be used to infer the likely error on the point of intersection. The residual errors from a real-world scene are shown to validate this process. Noting that the segmentation and scale errors are independent, the segmentation error function may then be convolved with the scale error function allowing the probabilistic entry to the location PHT to be determined.

As detailed in the previous section, sets of consecutively scaled PGHs are used to represent edge distributions through scale space. Because the errors are constant for each scaled PGH region, in effect, a rectangular bound is placed on the location of the hypothesised centroid position. Since the scale of the model must be the same relative to each line segment, the position of the centroid is further constrained to lie on an equal constraint line passing through this (skewed) rectangular region (Figure 3.6). The error distribution relating to line segmentation can therefore be convolved with this constraint, allowing the most likely position of the hypothesised centroid of the scaled object to be probabilistically determined for entry to the location PHT.

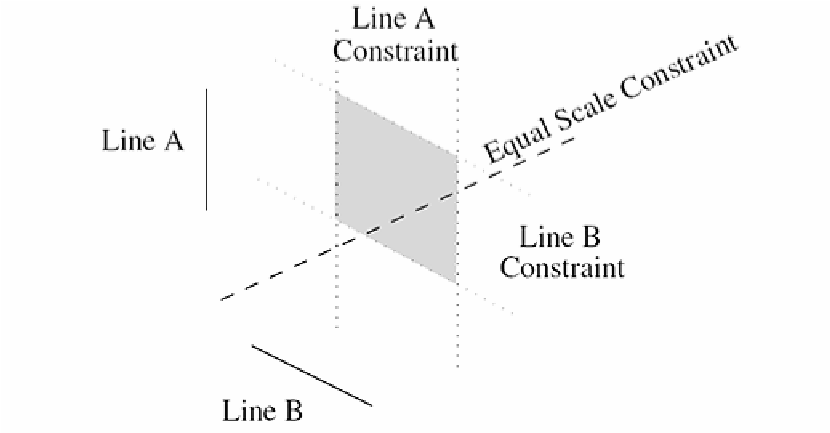
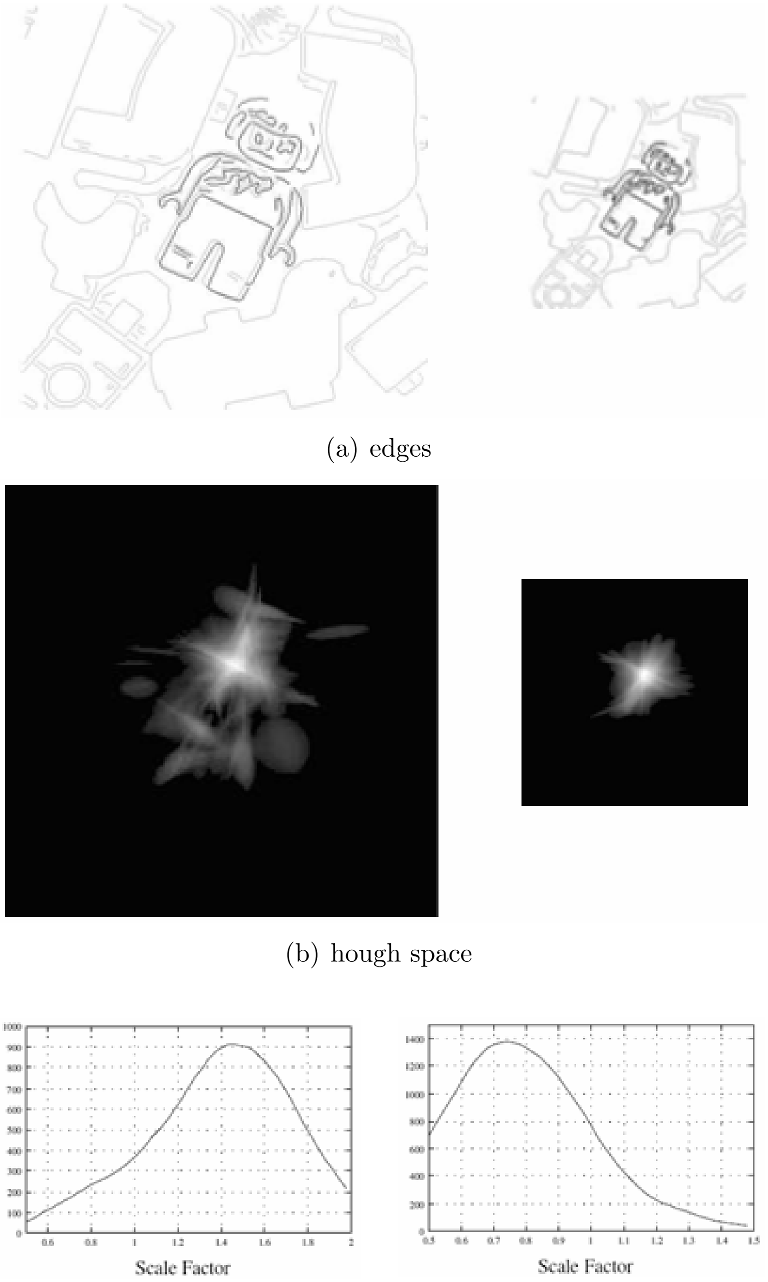


Figure 3.6: *The possible positions of a shape's origin relative to two reference lines are constrained to lie on the dotted line within the shaded region [46].*

Once the occurrence and location of a hypothesised model match are determined from a PHT, subsequent, single parameter PHTs are used to explicitly determine the scale and orientation of the match. Votes are cast from each scene line in accordance with the proposed model match. The scale parameter is simply determined from the perpendicular distance from the model line to the centroid, relative to the same distance in the image frame. The orientation is voted for according to orientation differences between scene lines and corresponding model lines.

The PHT has been shown to be a robust method of detecting, localising and parameterising any 2D model occurrences in cluttered and occluded images. Examples of location and scale Hough transform spaces are provided for a pair of scaled 2D objects in Figure 3.7. The PHT method is able to cope with multiple identical objects in a scene, but does however strictly require rigid 2D shape templates. Any deviations from rigidity will produce fragmented localisation peaks and sub-optimal object detection. In terms of scale-based analysis, current research instead investigates the potential of sampling the scene at multiple scales, in a manner similar to SIFT [72], rather than having to store multiple scaled entries for each object view.



(a) edges

(b) hough space

(c) scale estimate

Figure 3.7: *The above diagrams (taken from [46]) show 2D probabilistic Hough transform results (b) for the expected locations of the 2D projected models highlighted in the corresponding images (a) (at scales of 1.5 and 0.75). Although the peak positions (bright white highlights in (b)) are clearly identifiable, the images indicate the effects of variable sensitivity through scale space. The lower 2 diagrams (c) show the corresponding 1D Hough transforms used to determine the scales of the inferred model matches.*

3.7 Capacity of the PGH Representation

The suitability of the PGH representation for application to real-world-oriented computer vision recognition tasks also critically depends on the capacity of the representation, in terms of its ability to differentiate distinct shapes across very large datasets. As discussed previously, the capacity of the representation will be proportional to the resolution of the histogram and the levels of blurring used. It is however, regardless, very difficult to define such a measure of capacity. It has otherwise been shown [44] that processing requirements scale no worse than linearly with the number of stored models. Recognition reliability can even be shown to improve with larger model databases because votes from any spurious scene features' PGHs become more diluted in representation space, so that any consistent model match hypotheses stand out more distinctly.

One approach to capacity estimation is to estimate the area of a typical circular hypersphere surface patch representing a distinctive view and accounting for errors, relative to the surface area of the positive quadrant of the hypersphere - according to standard PGH dimensions (e.g. for a 20 by 64 bin (dual directed) PGH, the hypersphere is 1280 dimensional). The problem with this approach is that projected structural edge patterns are likely to be very correlated and the hypersphere is highly unlikely to be uniformly populated. Accordingly, associated research has focused on identifying the 'effective dimensionality' of the representation [44][61]. Deriving results from mismatch probability estimation curves for limited sets of objects, the effective local dimensionality of PGH encoded data is estimated at 16 [44]. This result is used to estimate that PGHs are typically capable of storing between 10^8 and 10^{13} distinct histograms. Given these numbers it is relatively easy to see why histogram matches are likely to be quite informative across large datasets, even when match scores are degraded by significant quantities of clutter and occlusion.

Although the proposed fragmented pattern encoding scheme (e.g. sampling PGHs for a 50 pixel perpendicular distance from the reference line) does mean that there are likely to be localised recognition ambiguities, these bounds on the capacity of the representation suggest that overall, multi-segment matching strategies should be robust to ambiguity in terms of application to real-world oriented learning recognition tasks. As discussed, PGH scope and resolution can otherwise be enhanced so that arbitrarily complex patterns can be differentially encoded.

3.8 PGH-Based 3D Object Recognition (Previous Research)

It should now be clear that PGHs are a valid solution for representing 2D projected shape as the basis of a multi-view-based real-world-oriented 3D object learning recognition system. In the simplest terms, given that we now have a system able to recognise views of objects, each object of interest could be finely and uniformly sampled around its view-sphere and a nearest neighbour matching strategy could be implemented to recognise any likely model occurrences. Although this would constitute a 3D object recognition system, there are many issues that require further attention in support of a more practically viable and physiologically plausible view-based recognition system. Notably, an inordinate number of views would be required by such a system at the accuracy levels required, whereas, many views may be redundant, supporting potential application of interpolation mechanisms to optimise the required number of views.

A PGH is essentially a high-dimensional vector, with dimensionality equivalent to the number of histogram bins. As such, the stability of the representation means that clustered views of an object should describe smooth low-dimensional shape manifolds in this high-dimensional vector space. This is strictly only true for objects' aspects, for which a continuous set of visible features is maintained. Otherwise, discontinuous sets of these manifolds will be formed. By further normalising any PGHs, we can constrain any such shape manifolds to lie on the surface of (the positive quadrant of) the unit hypersphere. This processing allows objects to be represented as (possibly sets of) continuous hypersurfaces, which can be extracted and analysed as much lower dimensional shape manifolds. The problem at hand is therefore appropriate learning, modelling and recognition of these object-specific, possibly fragmented, shape manifolds.

The utility of self organising neural networks has previously been investigated to perform this PGH to object identity mapping [25][37][41]. Appreciating that some objects' hypersurfaces may intersect and may be susceptible to interference from noise, the proposed neural network architecture was composed to return probabilistic estimates of object recognition hypotheses, according to approximation of Bayesian a posteriori probabilities. A 3 layer network was proposed for use.

While the cited neural network related publications specify a framework for detecting the presence of an object in an image without regard to the recognised 3D pose of the object, the methodology could be extended to learn specific views of objects (i.e. treating object views as distinct objects) so that recognition of such a view would indicate the approximate recognised pose of the corresponding object.

A shape representation layer is initially used by the network to automatically distribute reference units around the domain of hypersurfaces according to the distribution of PGH vectors sampled during training. Resultantly, a Voronoi tessellation of learned hypersurfaces is formed (a nearest neighbour mapping), so that the best matching node will be 'fired' upon presentation of a novel PGH vector in a 'winner-takes-all' manner. The expected output firing rate of a node is thus the probability of the node being the best description of the data. A Bhattacharyya distance metric is used as the basis of the discriminant function (in accordance with an assumption of frequency coding) and each node's weighted vector parameters are updated in a probabilistic manner, to account for those of the newly matched vector during training. This training process reduces both noise in stored patterns and the problems of missing or additional responses, via a process of 'resonance' (as per Grossberg's physiologically motivated ART network). In this way, subsequent responses of the network are robust to effects of occlusion and clutter in the viewed scene. A second layer (computationally identical to the first) is then used to learn the pattern of responses from the first, integrated over the features of the entire scene, in a manner consistent with a requirement of invariance under temporal partitioning. Again, the response is a probabilistic 'winner-takes-all' one.

The third layer of the proposed neural network is an object recognition one, which, in the simplest case, has a node for each object presented during training. Since the hypersurfaces of different objects' parts may intersect, shape units in the first shape representation layer may be representative of a number of different objects. The object recognition layer therefore essentially serves to learn the probabilities that different objects may be being viewed, given a firing pattern of nodes in the previous shape representation layer. This is achieved by connecting each shape representation node with each object node and, for each such connection, learning how frequently each object may win through labelled training data (this assumes that enough training data is learned to reflect these distributions). The estimated conditional probability that an object C is being viewed in the image data D ($P(C|D)$) can then be determined over the training data according to the ratio of the number of times the shape node j responded to an object $P(C|j)$ to the number of times that the node won $P(j|D)$. This probability value can then be assigned to the connection for subsequent output. Resultantly, presentation of a novel PGH vector to the network will result in output of a set of probabilities that each object learned in the network is being sampled, i.e.

$$P(C|D) \;=\; \sum_j P(C|j)P(j|D) \tag{3.4}$$

an approach known as probability recoding. The algorithms presented in [25] were designed to allow the robust estimation of $P(j|D)$ from the accumulated responses to a set of individual histograms. Given the outlined neural network framework, subsequent analysis [41] indicated that the accuracy of the representation will be relative to the number of nodes used in the representation layer and the number and typicality of the example PGHs used in training the network. It is suggested that the number of nodes used can be optimised (essentially by trial and error) to limit any recognition ambiguity across the objects considered [41]. The utility of the neural network system was however successfully demonstrated in this regard and performance accuracy was related to the number of nodes required by the network in order to accurately compute $P(C|D)$. It was found that the early layers of the network needed to converge to stable representations in order that later layers could be adapted to their task. In general, the full machinery of an adaptive system was therefore considered an unnecessary overhead for investigation of the remaining issues, provided that issues of occlusion and clutter were avoided during construction. Later work therefore dispensed with the network architecture and concentrated on the form of histograms required in order to unambiguously represent edge-based shape. Basic recognition experiments were run on a selection of wireframe models [41], which although indicative of PGH behaviour, were not suited to general 3D object recognition tasks.

3.9 Related Research

The only other prominent research associated with PGHs is that of Hancock and Huet [70]. This work aimed to optimise the use of PGHs for retrieval of 2D line patterns from large databases. Their proposed representation maintained the use of relative pairwise lengths and angles and the use of the Bhattacharyya metric for histogram comparison, but opted for a scale invariant histogram format using a single global histogram for each template. Observing that global histograms were however prone to saturation, thus making unambiguous recognition impossible in the general case, entries to the global histogram would instead be limited to the nearest (typically 5 or 6) local linear edge features. Because of this 'gating', it was proposed that scale invariance would be achieved. However, the associated research was based around matching fixed templates of black logos against plain white backgrounds and very similar aerial road photos, for which the generalised 3D object recognition issues of occlusion, background clutter and variable illumination were very limited or not encountered at all. For example, any scheme based upon selecting the nearest neighbours to an edge feature in an image will be hampered by broken edges, as so often occur in typical images of 3D objects, or the addition of background clutter, which would invalidate the solution for objects' extremal edge features against noisy backgrounds. Furthermore, by using a single histogram, the completeness properties of the PGH representation are invalidated. Whilst abstracting the recognition scheme proposed in the original work by Thacker et al. [38][37][41][42][51][50][46][44][61][64] in an attempt to offer a more compact and less specific representation, the work opens up susceptibility to ambiguity.

In the decade following the initial publications of the PGH method, numerous publications suggested techniques which recognise objects using local edge information encoded in histograms. What this later work lacks is an approach which integrates measurement uncertainty with representation, a theory of statistical matching, or notions of information as embodied here by the idea of completeness. Analysis suggests that the original research's proposed scheme remains a perfectly valid, complete and optimal one for representing and recognising the projected 3D shapes of objects in arbitrary scenes via PGHs.

3.10 Conclusions

The Pairwise Geometric Histogram (PGH) representation has been shown to possess all the characteristics required for application to a multi-view-based, real-world-oriented 3D object recognition system. PGHs are the result of an engineering effort orchestrated to that end and it is suggested that no alternative (essentially different) representation may exist under the constraints proposed. However, to date, no opportunities to actually implement a multi-view-based recognition system based upon PGHs have been forthcoming.

Although neural networks hold great promise in terms of generalised unsupervised learning, they are notoriously fiddly to train and implement and much more research is required before a finalised practical solution to the problem is realised. In current related research, although the same underlying processes are effected, a more direct and controlled form of 3D object learning is to be implemented.

Chapter 4

Quantitative Localisation and Verification of Projected Wireframe Edge Models

4.1 Introduction

The use of edge features for object detection and localisation tasks is prevalent throughout the history of computer vision. The edge-defined contours of projected objects can be seen as compact and powerful shape descriptors, conveniently offering a high degree of invariance to background clutter and environmental illumination. Furthermore, quantitative spatial information is concentrated at edges, with surface regions typically offering little information to help determine and localise 3D scene structure under arbitrary illumination. This chapter presents a methodology for representing, accurately-localising and verifying the presence of (rigid) edge-defined 3D objects in images. The work supports both the TINA stereo and view-based model matching systems, which are detailed in the following chapters.

Typically, a 3D object model matching system will return only approximate cues relating to the position and orientation of a recognised model in a scene. This may be due to approximations built in to the recognition architecture to limit processing and memory costs, inadequacies of any representative models or because of degraded image evidence due to illumination, clutter and occlusion. Because of this, a process of projected model alignment optimisation may be required to correctly align a hypothesised model with the corresponding image evidence. This allows for more precise 3D object localisation to be inferred, as may be required for interaction, and for more informed assessments of the validity of any model matches to be made. Accordingly, a quantitative statistical metric is presented herein to support the precise alignment and subsequent verification of any image-projected 3D wireframe object models. The novelty of the work centres around a bootstrap approach to the definition of edge detection and localisation coupled with probabilistic combination of corresponding edge orientation information. Emphasis is placed on quantitative testing of the assumed distributions in order to avoid any arbitrary distribution assumptions and weighting factors.

4.2 3D Object Wireframe Modelling

With initial regard to stereo model matching, 3D models are required as the basis of feature matching. Such processes are however better supported if only those features that are co-visible are modelled around the view-sphere; so called '2.5D vision', where self-occluded features are discounted from 3D matching. Historically, within the TINA stereo vision system, this has been achieved with reference to accompanying view-dependency files, which detail which features are likely to be visible from a finite number of sample view-points around the view-sphere. The closest stored view to a sample view-point is taken to indicate which features should be visible. Following 3D model matching, the same view-dependency files can be used to sample 2D projected views, so that any model match hypotheses can be projected against the image data in support of optimised alignment for localisation, verification and camera calibration purposes. 3D models have previously been composed of fixed linear and elliptical sections.

With further regard to development of the proposed view-based 3D object recognition system based upon Pairwise Geometric Histograms (PGHs), 3D information is not strictly required for recognition, as matching is based upon sampled 2D information. However, the view-based system is again geared towards accurate 3D model localisation, whereas the process of view-based model matching is approximation based, mediated by the minimum prescribed level of reconstruction accuracy tolerated across continuous interpolated manifolds. So while accord from a number of PGHs regarding a recognised object's hypothesised location, scale and orientation may be sufficient to indicate the reliability of object occurrence, further 3D projection alignment optimisation processes may be required in support of accurate 3D object localisation and pose inference. In difficult feature-based model match scenarios (i.e. in complex noisy scenes) it also makes sense to make direct use of the original image data, for which 2.5D edge data is again required for continuous view-sampling and projected model correspondence verification.

Because robotic handling and visual inspection machinery is unavailable for this project, there is a further requirement for 3D wireframes as surrogates for the corresponding tangible objects to support object manipulation and PGH encoding, i.e. learning. Although the system is to be tested on real-world images of tangible objects, such virtual object modelling conveniently allows for objects to be precisely manipulated in 3D for research analysis. Furthermore, for real scenes, each object may otherwise be required to be observed under a range of illumination conditions for each view in order that all potential edge features be observed and learned (i.e. many edge features may be 'washed out' in an image due to a lack of contrast).

As should now be clear, surface information is of little use for the 'bottom-up' model matching task because the bulk of quantitative image information relating to object shape and position is conveyed by edge features. Surface regions typically convey little or no discernible information, with it being fundamentally impossible to parameterise a featureless surface region from a single image of an unknown object under arbitrary illumination. Though surface information may be computed via 'top-down' strategies, in the context of a system that must learn and update any internal models from visual data, this precludes direct use of surface-based 3D models as a suitable basis of recognition. This however presents a potential problem in that edge feature visibility is directly related to surface occlusion. Each point along an object's projected edge features (i.e. structural discontinuities or extremal projected boundaries) can only be visible if a ray from that point to the camera does not intersect an object surface. In line with the view sampling scheme employed by the TINA stereo model matching system, aspect graphs [9] have therefore been proposed as a solution to the view-sphere feature visibility problem.

Since specific object features will be visible throughout certain ranges of view-point, aspect graphs serve to partition objects' view-spheres into regions for which sets of specified features are continuously visible. Aspect boundaries represent visibility transitions where features may come into or go out from view. Although such representations may become very complex for faithful mappings of intricately structured objects with lots of features, the methodology supports continuous assessments of feature visibility without the need for maintenance of surface information. The methodology is especially useful for simpler objects with well-defined boundaries of visibility. Because the aspect graph naturally segregates model features into consistently visible feature sets, aspect graphs are also useful as the basis of interpolative view-based recognition, avoiding discontinuities caused by changing feature samples.

Aspect graphs do not however lend themselves well to the accurate modelling of partial feature visibility through self occlusion. Such factors may be approximated by splitting features into segments, although at the cost of further complicating the assignment and composition of associated aspect boundaries. Fortunately, It can reasonably be argued that the requirement of perfectly faithful representations of feature visibility around the view-sphere is unnecessary, especially in the context of current research-oriented experimentation. A goal of faithfully modelling 90% of an object's projected features around the view-sphere is therefore proposed as being sufficient to support view-sampling, model matching, object recognition and subsequent verification. Beyond such levels, the modelling process becomes one of diminishing returns, especially considering that the proposed view-based recognition processes are designed to be robust to significant degradations in projected appearance (i.e. due to unfavourable illumination conditions and interference from scene clutter and occlusion).

The approach taken to modelling the visibility of 3D objects' features is therefore to sub-sample objects' view-spheres into approximately equally-spaced view regions, throughout which specified features are deemed visible. The nearest stored view to a sample viewpoint will therefore indicate which features should be visible. Although this differs from aspect graph theory in that view transition boundaries are implicitly encoded as medial axes between sampled view-points, this process is more attuned to the requirements of a system that must learn incrementally, potentially from a small number of sample view-points.

Although there is plenty of scope for adaptation and refinement of the proposed view-based sampling scheme with regard to the number and position of any learned reference view samples, more sophisticated aspects of representational refinement are deferred for future research. In current research, each object's view-sphere is uniformly sampled with 42 views, as sampled from the nodes of a triangulated icosahedral spherical mapping (see Figure 4.1). As discussed in the following section, further sub-view visibility constraints can be specified for elliptical features, along with other pertinent visibility information specific to each feature throughout the specified range of view.

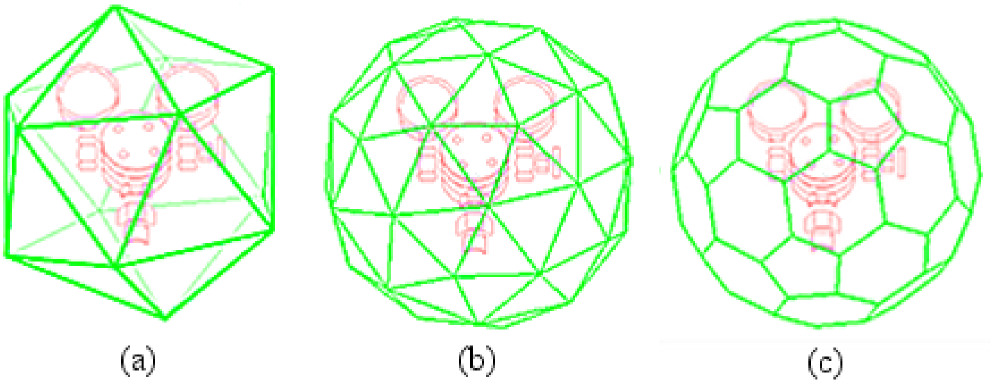


Figure 4.1: *The diagram above indicates how a 42 view spherical mapping may be sampled as a Voronoi tessellation (c) of a triangulated (b) icosahedron (a).*

4.3 View-Based 3D Feature Sampling

There are two main classes of features defining a 3D object's shape; fixed edge features, such as sharp planar surface discontinuities, and viewpoint dependent ones; representing the extremal projected boundaries of any smooth continuous surfaces. While the first class of edge features can simply be projected into the image from 3D, the second class presents more of a problem. To represent such features for arbitrary objects, a surface model is required from which to determine any such boundaries. In this work, mostly for convenience, we are concerned with the modelling and detection of man-made objects with well-defined simple geometrical structure. The current approach assumes that we can therefore account for a broad range of objects with line and ellipse segments and conical and cylindrical sections. The viewpoint dependent outer profiles of occluding boundaries can be accounted for by connecting the points on the supporting end ellipses that are most distant from the axis connecting the two ellipses' midpoints, as can be observed for the object in Figure 4.2 about 2 orthogonal axes. More complicated structures can be approximated, to a required degree of accuracy, using sets of object boundaries and resulting curves. The set of 15 manually constructed 3D wireframe test objects is presented in Figure 4.5.

Figure 4.2 indicates how 3D object feature visibility can be approximated by sampling feature visibility around an object's view-sphere with fixed spherical mappings, as derived from encompassing platonic solid constructs. Although simpler spherical mappings may be applicable for simpler objects (e.g. with sample views assigned to the 8 corners of a cube or 12 centre-faces of a raw dodecahedron), 42 views have been sampled for each object in the database used for this project (see Figure 4.5) to account for more faithful feature visibility through self occlusion (see Figure 4.3). If extra precision was required, any such view-sphere mapping could be iteratively sub-divided with more views being added in specific regions as required. Individual feature visibility has been set manually for each of the 42 views for each of the 15 manually modelled 3D objects in the test dataset. Despite an individual feature-based 'point and click' selection mechanism having been developed to support this process, many weeks worth of time were invested in the view-based wireframe model construction process. Although this chapter goes on to introduce a technique for semi-automatically assessing which 3D features are visible from a presented model viewpoint, this would ideally require robotic handling and visual inspection equipment which was unavailable for this project. The models were otherwise initially required for application to the pre-existing TINA stereo-based model matching computer vision system, which has historically required such manually composed view-based models.

The requirement for a reasonably high number of sample views is exacerbated when considering application to the proposed view-based recognition scheme, where features are sampled locally and may thus be more affected by inclusion of any invalid features in their local contexts. It is also desirable to not waste resources learning and recognising features that are falsely deemed to be visible. Because there is no mechanism to account for transitional feature occlusion across view regions, occluded features may be sub-sampled, e.g. with only one half of the feature being set as visible throughout the view. By maintaining more views, features can thus be more faithfully represented across the view-sphere, enhancing the practical acuity of the system.

Although single linear features are labelled as being visible or not within each view region, the nature of elliptical features lends them to more detailed visibility analysis.

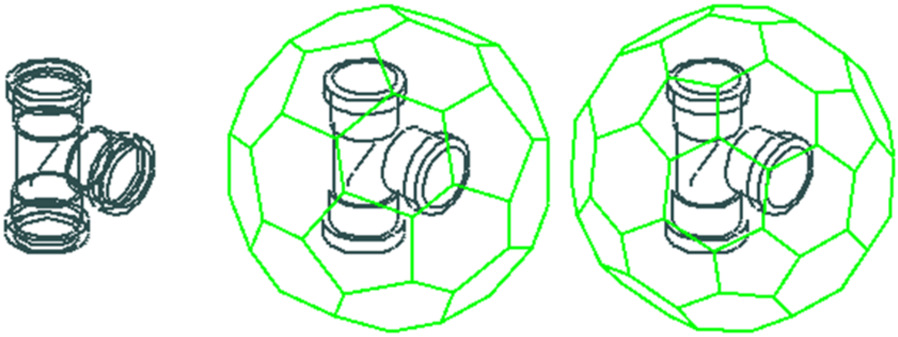


Figure 4.2: *A full 3D wireframe object is pictured on the left with occluding boundaries attached to specified projected extremal ellipse points. The 2 adjacent images show how the object's view-sphere can be partitioned into 32 or 42 approximately evenly spaced views (Voronoi cells derived from triangulated dodecahedral and icosahedral spherical mappings), each of which defines the features that are visible throughout that region of view-space. A 3D reference vector is assigned to each elliptical feature allowing for the front and back halves to be unambiguously defined (see Figure 4.4), further supporting specific-handed assignments of attached occluding boundaries in the image plane.*

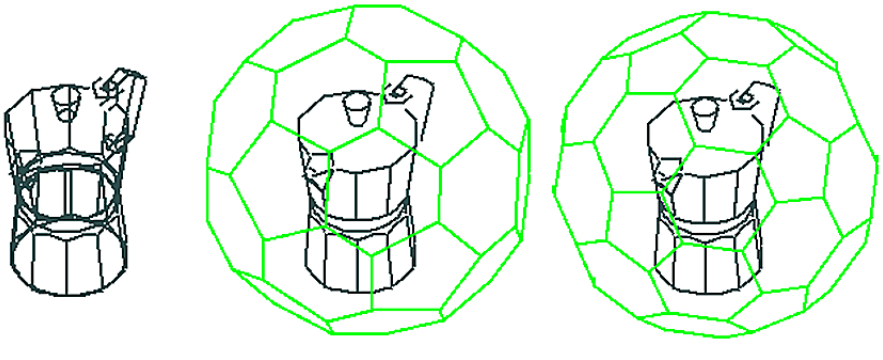


Figure 4.3: *The need for a high number of fixed views is exemplified with the coffee pot object pictured above. It is difficult to faithfully model the visibility of 3D objects' peripheral features around the view-sphere with fixed views even with 32 or 42 views as indicated.*

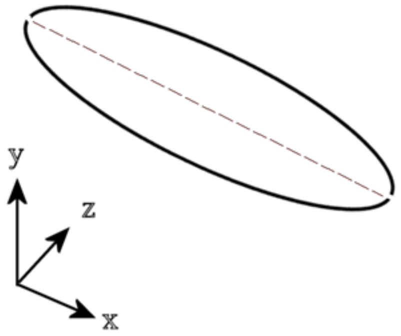


Figure 4.4: *3D circular features are visibly projected as ellipses in all but head-on views. By assigning each such model-based 3D feature a fixed 3D up direction (y), circular features' top and front halves' visibilities can be solved along the line of sight (z). Specific handed occluding boundaries can be attached to the outermost ellipse points (between pairs of ellipses).*

A circle, for instance, unless viewed head-on, will be visibly projected as an ellipse, meaning that it can be split along its major axis, allowing for visibility to be specified for both the front and back halves. As detailed in Figure 4.4, a 3D reference vector can be assigned for each 3D elliptical feature, pointing perpendicularly to the signed supporting plane, allowing for each half of a projected ellipse to be unambiguously assigned relative to the projected direction of this vector. Occluding boundaries can then be unambiguously assigned to particular handed sides of such supporting elliptical features. Similarly, upon viewing an object, a dot product test can be made between the 3D reference vector and the relative 3D model view-point so that visibility for each half of the ellipse can be further specified across the top and bottom of the feature as viewed. Coupled with attached, specific-handed occluding boundaries, the four modes of visibility defined by these two binary constraints allow for any 3D conical sections to be modelled continuously across the view-sphere. Any visibility constraints are stored in the view-dependency files relative to the specific feature.

Although beyond the scope of current research, the representational scheme could be further enhanced by treating line bounded planar regions in a similar manner to ellipses, with each such line segment's visibility being further specified relative to whether the planar region was being viewed from above or below, thus further obviating the need for exhaustive view-vector-based surface point intersection mechanisms. Such top/ bottom feature visibility constraints can otherwise be accounted for by assigning visibility boundaries in accordance with aspect graph theory, as described above.

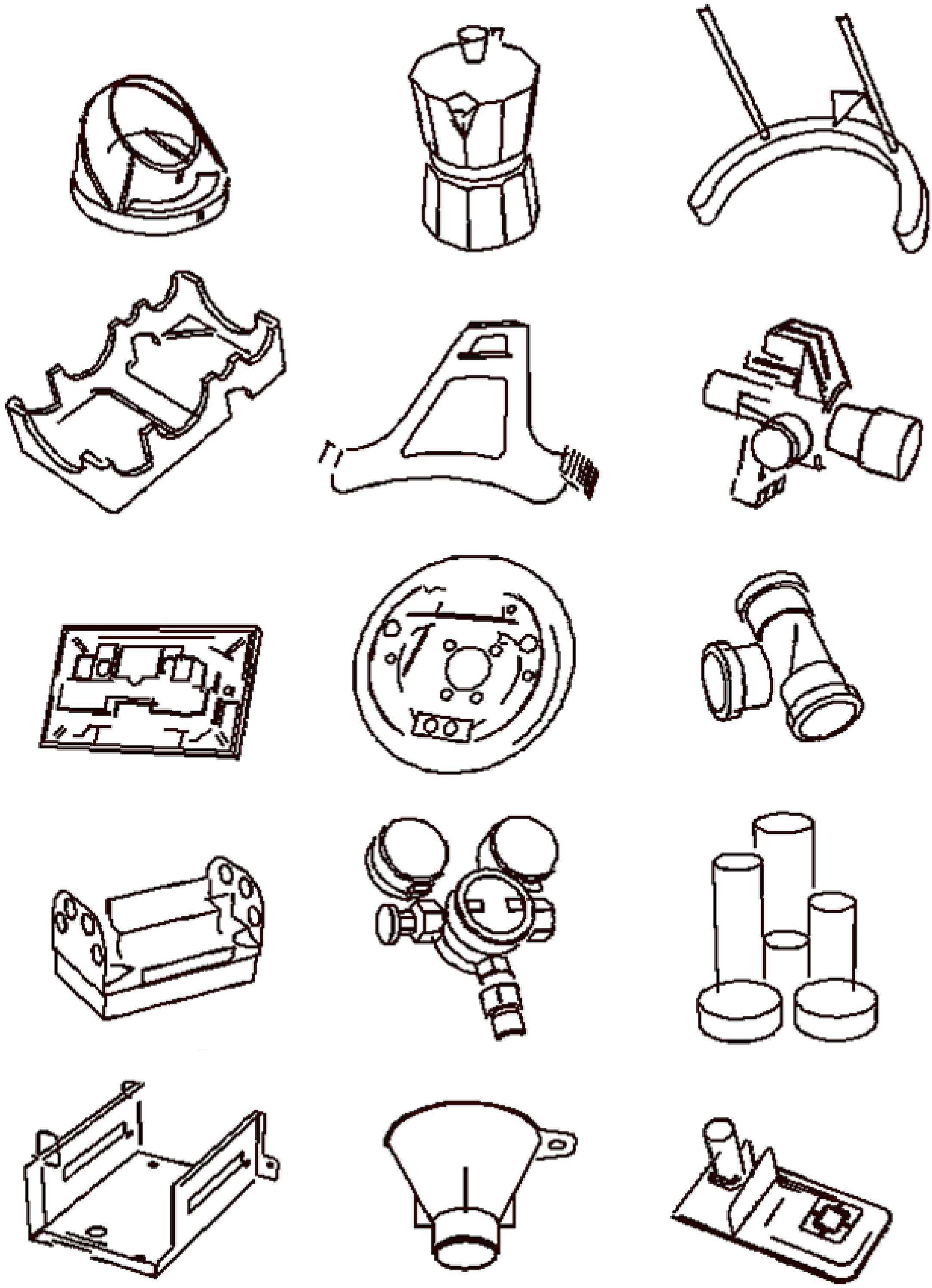


Figure 4.5: *The collection of 3D wireframe objects (manually) produced and utilised for research. Reading from top left to bottom right by referenced name; grill, pot, stand, guide, Aframe, pump, plug, brake, pipe, wiper, valve, (desk-) tidy, tray, funnel, widget.*

An object model must also account for image scale. The features describing an object's shape will be dependent on image scale, with finer features often being invisible at lower scales. Edge detection is also notoriously inaccurate at low resolution, where the predicted position of a step edge may be significantly distorted. Current research is focused around limited ranges of view-point (e.g. 1 to 2 metres for 0.25 mega-pixel images), for which objects' characteristic edge-based appearances are relatively stable. Well-defined physical edges such as sharp surface discontinuities and occluding boundaries are of most interest. Dual representations may otherwise be employed to account for feature visibility throughout scale space.

Another scale related consideration is that of thin plate representation. Unless infinitesimally thin, a planar region will have front and back boundaries of surface discontinuity. Representing both, especially relative to low scale imagery, would be unnecessary. Instead, a single feature can be modelled, which can be adaptively fitted, as will now be described, against the most prominent corresponding image edge data. This accords with a philosophy of developing simple, generalisable wireframe models.

4.4 Lateral Feature Shifting

The considerations outlined above provide a model for predicting feature visibility suitable for recognition, but for accurate localisation of objects, other effects must be modelled. Specifically, further consideration is required for representing the 'fixed' type geometrical features as described above. For many man-made objects, inter-surface regions will be slightly curved in nature. The positions of any corresponding edges will vary within these regions depending on the relative illumination. General behaviour can be predicted from simple illumination models. For distant illumination, proximal parts of an otherwise homogeneous surface at the same orientation are expected to have equivalent grey-level values. This ensures that any spatial effects apply systematically along an extended boundary, i.e. a lateral shift. Observation of these problems indicates that most of the errors of re-projection resulting from simple wireframe models correspond to such lateral edge feature displacements (see Figure 4.6).

In order to account for lateral shifting of wireframe edge features, the software has

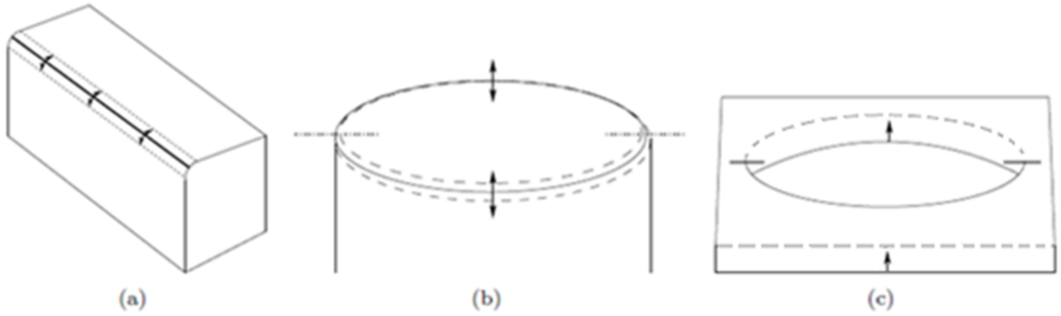


Figure 4.6: *Various sources of lateral feature shifting due to lighting and modelling simplifications; the location of a straight edge along a curved surface boundary (a), the end of a cylinder (b) and a hole in a thin plate (c). Circular features, visibly projected as ellipses, are split along their major axis, with each half being allowed to shift independently.*

been adapted to optimise the lateral locations of features in the image plane upon model to image projection, so that the most likely position of each feature can be determined (see Section 4.8). This modification is essential in order to quantitatively interpret any resulting likelihood scores. This strategy is also used to counter any imprecision of the manual wireframe construction process or the edge feature detection process. Without this mechanism, the location process may be dominated by systematic shifts in the extended features, with alignment driven to balance a variety of artefacts. This illumination dependency does not appear to have been considered previously in the computer vision literature. This is perhaps because any associated effects are unobservable when the object models are not accurate predictions of appearance, or when the objects chosen for demonstration are 2D or simple geometrical shapes with sharply defined features. The effects would also be hidden by conventional (loosely constrained) distance-based approximations to localisation and verification.

Unconstrained lateral shifting raises a possible problem for objects composed entirely of lines because it prevents unique definitions of object location (e.g. a centroid) and scale. This issue can be resolved either with use of 2nd order curves (constraining the alignment along the supporting elliptical feature's minor axis), or through the use of a localisation constraint during positional optimisation. Such a constraint can be obtained from accumulated statistics regarding the apparent position of features detected in a series of equivalent views and simply combined into the existing framework. In essence, we could learn which features are 'fixed' at their expected locations, i.e. those relating to accurately modelled and sharply defined physical structure, with this information being stored in the established view/ feature visibility files. Similarly, the degree to which any other features are expected or allowed to shift would be learned, with this information being stored in model-centric-metric units to account for projected scale variability. The utility of this approach for local feature positional optimisation is experimentally reviewed in the following sections in the context of the proposed supporting quantitative match quality metric.

4.5 Quantitative Edge Feature Localisation

The process of edge-based model matching is supported by a process of edge detection which serves to extract binary edge contours from an image. A version of the Canny edge detector [14] is utilised for this purpose in this work because it produces well localised and relatively unbroken edge segments. The process of model matching is detailed in Chapter 7 in terms of both view-based and stereo matching. The problem at hand is the optimised alignment of a hypothesised projected edge feature model against a purported occurrence of such a model in an image given an initial indication of the projection parameters by the model matching sub-system. An optimisation method such as 'simplex' [36] can be utilised to best align the image-projected 3D features, while simultaneously calibrating any associated camera parameters. The key issue is that of defining an appropriate, quantitatively valid cost-function.

Perhaps the most straightforward approach to accurately aligning a wireframe shape template with the corresponding image evidence is to minimise the distance of each projected sample point to a binary image edge point. This is effectively what is advocated by Chamfer and Hausdorff model matching techniques [21][39], where the mean and largest individual distances between edge-point sets are respectively minimised. Although these methods provide a tunable statistical system under restricted circumstances, they are not directly applicable to random images of objects, as the residuals from the predicted model values are not derived from a model of image formation. Furthermore, complications arise in generically determining which model feature points relate to which image feature points. Simply assigning the nearest image edge point to each sampled feature point and discounting these from further consideration may lead to unstable accounts of template correspondence; especially in noisy scenes.

The orientations of models' predicted edge features also convey a significant amount of information regarding model-match quality, as has been noted in a number of publications [55][80]. Without such constraints, distance transform techniques, for example, are prone to registering significant likelihood scores across their features in busy or highly textured images, where each predicted point would be close to an image edge. By ensuring that feature matches are only counted if at admissible orientations, any incorrect, coincidental feature match hypotheses can be more readily rejected. This is especially important when considering the typical flawed nature of objects' image-edge maps, where many of an object's features may not be detectable. Hausdorff methods have been adapted to account for edge orientation information, where a cost for disparate orientations is weighted into the distance transform function [55]. Similarly, truncation techniques have been developed where only a reduced set of the best point matches are considered [63]. It should however be noted that assessments of edge orientation are commonly dealt with without any account of propagated uncertainty [31][47]. For weaker edges, such errors may be significant, thus requiring appropriate consideration.

While the techniques described above do intuitively have some operational merit, i.e. accounting for whether each predicted edge point is supported to some degree by the image evidence, the inherent vagueness of each method can be seen to be a drawback in terms of precise and reliable quantitative localisation and prospective verification. In summary, the similarity distributions observed for such geometric approaches will vary as an unknown function of the illumination, the object and its orientation, making such schemes unreliable in the general case. The work presented in this chapter is concerned with analysing how well each predicted edge feature point is supported by the underlying image edge evidence. By basing the statistical distributions directly upon the measured evidence, quantitative tests can be constructed that follow directly from an understanding of image formation and feature detection. The methods are developed to quantitatively embody any degrees of uncertainty, such as the reliability of detection and errors encountered in the analysis of edge orientation. This is proposed as being the only theoretically justifiable way to perform such a task, further supported by the associated need to quantitatively verify the quality of any sampled model match hypotheses, as discussed in the following sections.

There are several important benefits expected for the proposed approach to quantitative model to image alignment. By taking explicit account of the original image data, the approach is expected to be more discriminatory than distance-based measures while being less dependent upon the details of image formation than appearance-based approaches. Secondly, the method is quantitatively testable. It is possible to verify that the assumed probabilistic metrics are indeed truly reflective of the behaviour of sampled edge data for real-world examples of imaged objects. Thirdly, unlike previous approaches, the theory gives an absolute calibration (i.e. relative weighting) for the orientation and localisation terms used in likelihood construction. The difficulties in attempting this approach are however significant. A successful theory must deal with the effects of arbitrary illumination and changes of physical appearance under projection while defining calculations that are computable using minimal knowledge of expected image content, specifically; predicted edge location, orientation and local image data.

In the first instance, a general definition of the edge feature localisation term can be defined in terms of probability. Assuming that the spatial location and orientation of an edge are uncorrelated (which is true for sufficiently large separations between projected points), and given a wireframe model and a set of camera parameters θ, we can define the probability of an edge pixel being present at a certain location (x, y), with an orientation ψ and within intervals $\Delta_x \Delta_y \Delta_\psi$ (since we are initially dealing with probability densities) as

$$\begin{gathered} P(x, y, \psi|\theta) = \\ \int_{x-\Delta x}^{x+\Delta x} \int_{y-\Delta y}^{y+\Delta y} \int_{\psi-\Delta\psi}^{\psi+\Delta\psi} p(x, y|\theta) * p(\psi|x, y, \theta) dx\ dy\ d\psi \\ \approx p(x, y|\theta) * p(\psi|x, y, \theta) 2\Delta_x 2\Delta_y 2\Delta_\psi \end{gathered} \tag{4.1}$$

Taking the logarithm

$$\begin{gathered} lnP(x, y, \psi|\theta) = ln\ p(x, y|\theta) + ln\ p(\psi|x, y, \theta) \\ -ln(8\Delta_x \Delta_y \Delta_\psi) \end{gathered} \tag{4.2}$$

In order to unambiguously define this probability, a methodology for selecting appropriate intervals is required. The intervals are therefore chosen to be proportional to the measurement accuracy ($var(x)$, etc.). So

$$lnP(x, y, \psi|\theta) = ln\ p(x, y|\theta) + ln\ p(\psi|x, y, \theta) \tag{4.3}$$

$$+\frac{1}{2}\ ln(var(x)var(y)var(\psi))\ +\ constant$$

Given this definition, the remaining issues involve computation of these probabilities.

4.6 Edge Location Analysis

In order to optimally align a projected wireframe template with an image, a simple binary edge template is suggested to be suboptimal for general application across interference corrupted arbitrary scenes (as typically realised in real-world oriented vision tasks). This is because the binary cut off employed by the edge detection mechanism will disregard any weaker supporting edge evidence whilst paying no account to the relative quality of any detected edge regions. As discussed, associated distance-based metrics are also deemed inappropriate for this purpose. Edge strength may be otherwise accounted for in terms of local gradient magnitude, but the problem here is that it is impossible to predict edge strength under arbitrary illumination, meaning, for instance, that other distracting scene elements with coincidentally high gradient edge strengths may distract any alignment optimisation schemes. What is required is a uniform, repeatable sampling of edge features across an image that directly accounts for the likelihood of each pixel being an edge feature with robustness to arbitrary illumination and background conditions. Such reasoning is further supported by the need, in this case, for the quantitative statistical combination of corresponding orientation information, as discussed in the following section, especially with regard to compatibility with any prospective verification procedures.

The edges of most interest for object detection are continuous step edges that define the contours of objects. By definition, a pixel can therefore be deemed to contain such an edge feature if its gradient value (edge strength) is both above the observed noise level and greater than 6 of its 8 immediate neighbours. This strategy should also conveniently allow for detection of corner and terminal edge features, with the only exception being T type junctions, where, a probability relating to 5 rather than 6 from 8 immediate neighbours would be required. Even if the scheme was invalidated by T junctions or similar interference from surrounding noise, the scheme should still return positive probabilities for such edge point features, at the cost of a slight imbalance in predicted probability.

On the assumption of approximate uniform Gaussian errors on the edge strength values, the probability that the pixel under consideration would be larger than one of its neighbours would be computed using the error function (erf()). The probability that the central pixel (g) is greater than one of the randomly selected neighbours (n) can therefore be calculated by (Appendix A.2)

$$P(g > n \in N) \;=\; \frac{0.5\; + \sum_{i=1}^{N} erf(\frac{g-n_i}{\sqrt{2}\sigma_I})}{N+1} \tag{4.4}$$

In principle, such a calculation is required at every image pixel, thus implying a considerable computational overhead. Approximating the error function with a linear ramp, set in accordance with the sampled image noise, allows the same calculation to be more quickly approximated via a unit-normalised 'soft' ranking process, i.e. as opposed to a 'hard' rank where each element is simply ranked in order without account of noise.

Assuming that a soft-ranked image gradient map is indicative of the probability that each pixel is greater than one of its randomly selected immediate neighbours (P), probability theory can be called upon to evaluate the chance that each pixel is greater than 6 from 8 of its immediate neighbours, as required. The probability that a value is greater than exactly 6 from 8 randomly selected neighbours is

$$P_{(6/8)} \;=\; 28P^6(1-P)^2 \tag{4.5}$$

The probability that the value is greater than 6 or more neighbours (P_{edge}) is then

$$P_{edge} \;=\; P_{(6/8)} \;+\; P_{(7/8)} \;+\; P_{(8/8)} \;=\; 28P^6 \;-\; 48P^7 \;+\; 21P^8 \tag{4.6}$$

with the corresponding hypothesis test (H_{edge}) (see Section 4.10) given by the normalised integral, obtained by application of the 'ordering principle'

$$H_{edge} \;=\; \frac{\int_0^P P_{edge}\,dp}{\int_0^1 P_{edge}\,dp} \;=\; 12P^7 \;-\; 18P^8 \;+\; 7P^9 \tag{4.7}$$

Each normalised entry (P) in the soft-ranked gradient image can therefore be transformed to the probability P_{edge}, thus indicating the probability that each pixel is greater in gradient value than 6 or more of its 8 immediate neighbours. Soft ranking schemes will however typically return approximately mean valued ranks (i.e. N/2) for featureless regions. Each pixel value can therefore be multiplied by the probability that the corresponding image gradient is above the noise threshold, as estimated here with a Rician integral $(T())$*.

A potential image $V(x, y)$ is accordingly formed for each image, directly representing an approximation of the (log) probability that each pixel contains an edge feature (see Figure 4.7).

$$V(x, y) \; = \; ln(T(g, \delta)P_{edge}(x, y)) \tag{4.8}$$

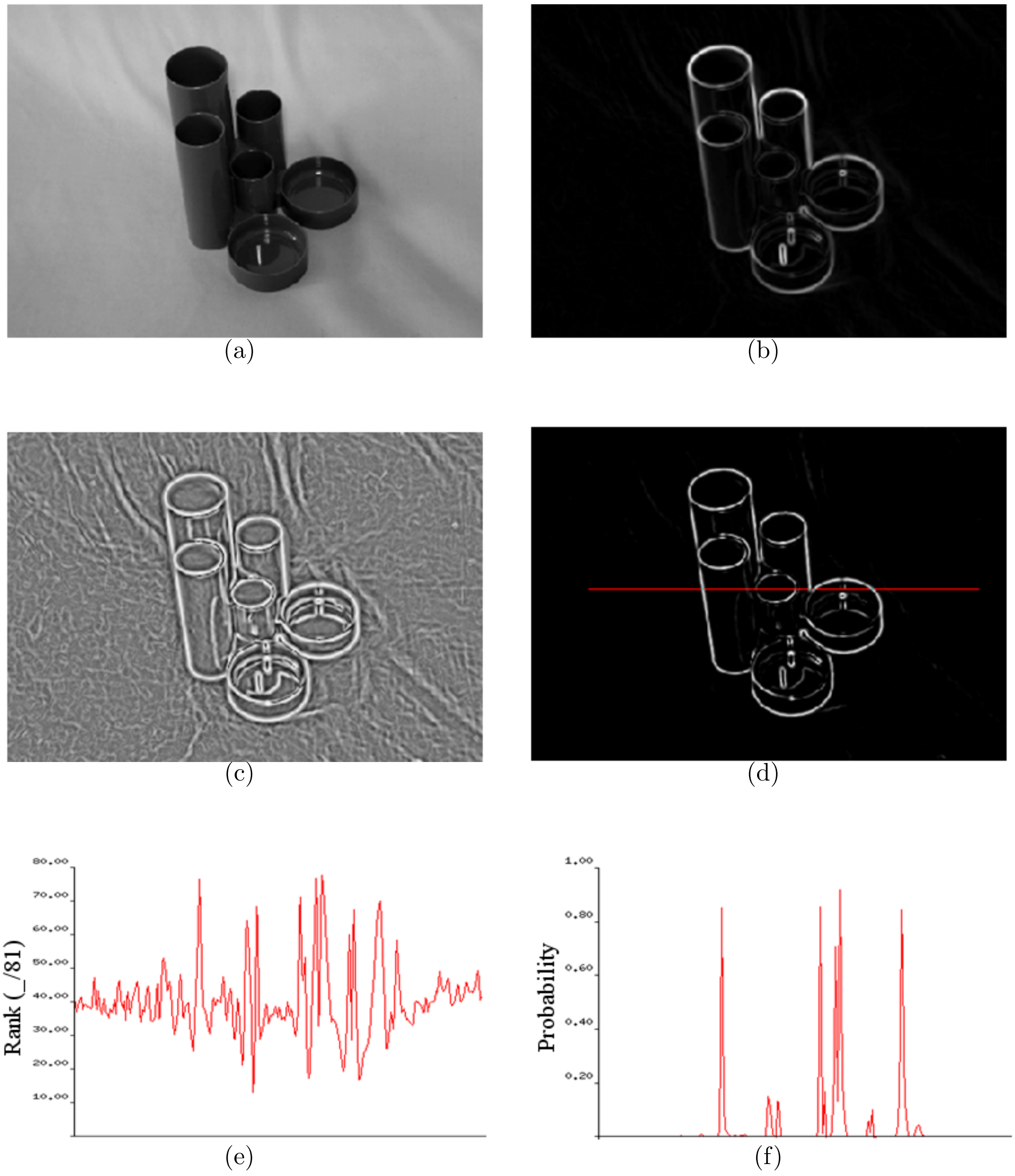
where δ is an estimate of the (gradient) image noise.

Therefore, for genuine edges, where $g > 2\delta$

$$V(x, y) \; = \; ln(P_{edge}(x, y)) \tag{4.9}$$

Using just 9 pixels with a ranking procedure is not however conducive to accurate calculation of likelihoods, since the rank is only based on 9 samples of evidence across a smoothed gradient manifold. More accurate probabilities can be sampled by using wider pixel reference regions, although it is still desirable to limit the scope of such sample regions so that adjacent edge structures do not interfere. Using 4 pixels each side and vertical half of a central reference region (i.e. a 9 by 9 pixel block) has proved to be a stable basis from which to infer the likelihood of the sampled pixel being greater than a randomly selected neighbour. The validity of the proposed scheme is evaluated in section 4.8.

*Specifically, a cubic approximation of a Rician integral with a limited lower bound is implemented in the software [1] for computational efficiency.



(a) (b) (c) (d) (e) (f)

Figure 4.7: *Image (a) represents a raw image of the (desk-) 'tidy' object accompanied by a smoothed gradient image (b). Image (c) represents the output of a rank filter, where each pixel in the gradient image is 'soft-ranked' relative to its neighbouring pixels. Image (d) is a 'potential image' (V), formed by multiplying the normalised rank image with the probability that each pixel is greater than at least 6 of its 8 immediate neighbours and the ranked gradient image noise floor (see Equation 4.8). The bottommost charts ((e) and (f)) represent a cross section of values conveyed by the rank (c) and potential (d) images, as sampled across the red horisontal bar indicated in the potential image. A 9 by 9 pixel block was used as the basis of ranking.*

4.7 Edge Orientation Analysis

The true distribution for estimation of the local gradient direction from separate x and y image derivatives has been described as a Von Mises distribution [53]. This is expected to be more detailed a model of the distribution than is needed for most practical purposes (i.e. away from low contrast regions). A first order (Gaussian) approximation can be derived using error propagation and tested empirically on real data. The local image orientation can be defined as

$$\psi = \arctan(\frac{dI/dx}{dI/dy}) \;=\; \arctan(\frac{u}{v}) \tag{4.10}$$

Given equal errors δ on the derivatives u and v

$$\delta_{\psi}^{2} = (\frac{v\delta}{v^2+u^2})^2 + (\frac{u\delta}{v^2+u^2})^2 \tag{4.11}$$

Since the variance is the square of the error, and taking 'r' to be the edge magnitude

$$var(\psi) \approx \frac{\delta^2}{r^2} \tag{4.12}$$

I.e. the error on the local edge orientation is inversely proportional to the edge strength r and proportional to the image noise δ. The edge strength can thus be considered as the Fisher information for calculation of image plane orientation [98] (i.e. strong edges are the best locations to apply an orientation test).

The orientation term can be selected in order to match the error distribution on orientation measurement ($var(\psi)$)

$$p(\psi|x,y,\theta) \;=\; \frac{1}{\sqrt{2\pi\ var(\psi)}} exp(-\frac{(\phi(x,y)-\psi(x,y))^2}{2\ var(\psi)}) \tag{4.13}$$

with $\phi(x,y)$ representing the edge orientation of a feature pixel. This is the likelihood for feature orientation. Equivalently,

$$ln\ p(\psi|x, y, \theta) = -\frac{(\phi(x, y) - \psi(x, y, \theta))^2}{2\ var(\psi)} \tag{4.14}$$
$$- \frac{1}{2}\ ln(2\pi\ var(\psi))$$

When this is substituted into the log probability for localisation, the second term effectively cancels with the interval for the angle measurement $\Delta\psi$ to give a chi-square like statistic[†]. The two remaining variance terms can be combined with other constant factors that play no further role in the estimation of parameters or parameter variances.

Unfortunately, the above measure contains a subtle but significant problem. Valid use of likelihood requires that the interval relating probability density to probability is not changed during the process of parameter estimation. One way to understand this is to observe that if the interval is defined using an estimate of variance at each location in the image, then perfectly good statistical matches will arise for data in regions of high variance. Simply optimising this function is likely to locate a curve over a featureless, noisy region. This is not because these measures are fundamentally wrong, rather that they are not suitable for relative comparison. Specifically, in this case, the true value of $var(\psi)$ for the optimal location of the curve is required prior to localisation. The problem of lack of knowledge of $var(\psi)$ at the solution can be avoided by observing that, for a wide range of edge strengths, the expected orientation error is approximately constant and scales with the intrinsic image noise, so that $var(\psi) \approx \kappa^2\delta^2$ where δ is the estimated pixel value error.

A quantitative hypothesis test (H_{angle}) for matching edge orientation (see Section 4.10) can accordingly be calculated via the (complementary) error function, i.e.

$$H_{angle} \ = \ 1 - erf(\frac{\phi(x, y) - \psi(x, y|\theta)}{\sqrt{2}\delta}) \tag{4.15}$$

[†] If the interval is not scaled in this manner, or if it is simply ignored under the conventional definition of likelihood as provided in standard texts [66], then one needs to explain where the probability density normalisation terms disappear to in the construction of standard statistical measures (i.e. squared difference divided by variance).

4.8 Quantitative Edge Feature Localisation Revisited

With probabilistic terms for edge location and orientation now defined, the combined likelihood L can be written as

$$ln\ L(x,y,\psi|\theta) = V(x,y) - \frac{(\phi(x,y) - \psi(x,y|\theta))^2}{2\kappa^2\delta^2} \tag{4.16}$$

Slight modification of this approach is necessary to deal with occlusion and specularity, so that valid feature alignments are not overly penalised. Residual truncation is therefore utilised by setting a maximum value χ^2_{max} for the contribution to the cost function from each data point.

As the likelihood is related directly to the quantitative probability of observing the data by a constant, it is understood that this satisfies the requirements for the optimisation to result in a 'consistent' estimate of the required parameters. The resultant cost for each of a feature's pixels can be summed to give a combined score for use in alignment, as explained below. Rather than exhaustively sampling every pixel separated point along each projected feature, sample points may be separated (e.g. every 4th pixel (relative to projected scale) in order to speed up processing.

In summary, the likelihood function defined in Equation 4.13 is used as the cost function supporting the proposed optimised projected wireframe model alignment routines. In essence, a view-based wireframe model is projected over the image data according to the particular model match and camera projection parameters and a match score is calculated as a sum of the likelihood values sampled at points uniformly spaced across each of the projected model features. The simplex algorithm [36] is then used to optimise any specified projection parameters (currently the location[‡] and orientation parameters of the matched model and the focal length of the camera) in order to maximise the match score realisable between the projected model and the underlying image data. To account for a range of modelling and illumination dependencies, rather than using a fixed global edge template, the position of each feature is independently optimally aligned with the underlying image data using a

[‡] The optimised distance of the object feature from the camera is constrained by 10% from the initial prediction in each direction. This, for instance, prevents the projected model converging to a point.

process of lateral feature shifting (see Section 4.4).

More specifically, with reference to Figure 4.8, the match score for a projected model ($ObjectView$) with n features is calculated as a sum of the match scores ($ln\ L_{max}$) for each projected feature (f_i) as optimised with a lateral shift in the image plane (for a fixed set of visible features and sample points). I.e.

$$ln\ L(ObjectView) = \sum_{i=1}^{n} ln\ L_{max}(f_i) \tag{4.17}$$

$$if(\ ln\ L_{max}(f_i)\ >\ \chi^2_{max}\ )\ ln\ L_{max}(f_i)\ =\ \chi^2_{max}$$

$$ln\ L(f_i) = \sum_{j=1}^{\#f_i} ln\ L(x_{f_j}, y_{f_j}, \psi_{f_j}|\theta) \tag{4.18}$$

where $ln\ L_{max}$ is the optimal value of $ln\ L$ as (quadratically) interpolated between the detected peak location (e.g. in pixel increments) and its immediate neighbours (see Figure 4.8) and # represents the number of sample points for the specified feature (f_i).

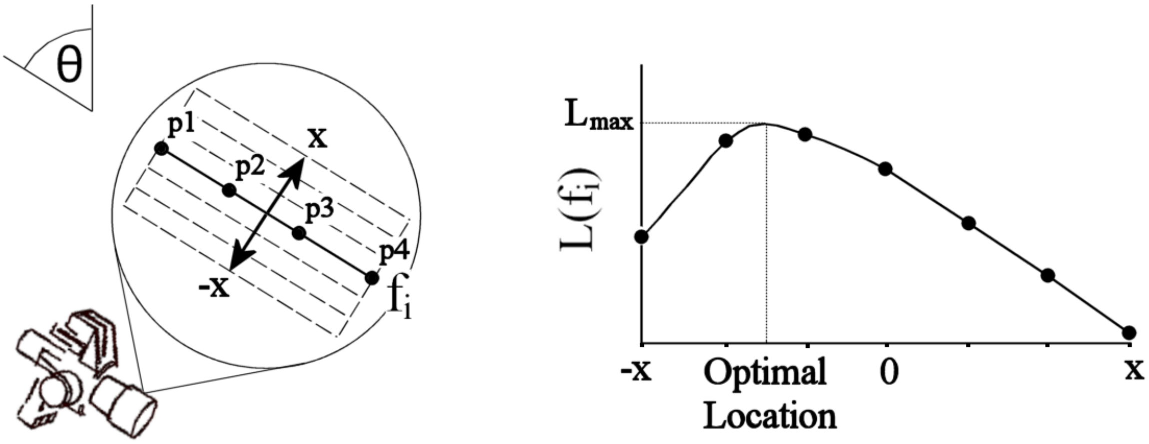


Figure 4.8: *The diagram above indicates how the process of lateral feature shifting is effected for a projected wireframe edge feature (f_i). The feature is sampled with a set of reference points (e.g. 4 pixels apart) ($p1-p4$) and the cumulative likelihood score ($ln\ L$)(see Equation 4.15) is assessed as the feature is laterally shifted in pixel increments (e.g. +/- 3 pixels) in the image plane. The optimal location of the feature and the corresponding likelihood score ($ln\ L_{max}$) are quadratically interpolated between the detected peak location ($x = -1$) and its immediate neighbours ($x = 0$ & $x = -2$).*

4.9 Validation of the Proposed Quantitative Metric for Oriented Edge Feature Analysis

Although having derived probabilistic terms in support of quantitative oriented edge feature localisation, the proposed metrics are premised upon a number of assumptions regarding image formation and edge detection. While empirically observed localisation results may qualitatively indicate the utility of the measures, the intrinsic validity of the proposed metrics may be verified by sampling the distributions of likelihood values observed across real-data. The conformity of the observed distributions can then be assessed relative to the theoretically predicted distributions. Not only does this process support verification of the validity of the proposed metrics, but it also allows for the corresponding location and orientation terms to be weighted appropriately without recourse to arbitrary scaling factors. Furthermore, given verification of the underlying metrics, statistical hypothesis tests may be implemented to sample probabilities of oriented edge feature occurrence by virtue of self-consistency with the underlying distributions.

The key observation regarding the proposed approach is that we have well defined predictions for the expected behaviour of the orientation and location distributions. The approximate behaviour of the orientation term is that of a Gaussian distribution with width $\kappa\delta$. As the process of edge location is a form of integral transform (i.e. analogous to histogram equalisation), the approach should deliver a uniform distribution of probabilities for true edges. Samples of log probability for localisation (represented as a potential image $V(x, y)$) should therefore have an exponential distribution.

4.9.1 Method

In order to evaluate the proposed oriented edge feature point similarity metrics, a random image of each of 15 objects in the test database has been sampled. The objects themselves have been selected so as to represent a diverse range of materials and fabrication processes. Using the feature visibility methods described earlier in the chapter (Sections 4.2-3), a corresponding view-sampled wireframe edge feature model has been aligned with each object (see Section 4.8). Each edge feature is downsampled with one from every 4 projected pixels (the images are 576 by 432 pixels).

Each 3D wireframe's projected location and pose has then been optimised in the image plane with the simplex algorithm [36] via edge strength correspondence (see Equation 4.17 (edge location only)) in accordance with the proposed potential image $V(x, y)$ constructs. As discussed in previous sections, each feature is allowed to shift laterally (see Figure 4.8), up to a maximum of 3 pixels, to compensate for modelling simplifications and possible illumination dependencies. For each object view, both the edge location and orientation likelihood values are sampled for each sample point along each feature and the values are amassed in corresponding histograms from which normalised cumulative distributions are sampled.

The cumulative distributions for sampled edge feature points for both edge location and orientation terms passing hypothesis tests at 0.1% confidence limits are plotted against the theoretically predicted cumulative distribution curves (i.e. cumulative-exponential and error ($erf()$) functions respectively) (Figure 4.9). The level of conformity of each such curve to the predicted distribution can be measured via the maximum vertical separation of the 2 curves in accordance with the Kolmogorov Smirnov test. In each case, the location and orientation terms are scaled to achieve best alignment with the underlying distributions (see Table (Figure) 4.11). In order to assess the utility of the proposed process of lateral feature shifting in such regards, the experiments have also been repeated without the lateral feature shifting mechanism, as detailed in Table (Figure) 4.12. In these cases, each model has been optimally realigned with the image data with fixed features.

4.9.2 Results and Discussion

With reference to Figure 4.9, the sampled distributions can be observed to be well-aligned with the theoretically predicted ones, thus supporting the validity of the proposed metrics. However, there is a reasonably wide degree of variance between specific object scaling factors, as required to obtain such a high degree of correspondence. Since we are dealing with a process of localisation and verification of a priori recognised objects, this is not necessarily a problem, in that the most accurate scaling factors could be learned for each object and called upon as required. Object-specific scaling factors are assumed to arise due to the different nature of the materials and physical edge feature types interfering with the ranking procedure. It would however be more convenient if universal scaling factors could be employed. To this end, the experiments have been repeated with the mean scaling factors sampled across the data set being factored into the calculations (Figure 4.10 and Table (Figure) 4.11). Although, for some objects, this evidently disrupts the degree of conformity observed, the net results are still monotonically related to the predicted distributions, with the net cost being slight distortion of the predicted probabilities.

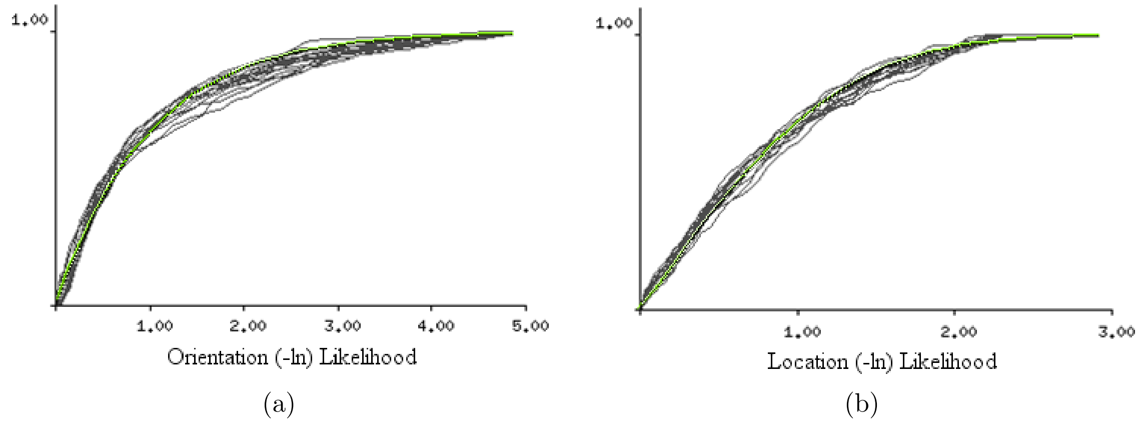
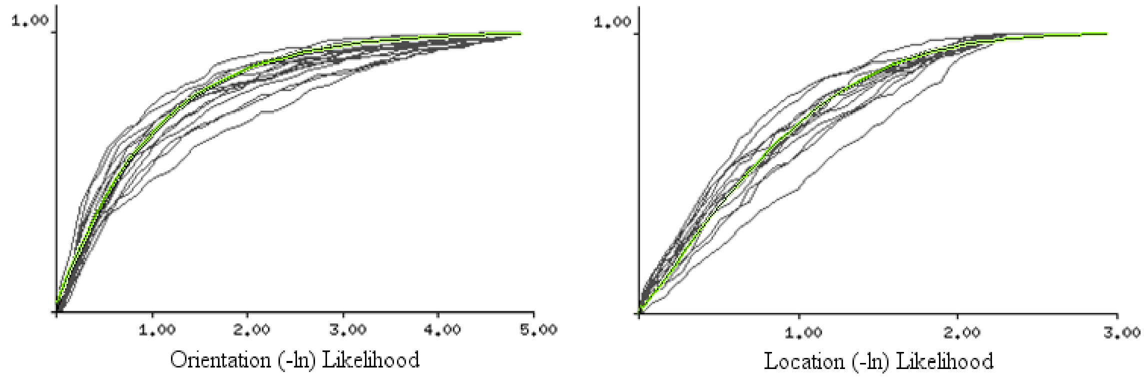


Figure 4.9: *With reference to the methods outlined in Sub-Section 4.9.1, the 2 charts above represent respective (normalised) cumulative distributions of the edge location and orientation likelihood values of the edge feature points passing a hypothesis test (at a 0.1% confidence limit) sampled across a well aligned instance of each of the 15 objects in the experimental database (Figure 4.5). The distributions are sampled from histograms representing the relative frequencies of the range of observed likelihood values. The edge location distribution (a) corresponds to a cumulative exponential function and the orientation distribution (b) to the error function. These theoretically predicted distribution curves are plotted in green as can just be observed underlying each sampled distribution set. The maximum vertical separation from the ideal curve is specified for each object in Table (Figure) 4.11, accompanying the scale factors used in each case to attain best alignment.*

Notably, the results presented in Table (Figure) 4.12 suggest that the process of lateral feature shifting does make a significant difference to the supporting metrics. Although the sampled orientation terms appear relatively unimpaired, the error (D-value) on the edge location likelihood terms can be seen to almost double in value on average. Furthermore, there is an average of 7% less feature points verified as being edge features if lateral feature shifting is not accounted for, with, as expected, some objects being more affected than others. Although this latter figure may appear marginal, as can be seen, typically only around 70% of an objects projected edge feature points will be detectable (due to illumination interference) and erroneous template matches are still expected to contribute, for instance, 30 or 40% correspondence in noisy scenes. Being able to account for an extra 10% of features for susceptible objects should



(a)

(b)

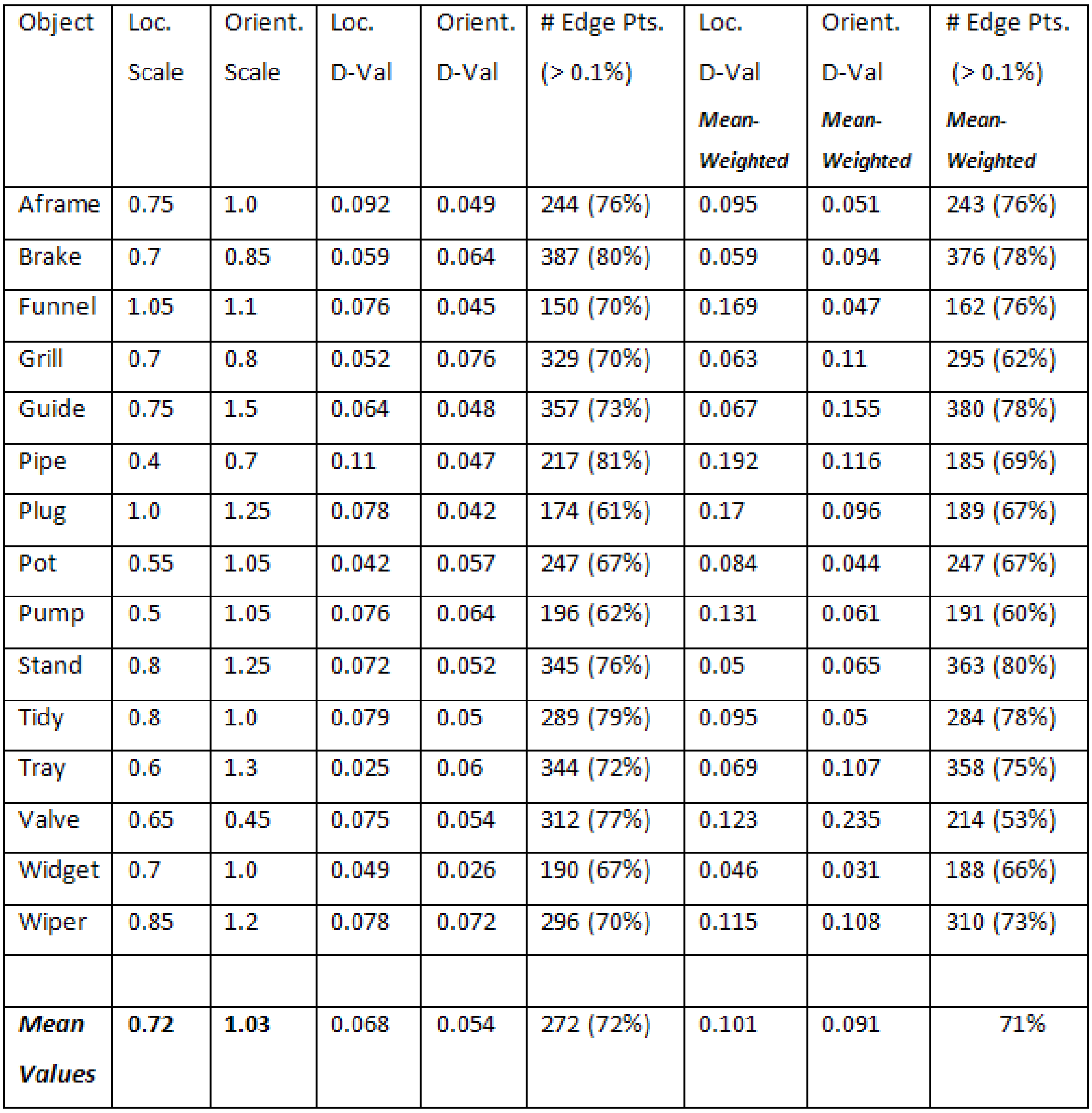
Figure 4.10: *The 2 presented charts are equivalent to those in Figure 4.9 with the edge-point location and orientation likelihood terms scaled by the mean scaling factors ((0.72 and 1.03) see Table (Figure) 4.11) sampled from the associated experiments. Quantitative degrees of conformity are indicated in Table (Figure) 4.12.*

therefore be beneficial to the differential hypothesis verification process.

4.10 Hypothesis Testing of Edge Location and Orientation

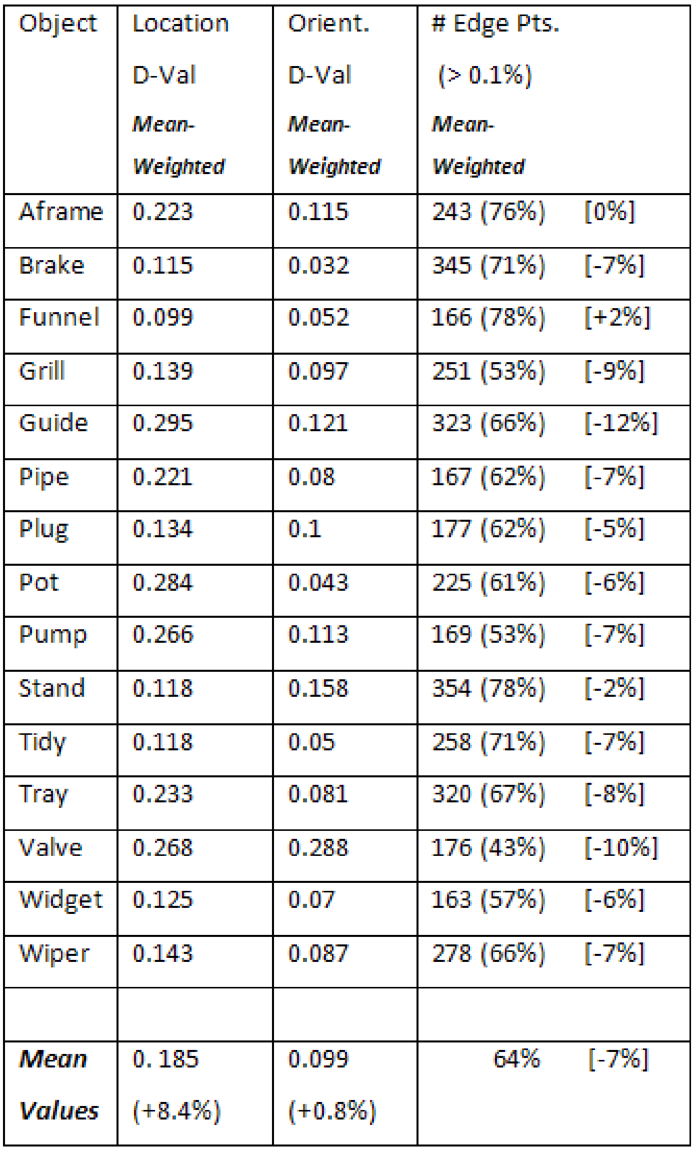
Hypothesis testing is a fully quantitative method to indicate how well sample data conforms to a predefined model. In this case, the statistical models relate to edge location and orientation, which are expected to correspond to (log-) exponential and Gaussian distributions respectively. In essence, the theoretically predicted cumulative distribution curves used in Figure 4.9, are used as probability look up charts for sampled edge likelihood values.

Probabilities resulting from hypothesis tests are expected to be uniformly distributed. In order to validly combine the terms for orientation and location, the following renormalisation formula can be derived [85]. Given n quantities, each having a uniform probability distribution $p_{i=1,n}$, the product $p = \prod_{i=1}^{n} p_i$ can be renormalised to have a uniform probability distribution $F_n(p)$ using



Object	Loc. Scale	Orient. Scale	Loc. D-Val	Orient. D-Val	# Edge Pts. (> 0.1%)	Loc. D-Val *Mean-Weighted*	Orient. D-Val *Mean-Weighted*	# Edge Pts. (> 0.1%) *Mean-Weighted*
Aframe	0.75	1.0	0.092	0.049	244 (76%)	0.095	0.051	243 (76%)
Brake	0.7	0.85	0.059	0.064	387 (80%)	0.059	0.094	376 (78%)
Funnel	1.05	1.1	0.076	0.045	150 (70%)	0.169	0.047	162 (76%)
Grill	0.7	0.8	0.052	0.076	329 (70%)	0.063	0.11	295 (62%)
Guide	0.75	1.5	0.064	0.048	357 (73%)	0.067	0.155	380 (78%)
Pipe	0.4	0.7	0.11	0.047	217 (81%)	0.192	0.116	185 (69%)
Plug	1.0	1.25	0.078	0.042	174 (61%)	0.17	0.096	189 (67%)
Pot	0.55	1.05	0.042	0.057	247 (67%)	0.084	0.044	247 (67%)
Pump	0.5	1.05	0.076	0.064	196 (62%)	0.131	0.061	191 (60%)
Stand	0.8	1.25	0.072	0.052	345 (76%)	0.05	0.065	363 (80%)
Tidy	0.8	1.0	0.079	0.05	289 (79%)	0.095	0.05	284 (78%)
Tray	0.6	1.3	0.025	0.06	344 (72%)	0.069	0.107	358 (75%)
Valve	0.65	0.45	0.075	0.054	312 (77%)	0.123	0.235	214 (53%)
Widget	0.7	1.0	0.049	0.026	190 (67%)	0.046	0.031	188 (66%)
Wiper	0.85	1.2	0.078	0.072	296 (70%)	0.115	0.108	310 (73%)
Mean Values	**0.72**	**1.03**	0.068	0.054	272 (72%)	0.101	0.091	71%

Figure 4.11: *For each of the 3D objects listed in the table above, a view-based sample of a 3D wireframe has been aligned with the corresponding image evidence, allowing for the distributions of edge location and matching-orientation terms used in the experiments to be sampled across the projected-upon image pixels. The scale terms represent those used to minimise the vertical separation of each sampled data distribution from the predicted curve as measured by the 'D-Vals'. The rightmost 3 columns represent the same data when the edge location and matching orientation likelihood terms are weighted by the mean sampled scaling factors (0.72 and 1.03).*



Object	Location D-Val *Mean-* *Weighted*	Orient. D-Val *Mean-* *Weighted*	# Edge Pts. (> 0.1%) *Mean-* *Weighted*
Aframe	0.223	0.115	243 (76%) [0%]
Brake	0.115	0.032	345 (71%) [-7%]
Funnel	0.099	0.052	166 (78%) [+2%]
Grill	0.139	0.097	251 (53%) [-9%]
Guide	0.295	0.121	323 (66%) [-12%]
Pipe	0.221	0.08	167 (62%) [-7%]
Plug	0.134	0.1	177 (62%) [-5%]
Pot	0.284	0.043	225 (61%) [-6%]
Pump	0.266	0.113	169 (53%) [-7%]
Stand	0.118	0.158	354 (78%) [-2%]
Tidy	0.118	0.05	258 (71%) [-7%]
Tray	0.233	0.081	320 (67%) [-8%]
Valve	0.268	0.288	176 (43%) [-10%]
Widget	0.125	0.07	163 (57%) [-6%]
Wiper	0.143	0.087	278 (66%) [-7%]
Mean Values	0. 185 (+8.4%)	0.099 (+0.8%)	64% [-7%]

Figure 4.12: *The table above represents the same data as in Table (Figure) 4.11 for the mean-weighted entries, except that lateral feature shifting has been deactivated to allow the utility of the process to be practically assessed. The difference in the number of data points passing the hypothesis test at a 0.1% confidence limit, as compared to the previous table, is detailed alongside each entry. The mean maximum divergence of each sample from the predicted distribution (D-Vals) relative to the laterally optimised features in the previous table is indicated in brackets at the base of the respective columns.*

$$F_n(p) = p \sum_{i=0}^{n-1} \frac{(-\ln p)^i}{i!} \tag{4.19}$$

i.e.

$$H_{angle,edge} = (H_{angle} H_{edge})(1 - \ln(H_{angle} H_{edge})) \tag{4.20}$$

where H_{edge} and H_{angle} are the hypothesis probabilities for edge location and orientation respectively.

This metric can be evaluated for any sampled edge feature point relative to a model feature prediction to reflect the probability that the point was indeed representative of a valid model match. Since the assumed likelihood models can be seen to closely reflect those observed, we can be confident that such a hypothesis test will give a good indication of the true probability of occurrence. In practical terms, the D-value errors sampled in Table (Figure) 4.11, represent the worst case error in probability for those sets of edge features. Each such value can simply be multiplied by 100 to give percentage errors. For instance, even when using the mean term weighting values, the worst case mean error across all the sampled objects is only around 10% for both edge location and orientation terms. The average error for each object is likely to be significantly less. If greater precision was required, individual scaling factors may otherwise be applied to specific objects. As part of a learning framework, any such scaling factors could be continuously moderated through temporal experience for optimised discernment.

4.11 Verification of Projected Views

Given the proposed metric for assessing point to point oriented edge feature correspondence, such evidence can be accumulated across sets of representative curves to support verification of detected object models in images. For instance, thresholds can be set against the proportion of model features passing the hypothesis test at a specified confidence limit. Although there is some degree of flexibility regarding confidence limits and proportionality constraints, in current work, a feature is deemed to be present if at least 45% of its sampled feature points pass a joint hypothesis test for edge location and orientation at a 1% confidence level. The confidence level allows virtually any suspected edge feature to be passed, while the 45% threshold accounts for feature degradation due to factors such as partial occlusion, interference from background clutter or specular highlights. This mechanism supports the semi-automated acquisition and refinement of any view-based visibility files, for which the system is able to assess which 3D wireframe model features are visible from certain views through repeated exposure.

$$\textbf{VerificationScore} = \frac{\sum_{i=1}^{n} S(H_i - H_{limit})}{n} \tag{4.21}$$

where $S()$ represents the Heaviside step function, i.e. returning 0.0 for a negative argument and 1.0 for a positive one, n is the (fixed) number of sample points for the projected view-based wireframe **model (or feature)** in the image plane (e.g. every 4th pixel along (or around) each image projected 3D edge feature) and H_i represents the hypothesis probability: either H_{edge_i} or $H_{angle,edge_i}$ depending on the specified test.

Further to the experimental findings in the previous section, the requirement for a process of lateral feature shifting is further exemplified by the examples in Figure 4.13. Using a maximum potential lateral shift of 3 pixels for each feature, the feature verification results show that a significant number of extra features are correctly verified as being present if using lateral shifting. Although this extra representational freedom may occasionally erroneously validate other features as being present, analysis through multiple views should alleviate any such problems. It is reiterated that degrees of lateral shifting should be learned specific to particular features through multiple views, with some features expected to be fixed in predicted location.

Directly accounting for the proportion of features deemed visible is not however the most appropriate way to assess the quality of match of a projected view, unless each feature is weighted by its length. More straightforwardly, for view verification or match disambiguation, the projected model features can be treated as a single feature, with the proportion of sampled feature points verified as valid being taken as a direct measure of match quality. Deciding what proportion of feature points is required for positive verification depends on the strength of the test and to some degree the nature of the object. In the examples detailed in Table (Figure) 4.11, for instance, only approximately 70% of each object's projected feature points are

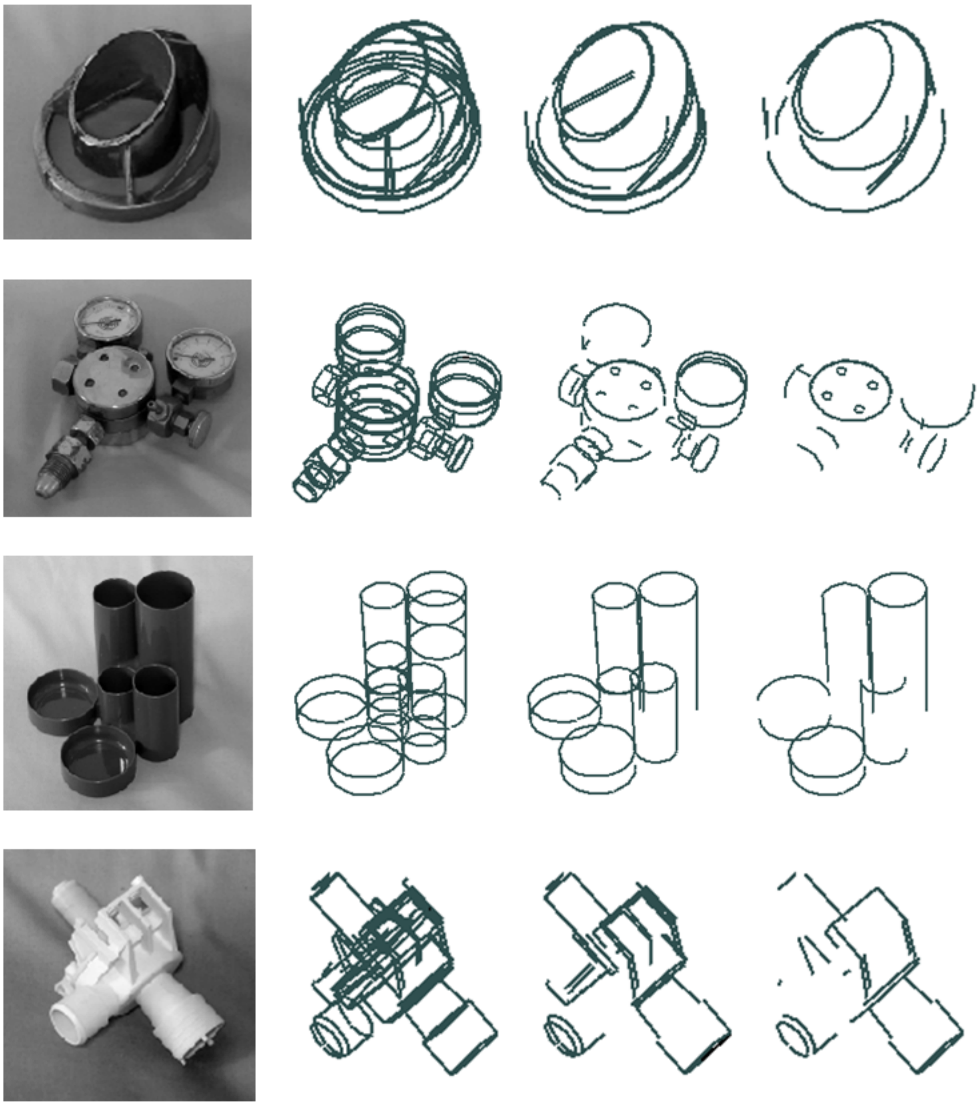


Figure 4.13: *The 4 imaged objects are accompanied by their full 3D edge feature-based wireframe models immediately to their right. The 2 rightmost columns represent those features from the full wireframes that are automatically deemed to be present with and without (left to right) a process of optimised lateral feature shifting (up to 3 pixels) in the image plane. Features are verified as present if 45+% of their sampled feature points pass a joint hypothesis test for edge location and orientation (weighted as per Figure 4.10) at a 1% hypothesis test confidence limit.*

verified as being present under the joint edge location and orientation test with a 0.1% threshold. This can be due to corroded physical features or interference from illumination or background elements. Further complications arise when attempting to verify the presence of heavily occluded objects, where, perhaps, only 30% of the features might be visible. This further justifies the expected enhanced discriminability offered by the supporting edge orientation constraint. In such extreme examples, it may be that other cues such as texture, surface shading or colour may be required in support of verification of the presence of any such object from significantly fragmented evidence.

The justification for the proposed approach to edge model match verification follows from that used in support of localisation. Indeed, the distance-based cost functions commonly used in the computer vision literature for localisation are typically directly used for subsequent verification purposes, i.e. with the optimised minimal distance representing the quality of match. Once again, such distance-based approaches and other ad hoc edge feature correspondence approaches are likely to be sub-optimal for the task, lacking the discriminability required as the basis of unambiguous differential model match verification in noise corrupted real-world scenes. In summary, it is proposed that the only theoretically justified means to assess edge feature occurrence is via direct account of the underlying image evidence for well aligned object models within a self consistent probabilistic framework. As indicated, under this model of feature detection, some edge features may require a degree of variability in their predicted positions, as modelled by a lateral shift.

4.11.1 Summary of the Proposed Method for the Verification of Projected Views

- Hypothesise the projected, view-sampled image pose of the 3D wireframe object (see Section 6.6 - to follow)
- Optimally localise the model in the image plane (See Section 4.8)
- Sample a verification score according to Equation 4.21 for a set of uniformly sampled reference points
  - e.g. every 4th pixel along each projected 3D edge feature (relative to projected scale)

The bearing of confidence limits on each sampled edge feature point hypothesis test and the utility of the supplemental edge orientation metric is experimentally reviewed in the context of the proposed view-based model matching system in Chapter 7 (7.7).

4.12 Conclusions

This chapter has introduced a strategy for the view-based wireframe modelling of rigid edge-defined 3D objects. These models are required as the basis of stereo model matching and view-sampling for localisation, verification and view-based learning purposes. In line with aspect graph theory, objects' view-spheres are partitioned into discrete view sets, with clustered feature visibility determined via a nearest view matching strategy. Techniques to represent view-based features such as occluding boundaries for conical sections have also been presented. In the context of precise quantitative model to image alignment, a process of lateral feature shifting has been presented as a solution to dealing with various modelling simplifications and illumination dependencies. In allowing lateral feature shifting, the proposed approach can be understood as an appearance model for detectable edge structure.

In attempting to model the features that can be practically extracted from images and their allowed variability in scenes, it is proposed that the methodology is well suited to the task of extracting quantitative information from images in the context of a general system that must learn. Though this avoids surface modelling, it is expected that it will be possible to infer some qualitative surface information from multiple views of an object under variable illumination later, once basic competences for recognition and location of objects and features have been established. The requirement to model surfaces is therefore moved out of the representation for 'bottom-up' view-based recognition and location, with this aspect of scene interpretation being left for later. It is suggested that surface identification in an image should proceed by recognition of scene contents followed by 'top-down' testing of image data with regard to generated surface hypotheses. This would lead to surface detection on the basis of an absence of quantitative evidence which invalidates the hypothesis, rather than measurement based upon the presence of detectable features.

Having established a representational basis for the view-based modelling of (rigid) 3D objects, the remainder of the chapter has detailed the proposed scheme for accurately aligning and quantitatively verifying image projected instances of such models. These tasks are intended as post model matching processes, in support of approximate localisation and orientation cues delivered by both the TINA stereo and view-based model matching routines. Observing that there are significant shortcomings to conventional edge distance-based approaches to edge feature model localisation and verification with regard to intrinsic validity and quantitation, a novel statistical framework has been developed which incorporates both the effects of image noise and local image structure. This approach supports the construction of a joint probability for the degree of conformity of image data to both edge orientation and location, without the need for arbitrary relative scale factors. The method has been validated on multiple views of man-made objects constructed from a variety of materials. It is understood that this is the first time in computer vision that the likelihood models used in object alignment have been shown to quantitatively accord with the corresponding sample distributions.

Chapter 5

3D Object View-Sphere Sampling with Pairwise Geometric Histograms (PGHs)

5.1 Introduction

Previous research has proven the viability of the Pairwise Geometric Histogram (PGH) representation for 2D edge-defined object recognition, or equivalently; single-view-based 3D object recognition (see Chapter 3). The main issue addressed by the research in this chapter is adaptation and extension of associated techniques for the modelling and recognition of 3D objects' projected appearances around their complete view-spheres with regard to in-depth rotation, scale and perspective. The chapter also introduces an amendment to the dual-directed PGH format (see Chapter 3) to solve for ambiguity in the assignment of histogram entries for sample lines lying parallel to the reference line.

As detailed in Chapter 3, PGHs are specified for linear edge segments, detailing the relative perpendicular positions and orientations of any other linearised edge segments in the reference line's local projected area. Curved edge features are approximated by line segments, with robustness to relative linear-misalignment induced by PGH bin entry blurring. An object's projected shape is represented by assigning multiple locally scoped PGHs across the 2D shape's constituent linearised edge segments. Although only a limited cross section of edge segments may be required to support object recognition, by encoding the majority of features, recognition is better supported through extreme cases of occlusion, where recognition must be based on a minimal random subset of features.

Despite there being a number of possible PGH formats, as detailed in Chapter 3, the most descriptive is the reference line directed format with separate histograms for each side of the reference line (Figure 3.4). This PGH format represents the limit of information that may be encoded by a PGH with the required invariances and has proven most useful and discriminatory for 2D object recognition in previous research. The proposed capacity and completeness properties of PGHs are also specific to this format. In keeping with development toward a real-world-oriented 3D object recognition system with a vast repertoire of recognisable objects, this more discriminating PGH format is therefore considered exclusively herein as the basis of the proposed view-based recognition system.

The PGH representation has been engineered to degrade stably and smoothly through locally continuous object rotation and deformation. This means that the established PGH-based techniques for fixed-view shape recognition may support recognition of views of 3D objects through proximal regions of view-space with precision commensurate with misorientation. The key issue is determining the minimum level of precision to be tolerated across objects' learned PGH-bound shape vector manifolds in support of recognition and localisation. If sufficient precision can be attained for wide enough centre-view deviations, then only a minimal number of raw PGH-based reference nodes may be required to support recognition of 3D objects' features across their view-spheres with a nearest neighbour search strategy. Otherwise, information from adjacent views may be interpolated across the continuous shape manifold over a reduced set of reference nodes to aid more efficient recognition and to support more informed object pose determination and localisation.

The overriding idea with this work is that accurate 3D wireframe models representing objects' defining edge features are available for research analysis to aid precise control over all projection and viewing parameters without the need for robotic handling and inspection equipment, which is unavailable for this project. It is assumed that such models are relatively trivially acquirable, for instance through stereo analysis, with this aspect of development being set aside for future research and wireframes being manually composed with view-dependency files. It is proposed that the techniques developed will be directly applicable for autonomously learning to recognise equivalent tangible objects with interactive robotic machinery. The learning methodology should also be extendible to deal with arbitrary 3D CAD models.

5.2 Shape Manifolds in PGH Vector Space

A PGH is essentially a high-dimensional vector with dimension of the number of histogram bins. Once formed, a rectangular PGH can be converted to a more conventional extended vector by placing adjacent rows end to end. Any pattern encoded across a PGH's bins can therefore be treated as a point in this high-dimensional parameterised space. By normalising the PGH, a point may be further constrained to lie on the positive quadrant of the surface of a normalised hyper-sphere, aiding regularity of the representation across the view-sphere.

Because the PGH representation has been engineered to deform relatively stably and smoothly with localised distortions in viewing parameters, as an object's normalised-PGH-encoded features are locally rotated or scaled, the point on the hyper-sphere surface in representation space should move accordingly in a smooth continuous form, essentially carving out a locally smooth object-view-representative shape manifold. The task of shape recognition can then be reduced to the modelling and recognition of such low-dimensional shape manifolds in the otherwise very high-dimensional representation space. In theory, by maintaining parameterised representations of these low-dimensional continuous manifolds, we can conveniently infer objects' continuous ranges of appearance for recognition without the need for storage and matching of potentially very many sample PGHs. We are essentially aiming at continuously interpolating objects' projected appearances across their view-spheres.

One problem with 3D projected shape interpolation is that new edge features may continuously come into or go out of PGH-bound view under object/view rotation so that arbitrarily complex histogram components may disrupt any more uniform interpolation procedures. This will result in discontinuities in the global manifold representing an object, partitioning the object into more stable 'aspects'. In contrast, for transparent objects, all features would be visible around the view-sphere so that no such discontinuities are introduced and a more continuous hyper-surface may be formed. Because of factors such as sub-object symmetries, shape manifolds may also be susceptible to self intersection. The possibility of arbitrary high-dimensional projected shape dictates that it is impossible to predict how a PGH-based shape vector manifold may manifest itself in terms of its form and number of PGH-bound dimensions, making it very difficult to learn to model such non-linear distributions in the general case. However, as observed in [37], a shape's representative hyper-surface will locally have as many dimensions as there are degrees of freedom in the viewing parameters, minus any invariances that the representation may have. Although many dimensions may be required to accurately model the extended projected effects of rotation about a point on an object's view-sphere, there are 2 main components of variance relating to the latitudinal and longitudinal directions of rotation at any point on a view-sphere's surface (i.e. we can only go about two orthogonal directions from any point on the surface of the earth, which is locally flat). Locally, at least, in accordance with the invariances of the PGH representation, there should therefore be 2 principal modes of approximately linear PGH variation relating to object rotation at each point on the view-sphere. This chapter goes on to investigate the potential for modelling such low-dimensional information in support of constructing continuous shape manifolds mapping out 3D objects' projected edge-based appearances for recognition.

The assumption that PGH-vector-based shape manifolds are locally smooth, continuous and low-dimensional can be validated by examining sets of PGHs uniformly sampled across localised regions of the view-sphere. Principal Components Analysis (PCA) is a standard statistical method for analysing high dimensional sets of data in order to identify any significant global modes of variance. PCA provides a convenient means of observing the statistical character of any presented data sets in terms of a low-dimensional orthogonal linear transformation. If any linear global modes of variance are observable then they may be sampled as the basis of recognition. With regard to the current problem of characterising the nature of any shape manifolds formed through 3D object rotation, clusters of uniformly spaced views can be sampled around specified object viewpoints. The eigenvectors from the covariance matrix of the set of sampled PGHs can then be extracted and those with the highest eigenvalues can be coordinated as the low-dimensional axes of global variance. The working hypothesis is that projected shapes' features' manifolds should be locally planar on account of the latitudinal and longitudinal dimensions of rotation realisable at the surface of the view-sphere.

For these experiments, as explained in context in Figures 5.1 and 5.2, a selection of object views have been taken and a PGH has been sampled for a specified (linear) edge feature in each case. A pentagonal reference region has then been sampled around each initial PGH-sampled viewpoint (with adjacent vertices spaced approximately 24 degrees apart), which has then been iteratively triangulated to give 181 uniformly distributed sample vertices. A PGH has then been sampled at a viewpoint corresponding to each vertex and the set of 181 PGHs has been analysed using PCA. The process has then been repeated for 2 similar smaller pentagonal reference regions, maintaining 181 sample viewpoints throughout each.

For the broader pentagonal PGH sample regions in these experiments, there is more opportunity for features to come into or go out from PGH-bound view, thus disrupting any uniform interpolation. To aid wider field interpolation, only features that are visible throughout all sampled viewpoints can be sampled across the manifold. This further aids the assignment of localised feature sets in support of occluded object recognition where recognition may otherwise be confused across full model templates. The bottom left diagram in Figure 5.1 highlights in blue the reduced set of features jointly encoded by PGHs across the widest sampled view-sphere manifold. As indicated in the corresponding PGH vector manifold graph this feature consistency constraint improves the smoothness, scope and accuracy of the 3D eigenvector space representation for wider views. The object features similarly examined in Figure 5.2 are accordingly sampled with reduced sets of features that are visible throughout each aspect. The results are quantitatively analysed in Figures 5.3 and 5.4.

It is observable that some objects' features' PGH-based planar manifolds' third dimensions are fourth (or potentially even higher) order. There is evidently a reasonably regular occurrence of interference from other sources of correlated complexity, typically when the 3rd eigenvector's eigenvalue is confused in the noise field. Many of the interfering dimensions pose highly correlated curved manifold structures. This raises the question whether more analogous free-form manifold modelling processes could be employed to model these other dimensions, or indeed all dimensions coherently as prospectively required for high fidelity wide-field interpolation. Relatively high degrees of correlated accuracy can however be attained in the first 3 dimensions for wide enough sampled regions to support triangulation of the manifold for recognition.

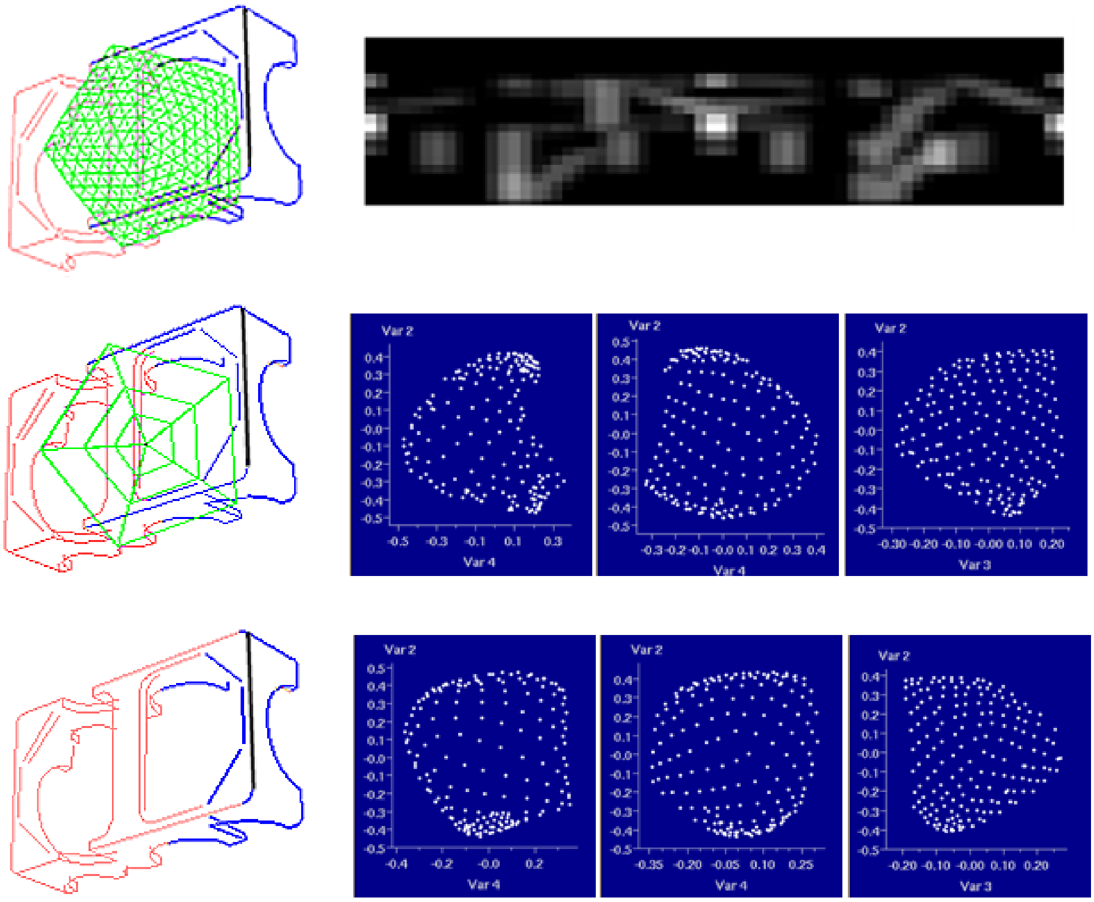


Figure 5.1: *The upper row of the diagram above indicates a 3D wireframe object with a corresponding dual-directed format PGH (20 by 64 bin). The PGH is sampled for the rightmost vertical feature shaded in black, with the features included in the PGH being shaded in blue. The green web imaged over the object shows the largest of 3 pentagonal reference regions sampled with 181 PGH reference points at the vertices. The 3 scaled pentagonal reference regions are displayed in the wireframe diagram beneath, with 181 PGHs being uniformly sampled across each. The blue shaded data plots indicate head-on orthogonal projections of the 3D manifolds pertaining to the top 3 (or 1st, 2nd and 4th) eigenvectors (including the mean) of the 181 PGHs sampled throughout each of the 3 scaled pentagonal sample regions (largest to the left). Table (Figure) 5.3 details the percentage of information (in terms of an eigenvalue sum) embodied by each 3D manifold relative to scale. The object image at the bottom left shows the reduced set of object features that are mutually visible by each view in the widest pentagonal aspect. The top set of data plots indicate that the planar projected manifold is disrupted by inconsistent incorporation of extended sample line features for wider field interpolation. By removing any inconsistently visible features from the interpolative sample set, as indicated in the bottom row, the planar manifold can be better approximated over wider fields with less distortion.*

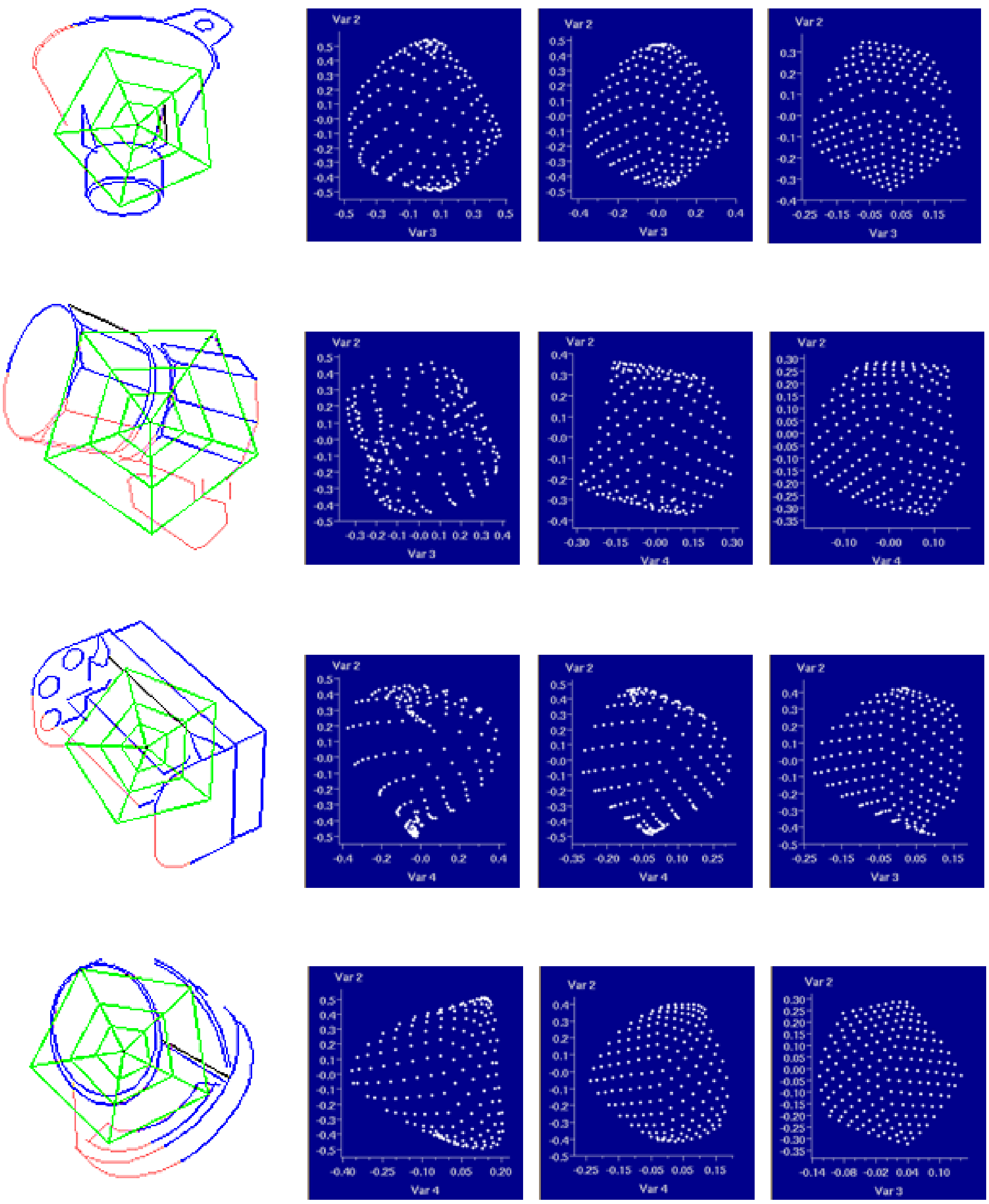
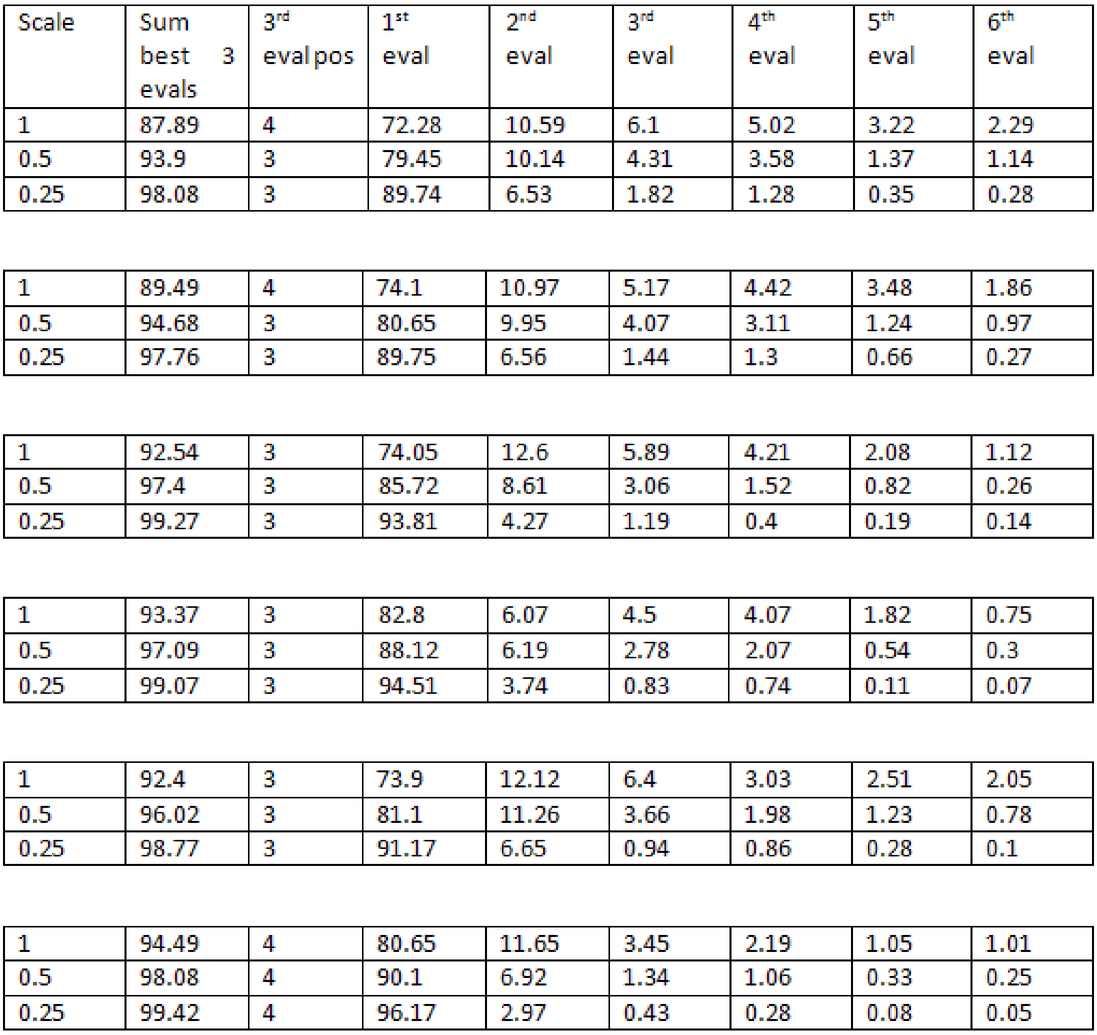


Figure 5.2: *For each 3D wireframe object in Figures 5.1 and 5.2, 181 PGHs are sampled uniformly throughout each of the (green) scaled pentagonal sample regions for the black line reference features and associated blue shaded reference regions. Only geometrical features that are visible throughout each pentagonal sample region are included in the respective PGH sample sets. The left column of data plots corresponds to the widest sample region with approximately a 24 degree outer pentagon-vertex separation about the view-sphere origin. For these wider fields of PGH scope it is observable that the isotropic planar projected distribution can be distorted in irregular non-linear 3D form. The uniform 2D projected mapping can however be more accurately maintained for more localised sample regions as indicated in the rightmost data plots. The bottom right data plot, for instance, displays a very uniform mapping of view-space. Each pentagonal region would be split into its 5 constituent base-triangles, allowing for the manifold to be conveniently sampled for continuous interpolative triangulation.*



Scale	Sum best 3 evals	3rd eval pos	1st eval	2nd eval	3rd eval	4th eval	5th eval	6th eval
1	87.89	4	72.28	10.59	6.1	5.02	3.22	2.29
0.5	93.9	3	79.45	10.14	4.31	3.58	1.37	1.14
0.25	98.08	3	89.74	6.53	1.82	1.28	0.35	0.28

1	89.49	4	74.1	10.97	5.17	4.42	3.48	1.86
0.5	94.68	3	80.65	9.95	4.07	3.11	1.24	0.97
0.25	97.76	3	89.75	6.56	1.44	1.3	0.66	0.27

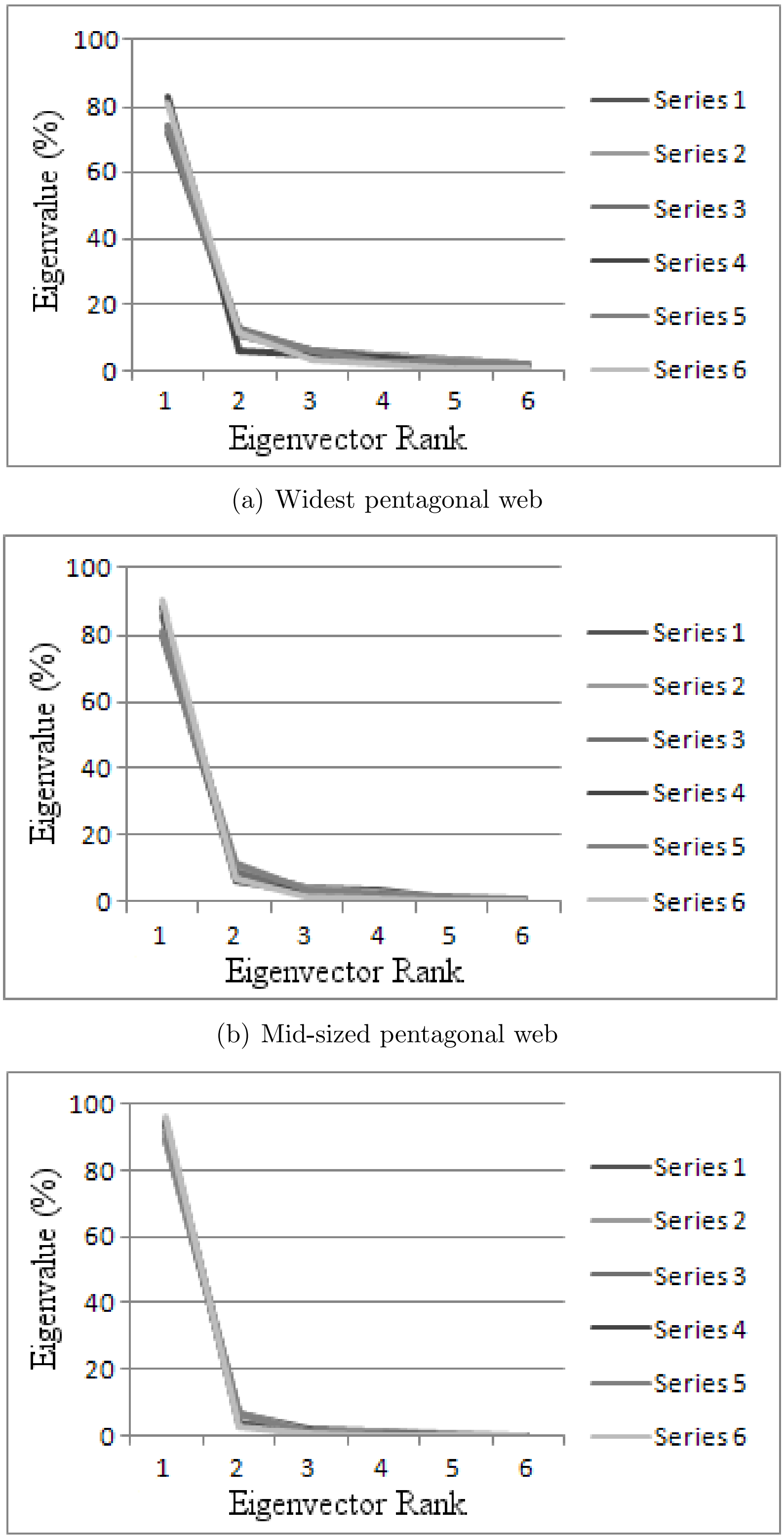
1	92.54	3	74.05	12.6	5.89	4.21	2.08	1.12
0.5	97.4	3	85.72	8.61	3.06	1.52	0.82	0.26
0.25	99.27	3	93.81	4.27	1.19	0.4	0.19	0.14

1	93.37	3	82.8	6.07	4.5	4.07	1.82	0.75
0.5	97.09	3	88.12	6.19	2.78	2.07	0.54	0.3
0.25	99.07	3	94.51	3.74	0.83	0.74	0.11	0.07

1	92.4	3	73.9	12.12	6.4	3.03	2.51	2.05
0.5	96.02	3	81.1	11.26	3.66	1.98	1.23	0.78
0.25	98.77	3	91.17	6.65	0.94	0.86	0.28	0.1

1	94.49	4	80.65	11.65	3.45	2.19	1.05	1.01
0.5	98.08	4	90.1	6.92	1.34	1.06	0.33	0.25
0.25	99.42	4	96.17	2.97	0.43	0.28	0.08	0.05

Figure 5.3: *The data above corresponds (respectively) to the 6 sets of multi-scale data plots presented sequentially in Figures 5.1 and 5.2. The scale relates to the size of the pentagon sample region as indicated in green over the specified models, with 1.0 being the outermost (largest) pentagon. The top 6 eigenvalues for each sample set are given in percentage terms with a percentage sum of the total information held by the top 3 eigenvectors (including the sample mean) supporting an approximately isotropic planar mapping. It can be seen that nearly all of the information in a sampled PGH data set is held by these 3 dimensions for more localised regions of view-space.*



(a) Widest pentagonal web

(b) Mid-sized pentagonal web

(c) Smallest pentagonal web

Figure 5.4: *The graphs above indicate the eigenvalues (in relative percentage terms) of the top 6 eigenvectors for the 3 sizes of pentagonal reference region in Figures 5.1 and 5.2 (1.0 is the widest and 0.25 the smallest) for the 6 randomly sampled objects. Almost all of the information can be seen to be contained in the top 3 eigenvectors for the smallest pentagonal sample regions.*

5.3 Amendments to the PGH Format

As indicated in the previous section, local view point clusters for fixed sets of features should carve out smooth, continuous manifolds in PGH vector space. Analysis of such distributions during project development however identified a subtle contravening problem with anomalous data points relating to histogram entries for lines lying parallel to the reference line. With initial, incidental regard to the rotation-invariant PGH format (see Chapter 3), pairs of lines are treated as vectors that point away from the point of intersection of the two (extended) lines. Relative angles and positive and negative handedness can only be unambiguously determined relative to the directions of these vectors. For parallel lines, there can however be no point of intersection, meaning that any assessments of handedness (i.e. positive or negative relative distances) are arbitrary. Without appropriate consideration, this means that arbitrary shifts may be introduced across positive and negative distance axes for histogram entries relating to parallel lines. To deal with this PGH instability problem, any values occurring at histogram bin angle extrema (relative to bin blurring parameterisation) can be weighted across negative and positive axes. Alternatively, it could be suggested that parallel lines be assigned a particular histogram side. In this case, the potential for noise on any line parameterisation assessments would mean that almost parallel lines could be assigned to the wrong side of the histogram, invalidating the robustness of such an approach for consistent matching.

Although such issues are exempted by the directed PGH format because the line direction and handedness are fixed, the full extended version of the directed PGH suffers from related problems. As detailed in Chapter 3, this PGH format has 2 distinct base-histograms that separate sample lines directed towards each side of the reference line. Matching is achieved by sampling each image line against this representation in both directions. Because it is impossible to determine which side of the reference line a parallel sample line points, arbitrary shifts may be introduced across the two halves of the representation. To remedy this, the weights of any sample lines with an incident angle of less than the angular bin width ($\pi/32$) can be proportionally distributed through the respective halves of the PGH, ensuring a smooth transition of data across connected bins. For instance, if the line was sampled as being perfectly perpendicular, it would be weighted equally in each half, or a sample line that was estimated to be $\pi/64$ degrees apart from the reference line would be three-quarter weighted in the side from where it came and one-quarter weighted in the opposite half of the PGH. The solution is explained in more detail in Figure 5.5.

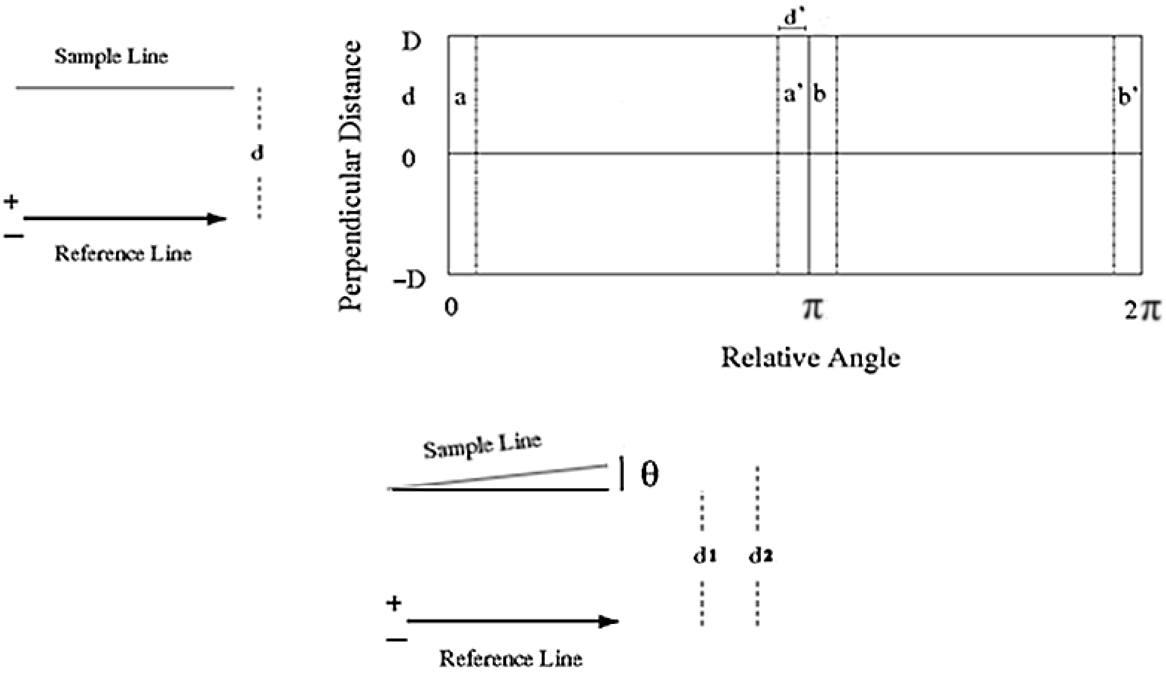
5.4 Interpolation of Projected Views

The process of interpolation serves to support continuous modelling of object representations via a reduced number of reference nodes mapping out the shape manifold. Object recognition should therefore be more efficient, because less information is required for storage and matching, and more accurate, because the precise corresponding 3D view of the object can be estimated and used for localisation and verification. Previous research [46] has shown how scale invariance can be approximated by interpolating between PGHs sampled at a range of desired scales. Before addressing any issues relating to scale deformation for 3D objects, the more immediate problem of modelling PGH deformation through in-depth object rotation is now addressed.

In context, Figure 5.6 shows a 3D object with 3 PGH sample points at the viewpoints indicated at the corners of the green reference triangle. The corresponding PGHs for each sample point are detailed alongside. It should be clear that various elements of each PGH are distorted and displaced through object/view-point rotation. The task of interpolation therefore involves accounting for how such PGH structure deforms and displaces continuously across such sparsely sampled regions. As indicated in section 5.2, these manifolds are expected to be locally planar, distorting progressively more non-linearly through extension around the view-sphere.

Perhaps the main issue with interpolating histogram structure is that of linear versus non-linear methods. Linear methods are popular in computer vision for such purposes because all they require is a set of raw sample nodes from which enclosed correspondence is attained directly, for instance via triangulation. Data is thus treated uniformly, affording a high degree of control over system uncertainty. However, as shown in Figure 5.7, linear interpolation methods are restricted to localised regions of the view-sphere, as further illustrated in Figure 5.8. While it may be convenient to build a system around low-dimensional locally-linear connected sub-systems, ideally, one would be able to represent PGH deformation more continuously without any fall off in accuracy.

In order to continuously model objects' features' PGHs it would be necessary to spe-



```
angle_bin_weight = line_length( Reference Line ) * line_length( Sample Line )

d' = Angle axis bin width (in accordance with the prescribed level of blurring (currently π/32))

α = θ / d'

if ( α < 1 )
{
  weight_a = weight_a' = ( angle_bin_weight / 2 ) + ( α * angle_bin_weight / 2 )
  weight_b = weight_b' = angle_bin_weight - weight_a
}
```

Figure 5.5: *For the dual-directed PGH format for which sample lines are partitioned into respective halves of the histogram (0 to π or π to 2π) according to which side of the reference line they are directed, instability problems may arise when assigning values for sample lines lying parallel to the reference line, where $\theta = 0.0$. Image sample lines that are almost perpendicular can therefore be weighted across corresponding halves of the PGH to approximate the likelihood that they actually emanate from that half of the representation. The formula used to linearly approximate the spread of such weightings across corresponding extremal PGH bins is presented as pseudo-code.*

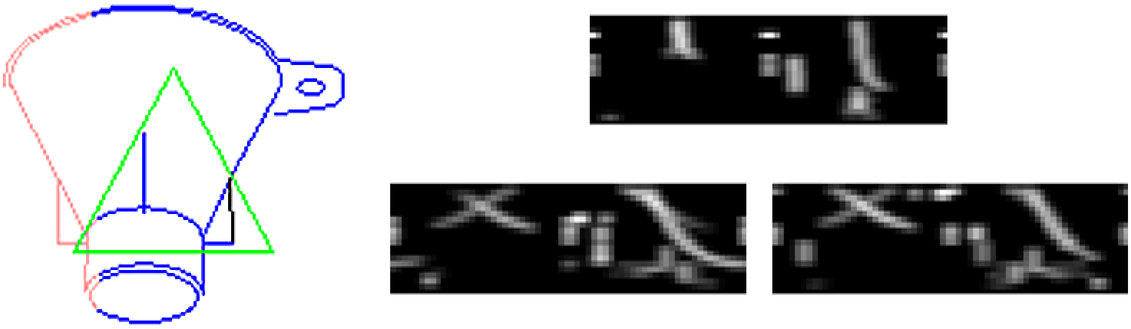


Figure 5.6: *For the 3D wireframe object imaged to the left, the vertical line feature in black to the mid-right of the object has been selected and a corresponding PGH has been sampled at a viewpoint corresponding to each corner of the green triangle, as indicated on the right hand side. The object features in blue are those that are included across each PGH sample. The deformation of PGH structure can clearly be observed across the three PGH samples.*

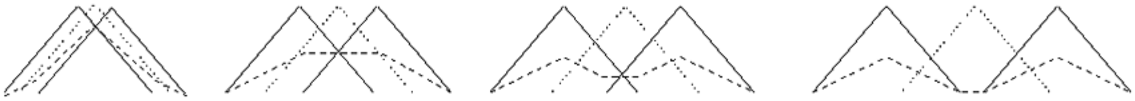


Figure 5.7: *The triangles above in bold represent pairs of displaced data distributions that are used as samples for exemplary linear interpolation. The dotted triangle in-between each pair represents what the data would like if it could be perfectly interpolated. The corresponding linearly weighted interpolation (mean) is shown as the overlaid broken line. A different linear scale of interpolation would otherwise skew the reconstruction to one side. Although linear methods can approximate local overlapping shifts in structure as indicated in the leftmost diagram, as reference structure diverges, linearly weighted interpolations are invalidated for displaced structure as indicated in the rightmost diagram.*

cify the geometrical particulars of the possibly high-dimensional curved manifold for each region of referenced view-space (somehow). Kernel-PCA, for instance, defines a kernel (a fixed non-linear continuous surface primitive) about which data may be reconstructed (effectively pre-transforming the data) before more traditional linear methods are implemented. If our local rotation sample data was for instance bound to lie uniformly on a sphere in representation space, we could use spherical geometry to uniformly cast each sphere-surface point into an isotropic Euclidean 2D mapping (maintaining neighbouring point separations) and then perform triangulation across this uniform planar manifold. As the data presented earlier in the chapter suggests, the problem here is that the high dimensionality of the PGH representation space and the potential for arbitrary PGH-sampled shape dictates that any shape representative manifolds may be arbitrarily complex high-dimensional free-form surfaces that simply cannot easily be modelled and parameterised with any reasonable degree

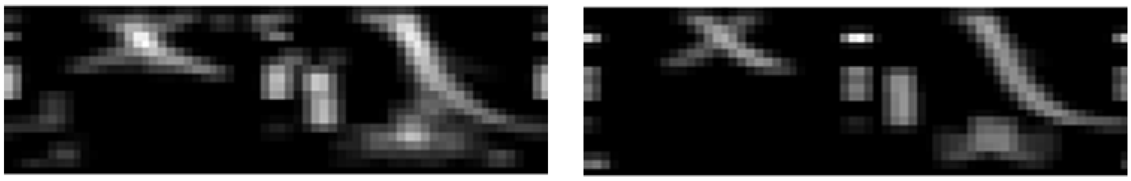


Figure 5.8: *The PGH to the left is the mean PGH sampled from the corners of the triangle in Figure 5.6. The PGH to the right is the actual PGH sampled at the corresponding position at the centre of the triangle. The limitations of linearly weighted reconstructions are evidenced by a blurring of corresponding PGH elements (in the left PGH) and other erroneous structure. Linear weightings cannot faithfully predict how structure may deform between samples. A reasonably accurate reconstruction of the main elements has however been maintained with the Bhattacharyya match score between the 2 presented histograms being maintained at 0.9, which should be sufficient to support object recognition.*

of precision. In attempting to continuously model such highly correlated, complex, high-dimensional free-form manifolds we introduce a whole new realm of abstraction and complication to the process of recognition.

Ideally, a graphical ‘morphing’ type process could be inferred to predict exactly how a PGH may appear at any point within a specified (outer-sampled) reference region. The problem with morphing is that correspondence needs to be inferred between corresponding features. While this may be tractable in a simple case with discrete blocks of histogram structure, PGHs tend to be of a highly complex, interwoven form, making any such task generally intractable. Other researchers have attempted to account for matching of such complex non-linear histogram-bound distributions with algorithms such as ‘earth moving’ [97] and ‘diffusion’ [95]. As the name suggests, the earth moving algorithm solves the transportation problem of moving one distribution to the space of the other (in terms of bin values), which can be a very costly process. Diffusion-based algorithms treat the two distributions as heat fields, assessing how much heat diffusion is required to balance the two opposing distributions. Despite these methods exhibiting facilitation of non-linear complex histogram matching, they are not statistically principled metrics for this purpose and do not produce results in accordance with a consistent statistical (probabilistic) interpretation as required for statistically valid interpolation and interpretation.

There are potentially many ways in which loose correspondence may be sampled between histograms for recognition, but clearly there should be one appropriate underlying metric that all others tend to. In previous research, the Bhattacharyya metric has been shown to be the appropriate metric for comparing histogram distributions

[45] in Euclidean terms, i.e. in line with probability; scaling linearly from 0.0 to 1.0. The problem with the Bhattacharyya metric is its bin to bin correspondence nature. Often, correlated PGH structure may shift laterally across the PGH (cross-bin deformation), meaning that despite blurring and bin quantisation, 2 topologically similar PGHs may have a very low Bhattacharyya overlap score. In keeping with use of the Bhattacharyya match metric, the task becomes one of inferring continuous ranges of PGH-bound appearance so that a PGH can essentially be continuously reconstructed across the reference region and compared via the Bhattacharyya metric to indicate the quality of match. This differs from methods such as diffusion and earth moving because we aim to continuously model how a PGH may legitimately manifest itself (to an appropriate degree of precision).

Avoiding complications embedded in the domain of high-dimensional free-form manifold modelling, it is suggested that PGH interpolation be mediated by a piecewise, locally planar approximation, effected by view-sphere triangulation. By ensuring that a tolerable reconstruction error is maintained across each sampled triangular region it should be possible to continuously and accurately model objects' PGH-bound shape manifolds, with the only inconvenience that a relatively large number of samples may be required as compared to any hypothetically valid non-linear modelling schemes. Although a perfectly complete and faithful continuous representation of an object's features would be convenient for recognition, this would be too rigid a constraint in that the same associated tasks could be performed without any significant loss of performance if using, for example, 98% of the base information. Such fall-offs in information should effectively be negligible in the framework of a recognition system that is geared to be robust to relatively extreme distortions in input data (i.e. depleted, fragmentary object-edge profiles). From the outset, object recognition can therefore be based on the assumption that a minor fall off in representational accuracy can be tolerated without any significant impairment to system performance, validating the use of linear approximations.

The level of precision required for object feature recognition critically depends on the diverseness of the object library and the accuracy required in support of 3D localisation. If only a few visually distinct objects are required for recognition, then relatively lax constraints on representational accuracy may be tolerated without much effect on differential recognition. If the recognition system is set to utilise more global edge template optimisation schemes for final model-scene alignment (See Chapter 4), then so too can looser constraints on the faithfulness of the representation be imposed, without any significant loss to overall recognition and localisation performance.

Another key issue is that of whether interpolation is even required, with regard to the potential of an alternative nearest-neighbour type PGH sampling strategy. Although many more raw PGH nodes may be required for a nearest neighbour recognition system (maintaining a common minimal matching error), no work would be required in maintaining and matching the connected manifold. The time spent analysing potentially complex individual manifolds could therefore be invested in raw matching. This computational trade-off is examined later in the chapter (5.8). Although if a fine enough sampling of PGHs was maintained, a recognised object's corresponding view-point could be reliably inferred, there remains the issue that it is desirable to have as few nodes as possible meaning that interpolation may always be required as an end process to infer the matching view-point of an imaged 3D object. Although global projected wireframe view alignment optimisation processes may be implemented after recognition to aid precise object localisation (see Chapter 4), more complex scenes may require accurate initial predictions of model localisation so as not to be caught up in local minima in the noise field.

5.5 Linear Interpolation

As discussed, PGHs are treated as traditional (high-dimensional) vectors, with adjacent rows of each PGH being sequentially concatenated. As indicated above, PGHs should deform approximately linearly with localised in-depth object rotation, forming a regular planar mapping of the data about a reference vector. Local PGH deformation, relative to viewpoint, can therefore be modelled in a tangent space with interpolation performed via triangulation across the plane.

In accordance with the Bhattacharyya match quality metric, each sample PGH vector ($\mathbf{v}$) is initially normalised and square rooted ($\mathbf{h}$)

$$h_i = \sqrt{\frac{v_i}{\sum_{j=1}^{n} v_j}} \tag{5.1}$$

where h_i is the value of the i^{th} histogram bin for $\mathbf{h}$ (similarly for v_i and v_j).

Variation in a histogram ($\mathbf{h}$) can therefore be modelled across a reference region as a weighted sum of a base vector ($\bar{\mathbf{h}}$) and two orthogonal reference vectors ($\mathbf{e_1}, \mathbf{e_2}$).

$$\mathbf{h} \ = \ \lambda_0 \bar{\mathbf{h}} \ + \ \sum_{i=1}^{2} \lambda_i \ \mathbf{e}_i \tag{5.2}$$

The base vector ($\bar{\mathbf{h}}$) can be taken as any of the 3 PGHs making up the reference triangle. The other 2 PGHs need to be extended in length so that they point to a plane lying tangentially to the end of the base vector. The first reference vector ($\mathbf{e_1}$) can then be sampled as a subtraction of one of these extended vectors with the base vector, the second ($\mathbf{e_2}$) being inferred as lying perpendicular to the first on the tangent plane. I.e. (for 3 local reference PGHs ($\mathbf{f}_1$, $\mathbf{f}_2$ & $\mathbf{f}_3$))

$$\bar{\mathbf{h}} \ = \ \mathbf{f}_1 \tag{5.3}$$

$$\mathbf{f}'_2 \ = \ (\mathbf{f}_1.\mathbf{f}_2) \ \mathbf{f}_2 \tag{5.4}$$

$$\mathbf{f}'_3 \ = \ (\mathbf{f}_1.\mathbf{f}_3) \ \mathbf{f}_3 \tag{5.5}$$

$$\mathbf{e}_1 \ = \ \frac{\mathbf{f}'_3 \ - \ \mathbf{f}_1}{|\mathbf{f}'_3 \ - \ \mathbf{f}_1|} \tag{5.6}$$

$$\mathbf{e}_x \ = \ \frac{\mathbf{f}'_2 \ - \ \mathbf{f}_1}{|\mathbf{f}'_2 \ - \ \mathbf{f}_1|} \tag{5.7}$$

$$\mathbf{e}_2 \ = \ \frac{\frac{\mathbf{e}_x}{\mathbf{e}_1.\mathbf{e}_x} \ - \ \mathbf{e}_1}{|\frac{\mathbf{e}_x}{\mathbf{e}_1.\mathbf{e}_x} \ - \ \mathbf{e}_1|} \tag{5.8}$$

Having established an orthogonal representation space for PGH modelling, as defined above, a sample PGH ($\mathbf{g}$) can be projected into this space, i.e. so that

$$\lambda_0 \;=\; \mathbf{\bar{h}.g} \;,\;\; \lambda_\mathbf{1} \;=\; \mathbf{e_1.g} \;,\;\; \lambda_\mathbf{2} \;=\; \mathbf{e_2.g} \tag{5.9}$$

with the sample PGH's reconstruction, or closest approximation ($\mathbf{h'}$) being given by

$$\mathbf{h'} \;=\; \lambda_0 \mathbf{\bar{h}} \;+\; \sum_{i=1}^{2} \lambda_i \; \mathbf{e}_i \tag{5.10}$$

This PGH ($\mathbf{h'}$), reconstructed in terms of linear weightings of the 3 orthogonal reference vectors, can then be normalised and matched directly to the original sample PGH ($\mathbf{g}$) with the Bhattacharyya metric to indicate the best match score (B) across the triangular reference region. I.e.

$$B \;=\; \frac{\mathbf{h'}}{|\mathbf{h'}|} \cdot \mathbf{g} \tag{5.11}$$

where $\mathbf{g}$ is (square rooted and) normalised to unit length from the outset. The corresponding viewpoint follows as an interpolation across the triangular view region relative to the weights λ_1 and λ_2.

By ensuring that all points within any such triangular reference region are supported to a predefined level of reconstruction error, we are able to model a continuous view-sphere-based shape manifold by triangulation to a moderated degree of precision. As discussed in more detail in Chapter 6, continuous triangulated mappings can be formed by iteratively growing connected triangular regions, or more straightforwardly by iteratively sub-sampling an approximately uniform spherical mapping such as an icosahedron.

To ensure that a sampled triangular patch maintains a consistent level of approximation error (i.e. distortion) across its bounds, sample points can be checked as the triangular region is plotted. Analysis of triangular plots of reconstruction error such as those in Figure 5.9, show that, typically, the worst points are mid-way between specific vertices or in the middle of the smoothly varying manifold. By ensuring that the reconstruction error (Bhattacharyya score) sampled at the centre (mean viewpoint) and mid-outer edges of each triangular reference region is above threshold, we can be assured that all points within the manifold are likely to be accurate enough to support continuous recognition across the triangular region. If the parameters of a sample point lie outside of the specified triangular region, the best match along the closest edge can be taken as a 1D interpolation between the 2 corresponding vertices. In these cases, recognition should be better supported by a neighbouring reference triangle - if one exists.

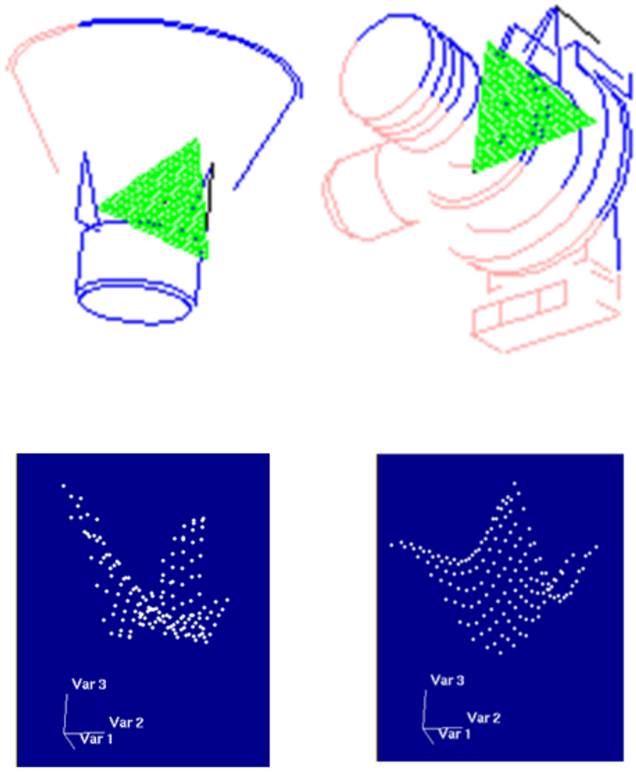


Figure 5.9: *The 2 data plots above represent the inverse Bhattacharyya linear reconstruction error (in the 3rd dimension (Var 3)) for PGHs sampled across the green shaded triangular partitions of view-space for the 2 pictured wireframe objects' respective blue shaded features (sampled for a single line feature). Each data point represents the Bhattacharyya score between the PGH sampled at that point on the view-sphere and its triangulated reconstruction between the outer 3 corners of the reference region. It can be seen that the accuracy of any triangulated PGHs tends to be worse at the centre of the manifold or mid-way along each outer edge. These 4 critical points can therefore be sampled for reconstruction to ensure that a given triangular reference region's continuous manifold maintains a specified level of precision. All the data points in the 3rd dimension of the presented distributions measure over 0.95 in terms of a Bhattacharyya match score.*

The diagrams in Figures 5.10 and 5.11 indicate the utility of the proposed linear interpolation scheme for continuously encoding objects' edge-based appearances around the view sphere with a reduced number of reference nodes relative to a more primitive nearest neighbour mapping. For these experiments, 8 essentially random object views and feature-based PGHs have been selected and triangular reference regions have been grown around the initial viewpoint to maintain a minimum specified reconstruction error across the triangular view patch (relative to the fixed-scale geometry sampled at 4 critical view points (i.e. at the centre and mid-edges of the triangular region (see Figure 5.9)). The process is performed using the proposed linear interpolation scheme (see Section 5.5) and also for a simpler nearest neighbour matching scheme where the match score is sampled as the best match of the sample view with each of the (3) reference views. The experiments are repeated for 3 different accuracy levels (i.e. 0.9, 0.95 and 0.99 minimum Bhattacharyya match scores). The corresponding relative sizes of the supported reference triangles are qualitatively illustrated in Figures 5.10-11. Although the difference in scope of the 2 view-sphere modelling methods is obvious for the coffee pot object, the effects are less pronounced for the other objects. The effective utility of the two approaches to object recognition is examined in more detail for a larger data set in Section 5.8.

5.6 Scale Interpolation

Given the proposed linear system for triangulating local regions of the view-sphere, the associated issue of scale space analysis requires further consideration. The PGH representation is not directly scale invariant, but robustness to changes in projected scale can be gained by sampling PGHs at a range of desired scales and interpolating between them. In previous research regarding 2D object recognition, series of scaled PGHs were stored in memory, with any sample scene lines being best matched throughout each set via linear interpolation. For large object databases, however, this requires a massive amount of storage, when conversely, a single scale could be sampled for each object with any scene data instead being scaled to match. As will be discussed, this will however restrict us to use templates with a fixed-distance perspective distortion.

In a general sense, the term 'scale invariance' is really a misnomer in that a single representation of a shape's features could never be used to accurately sample an object's appearance through scale space extrema with a sensor of finite resolution. For instance, edge-based recognition would be impossible if using only a handful

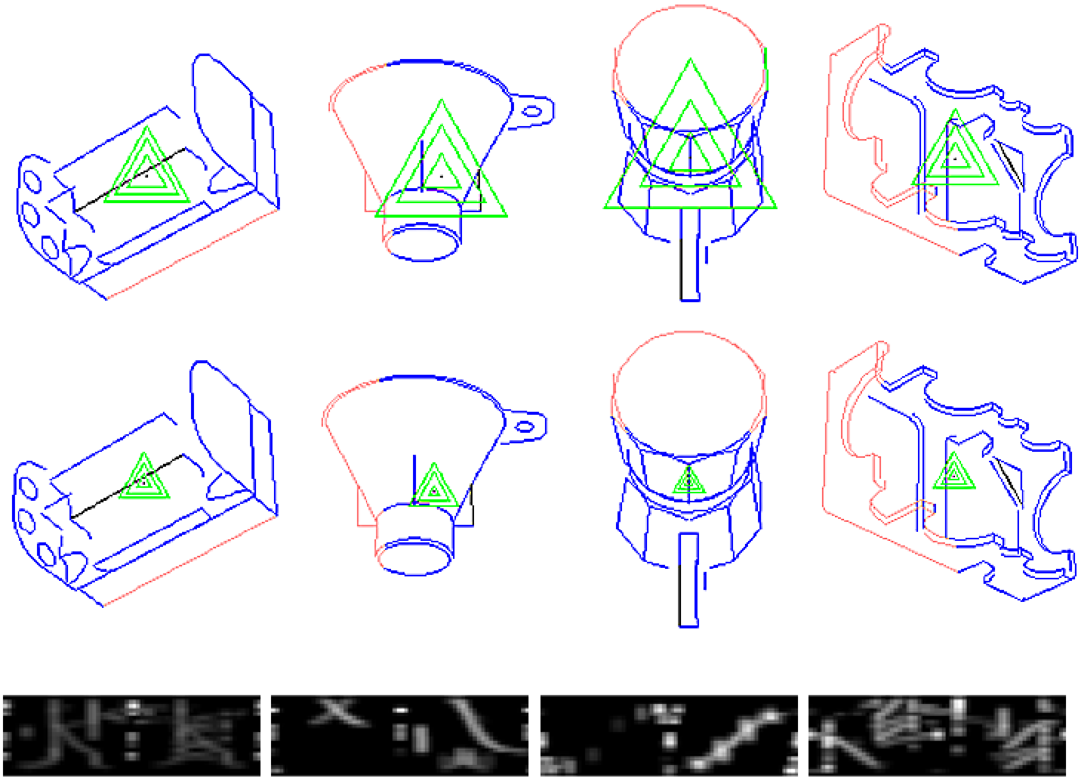


Figure 5.10: *PGHs have been sampled for individual features from each of the wireframe objects pictured above (bottom row), with features included in each PGH shaded in blue. The three triangles displayed over each object in the top row represent regions of the view-sphere for which all enclosed view-points' corresponding PGHs can be triangulated to 0.99, 0.95 and 0.9 degrees of precision (outwards respectively) in terms of a Bhattacharyya match score according to the linear model of PGH deformation presented above. The corners of the largest triangles in the top row, from left to right, are approximately 24, 37, 34 and 21 degrees apart from the view-sphere origin. The bottom row represents the same information without use of interpolation across the triangular reference region. Instead, the triangles extend to the point at which the minimum raw match score sampled at the corner of each triangle, relative to the triangle centre (mean viewpoint), drops below the threshold. The diagrams indicate the utility of using PGH interpolation for populating objects' view-spheres.*

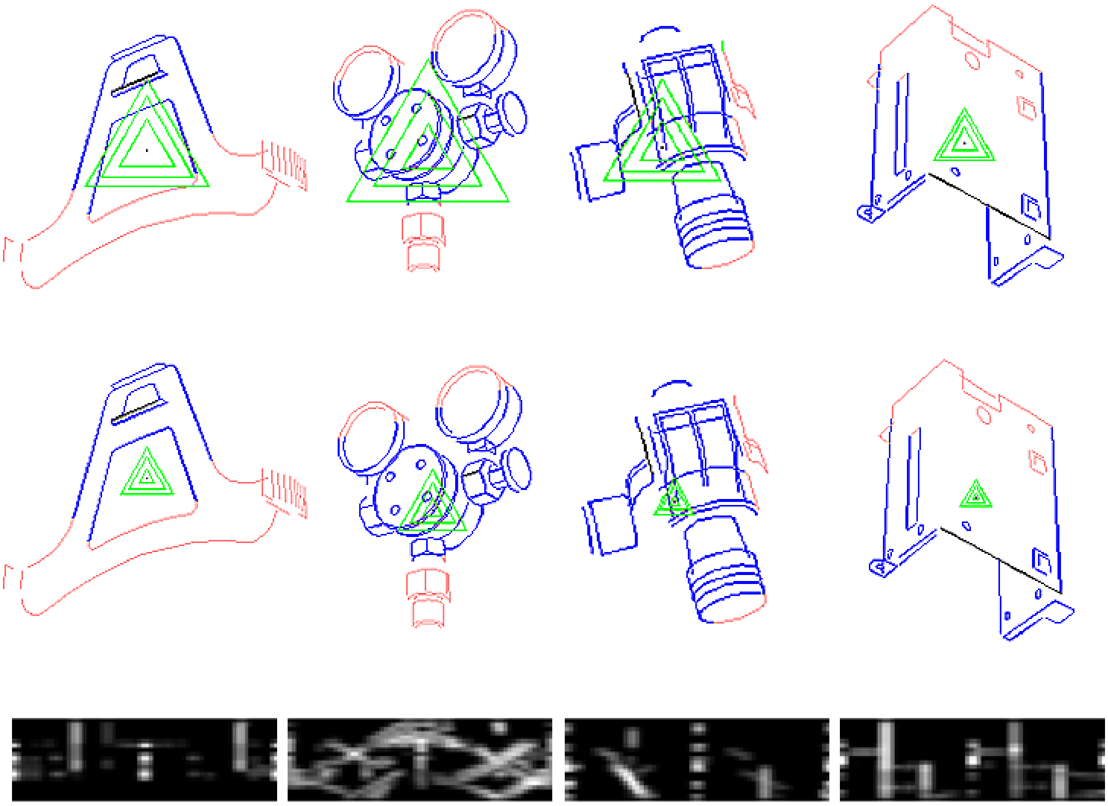


Figure 5.11: *The images above accord with those presented in Figure 5.10 and are those making up the set of objects sampled throughout the following experiments.*

of pixels. Furthermore, as objects' appearances change in size, any projected edge representations may undergo substantial relative distortion, with the potential for edge features to blur together or disappear. The assumption made here is that objects' appearances will however be stable, according to a prototypical set of edge features that define the object's characteristic shape, through a reasonably wide range of typical viewpoints for which analysis is currently directed.

In the same way that a single PGH can be linearly interpolated through scale, so too can the previously outlined linear system for representing in-depth rotation. The proposed representation is thus broken down into 2 linear systems, cumulatively forming a global piecewise linear model approximating continuous PGH-bound appearance for objects' features' projected edge-based appearances through in-depth rotation and image scale.

To recap, the variation of a histogram h is modelled by linearised 2D variation according to

$$\mathbf{h} \ = \ \lambda_0 \bar{\mathbf{h}} \ + \ \sum_{i=1}^{2} \lambda_i \ \mathbf{e}_i$$

where $\bar{\mathbf{h}}$ is the base PGH and $\mathbf{e_1}$ and $\mathbf{e_2}$ are orthogonal reference vectors, weighted respectively by $\lambda_{(0,1and2)}$. The three orthogonal axes are defined relative to the PGHs observed at the corner coordinates of a sampled reference triangle.

The closest match to this model for a sample PGH $\mathbf{g}$ is

$$\mathbf{h}' \ = \ \bar{\mathbf{h}}.\ \mathbf{g}\ \bar{\mathbf{h}} \ + \ \sum_{i=1}^{2} (\ \mathbf{e}_i.\ \mathbf{g}) \ \ \mathbf{e}_i \tag{5.12}$$

normalised to the unit hyper-sphere.

In accordance with use of the Bhattacharyya metric for histogram comparison, the metric used to match sample PGHs to corresponding manifold approximations is of the form

$$B^2 \ = \ \frac{(\mathbf{h}.\mathbf{g})^2}{|\mathbf{h}|^2|\mathbf{g}|^2} \tag{5.13}$$

where the square is taken to avoid any square roots (associated with PGH vector lengths) in the following mathematics for convenience. (This will not affect the relative location of the peak response.)

The best match attainable with the linear model is therefore

$$\begin{aligned}\frac{(\mathbf{h}'.\mathbf{g})^2}{|\mathbf{h}'|^2|\mathbf{g}|^2} \ &= \ \frac{[\mathbf{g}.(\bar{\mathbf{h}}.\mathbf{g}\ \bar{\mathbf{h}}) \ + \ \mathbf{g}.(\sum_i^2(\mathbf{e}_i.\mathbf{g})\ \mathbf{e}_i)]^2}{|\mathbf{h}'|^2|\mathbf{g}|^2} \\ &= \ \frac{[(\mathbf{g}.\bar{\mathbf{h}})^2 \ + \ \sum_i^2(\mathbf{e}_i.\mathbf{g})^2]^2}{[(\mathbf{g}.\bar{\mathbf{h}})^2 \ + \ \sum_i^2(\mathbf{e}_i.\mathbf{g})^2]|\mathbf{g}|^2} \\ &= \ \frac{[(\mathbf{g}.\bar{\mathbf{h}})^2 \ + \ \sum_i^2(\mathbf{e}_i.\mathbf{g})^2]}{|\mathbf{g}|^2} \ = \ \frac{|\mathbf{h}'|^2}{|\mathbf{g}|^2}\end{aligned} \tag{5.14}$$

The possible variation of the observed data can then be represented with a second independent linear model.

$$\mathbf{g} \;=\; \bar{\mathbf{g}} \;+\; \sum_k \delta_k \mathbf{f}_k \tag{5.15}$$

Because the overall system is still in effect a linear model and the similarity metric is quadratic, a unique best match for this second model with the view model can be obtained by determining the δ_k that gives the best match score

$$\frac{|\mathbf{h}'|^2}{|\mathbf{g}|^2} \;=\; \frac{[(\bar{\mathbf{g}} \;+\; \sum_k \delta_k \mathbf{f}_k).\bar{\mathbf{h}}]^2 \;+\; \sum_i [(\bar{\mathbf{g}} \;+\; \sum_k \delta_k \mathbf{f}_k).\mathbf{e}_i]^2}{1 \;+\; \sum_j \delta_j^2} \tag{5.16}$$

Approximating the effects of a change in scale on the input image data as a 1D model (δ_k), we can write

$$\frac{|\mathbf{h}'|^2}{|\mathbf{g}|^2} \;=\; \frac{[(\bar{\mathbf{g}} \;+\; \delta\mathbf{f}).\bar{\mathbf{h}}]^2 \;+\; \sum_i^2 [(\bar{\mathbf{g}} \;+\; \delta\mathbf{f}).\mathbf{e}_i]^2}{1 \;+\; \sum_j \delta^2} \tag{5.17}$$

$$= \;\frac{(\bar{\mathbf{g}}.\bar{\mathbf{h}})^2 \;+\; 2\bar{\mathbf{g}}.\bar{\mathbf{h}}\delta\mathbf{f}.\bar{\mathbf{h}} \;+\; (\delta\mathbf{f}.\bar{\mathbf{h}})^2 \;+\; \sum_i^2 [(\bar{\mathbf{g}}.\mathbf{e}_i)^2 \;+\; 2\bar{\mathbf{g}}.\mathbf{e}_i\delta\mathbf{f}.\mathbf{e}_i \;+\; (\delta\mathbf{f}.\mathbf{e}_i)^2]}{1 \;+\; \delta^2}$$

The value of d that maximises this expression can then be obtained when

$$\partial B^2/\partial\delta \;=\; \frac{[1 \;+\; \delta^2][2\bar{\mathbf{g}}.\bar{\mathbf{h}}\mathbf{f}.\bar{\mathbf{h}} \;+\; 2\mathbf{f}.\bar{\mathbf{h}}\delta\mathbf{f}.\bar{\mathbf{h}} \;+\; \sum_i^2 (2\bar{\mathbf{g}}.\mathbf{e}_i\mathbf{f}.\mathbf{e}_i \;+\; 2\mathbf{f}.\mathbf{e}_i\delta\mathbf{f}.\mathbf{e}_i)]}{[1 \;+\; \delta^2]^2} \tag{5.18}$$

$$-\;\frac{2\delta[(\bar{\mathbf{g}}.\bar{\mathbf{h}})^2 \;+\; 2\bar{\mathbf{g}}.\bar{\mathbf{h}}\delta\mathbf{f}.\bar{\mathbf{h}} \;+\; (\delta\mathbf{f}.\bar{\mathbf{h}})^2 \;+\; \sum_i^2 [(\bar{\mathbf{g}}.\mathbf{e}_i)^2 \;+\; 2\bar{\mathbf{g}}.\mathbf{e}_i\delta\mathbf{f}.\mathbf{e}_i \;+\; (\delta\mathbf{f}.\mathbf{e}_i)^2]]}{[1 \;+\; \delta^2]^2}$$

$$= \;0$$

which can be re-written as a quadratic equation in terms of d

$$[\bar{\mathbf{g}}.\bar{\mathbf{h}}\mathbf{f}.\bar{\mathbf{h}} \;+\; \sum_i^2 \bar{\mathbf{g}}.\mathbf{e}_i\mathbf{f}.\mathbf{e}_i] \;+\; \delta[(\mathbf{f}.\bar{\mathbf{h}})^2 \;+\; \sum_i^2 (\mathbf{f}.\mathbf{e}_i)^2 \;-\; (\bar{\mathbf{g}}.\bar{\mathbf{h}})^2 \;-\; \sum_i^2 (\bar{\mathbf{g}}.\mathbf{e}_i)^2] \tag{5.19}$$

$$- \ \delta^2 [\sum_i^2 \bar{\mathbf{g}}.\mathbf{e}_i \mathbf{f}.\mathbf{e}_i \ + \ \bar{\mathbf{g}}.\bar{\mathbf{h}}\mathbf{f}.\bar{\mathbf{h}}] \ + \ \delta^3[0] \ = \ 0$$

i.e.

$$A \ + \ B\delta \ - \ A\delta^2 \ = \ 0 \tag{5.20}$$

with

$$A \ = \ \bar{\mathbf{g}}.\bar{\mathbf{h}}\mathbf{f}.\bar{\mathbf{h}} + \sum_i^2 \bar{\mathbf{g}}.\mathbf{e}_i\mathbf{f}.\mathbf{e}_i \ , \quad B \ = \ (\mathbf{f}.\bar{\mathbf{h}})^2 + \sum_i^2 (\mathbf{f}.\mathbf{e}_i)^2 - (\bar{\mathbf{g}}.\bar{\mathbf{h}})^2 - \sum_i^2 (\bar{\mathbf{g}}.\mathbf{e}_i)^2 \tag{5.21}$$

so that

$$\delta \ = \ \frac{B \ \pm \sqrt{B^2 \ + \ 4A^2}}{2A} \tag{5.22}$$

The 2 possible solutions to this quadratic equation (δ_1, δ_1) can be fed back into the linear scale model, so that the reconstructed PGHs can be directly matched against the reference triangle's manifold, returning the corresponding Bhattacharyya match score (B). The better of the two possible solutions can be taken as the final match and the other can be disregarded. I.e.

$$\mathbf{g}_1 \ = \ \bar{\mathbf{g}} \ + \ \delta_1 \mathbf{f} \ , \quad \mathbf{g}_2 \ = \ \bar{\mathbf{g}} \ + \ \delta_2 \mathbf{f} \tag{5.23}$$

$$\mathbf{h}'_1 \ = \ \bar{\mathbf{h}}.\mathbf{g}_1 \ \bar{\mathbf{h}} \ + \ \sum_{i=1}^2 (\mathbf{e}_i.\mathbf{g}_1) \ \mathbf{e}_i \ , \quad \mathbf{h}'_2 \ = \ \bar{\mathbf{h}}.\mathbf{g}_2 \ \bar{\mathbf{h}} \ + \ \sum_{i=1}^2 (\mathbf{e}_i.\mathbf{g}_2) \ \mathbf{e}_i \tag{5.24}$$

$$B_1 \ = \ \frac{\mathbf{h}'_1.\mathbf{g}_1}{|\mathbf{h}'_1||\mathbf{g}_1|} \ , \quad B_2 \ = \ \frac{\mathbf{h}'_2.\mathbf{g}_2}{|\mathbf{h}'_2||\mathbf{g}_2|} \tag{5.25}$$

$$B \ = \ \max\{B_1, B_2\} \tag{5.26}$$

The best matching viewpoint is accordingly inferred as a corresponding triangulated point on the view-sphere (i.e. with reference to the orthogonal weightings ($\mathbf{e}_1.\mathbf{g}$) and ($\mathbf{e}_2.\mathbf{g}$)). As discussed, if the best matching view-point is inferred outside of the reference triangle, the best match is inferred as a 1D interpolation along the adjacent triangle edge. In these cases, the best match should be realised by an adjacent triangular region (if one exists).

5.6.1 Experimental Evaluation

In order to verify the proposed methodology, a selection of views of wireframe objects have been sampled and a triangular region has been grown about the principal viewpoint, maintaining a specified minimum reconstruction error at the centre and mid-sides of the triangle (see the top row of Figures 5.10 and 5.11). Maintaining a central viewpoint, the objects have then been scaled about their original sizes and the corresponding scale parameters and reconstructed match scores have been sampled.

With reference to Figure 5.12, by setting the pairwise respective scale of the test objects to certain ratios of the mean size, e.g. for a 40% scale change, the pairs of scales can be sampled as 1.0 to 1.4, 0.9 to 1.3, 0.8 to 1.2, 0.7 to 1.1 and 0.6 to 1.0 of the original size, representing the intermediate points 0.0, 0.25, 0.5, 0.75 and 1.0 along the 1D scale-based PGH-vector manifold. The scale value recognised as the best match across the triangular PGH reference patch using the mathematics outlined in this chapter can accordingly be plotted as the orthogonal axis. If the theory outlined in this chapter is valid, the best matching scale value returned by the proposed system will correspond linearly to the scale ratio of the sample, evidenced by a diagonal straight line from the bottom left to the top right corner of the graph. The experiments are run with various parameterisations of scale and accuracy so that the bounds of the linear modelling strategy can be identified as the basis of the proposed generic 3D edge-projected shape recognition framework. Corresponding 'Bhattacharyya Reconstruction Accuracy' graphs (Figure 5.13) detail the best scaled match scores for the sampled objects' features across the triangular reference regions. Because scale is changed in both directions in the given series to maintain the specified relative scale interval, the experiments highlight the stability of the methodology over double the specified interval as indicated in Figure 5.12. The experiments are repeated through scale changes spanning 0.3, 0.2 and 0.1 of the original size in each direction.

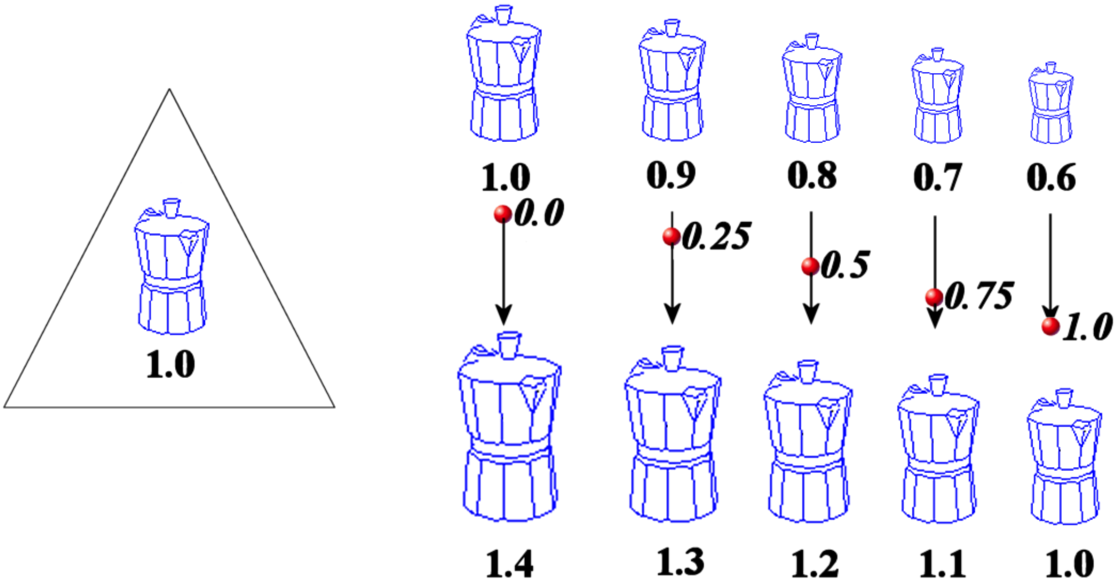
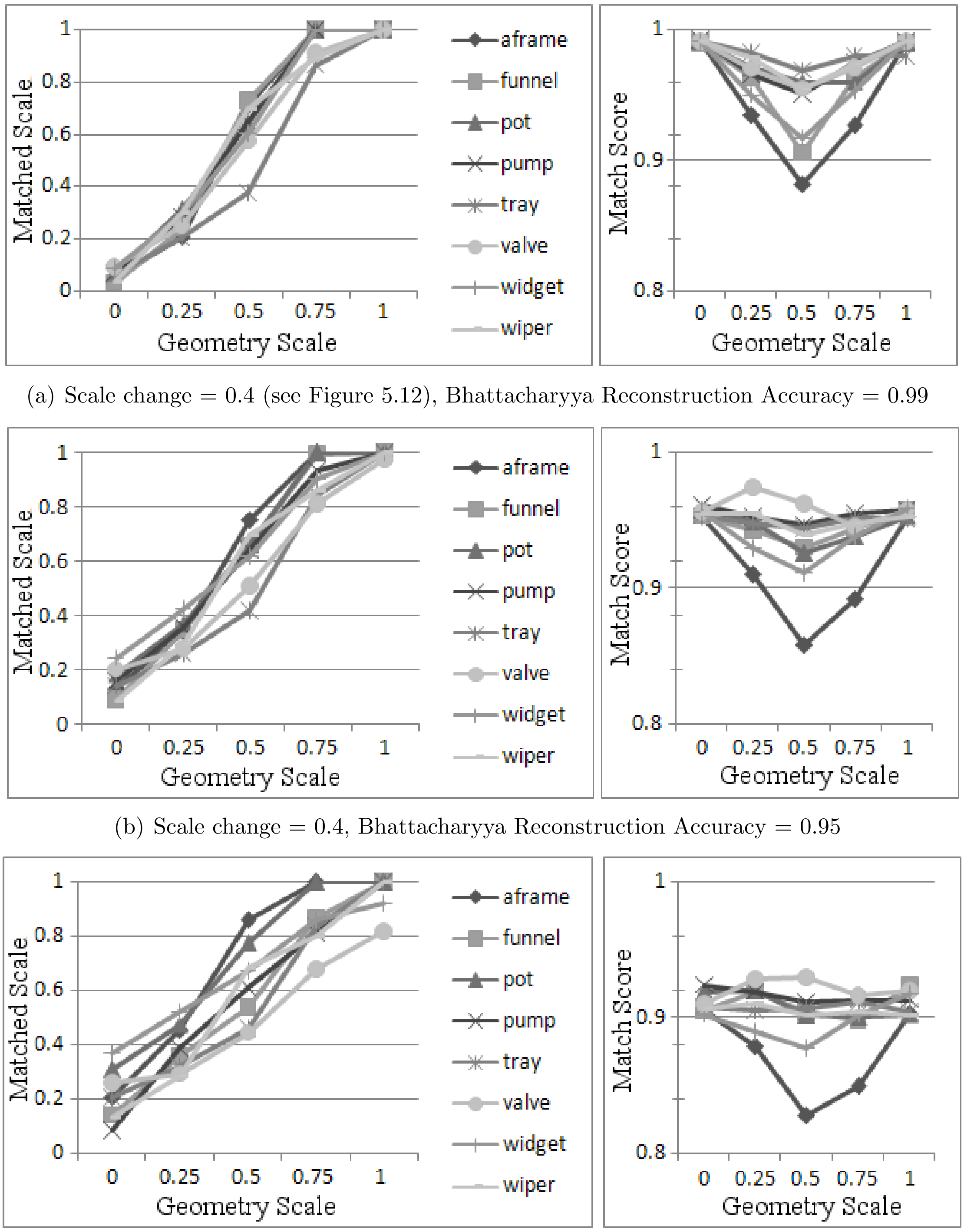


Figure 5.12: *The diagram above indicates the basis of the experiments supporting the graphs presented in Figures 5.13-16. A unit scaled triangular reference patch (supported by 3 PGHs) modelled at a certain precision is matched to pairs of PGHs representing the same geometry at a range of encompassing scale intervals. The best matching scale interpolated between the scaled reference pair (**Matched Scale**) is plotted against the corresponding scale of the sampled edge geometry (**Geometry Scale**) to allow the linearity and thus validity of the mechanism to be observed. In the example above, the unit object is scaled by 40% in each direction, with the valid corresponding scale being sampled at the points **0.0**, **0.25**, **0.5**, **0.75** and **1.0** along the single vector manifold modelled between them. The experiments are repeated with 10, 20 and 30% scale changes (in each direction) and also for 3 precision levels (0.9, 0.95, 0.99) modelled as minimum Bhattacharyya reconstruction errors between the triangular bounds of the 3 reference PGHs. The accuracy of the representation is assessed by virtue of consistency of the reconstruction error sampled as a Bhattacharyya match score (**Match Score**) between the linearly reconstructed (triangulated) PGH and the corresponding view-sphere sampled PGH. These match scores are plotted as adjacent corresponding connected graphs.*

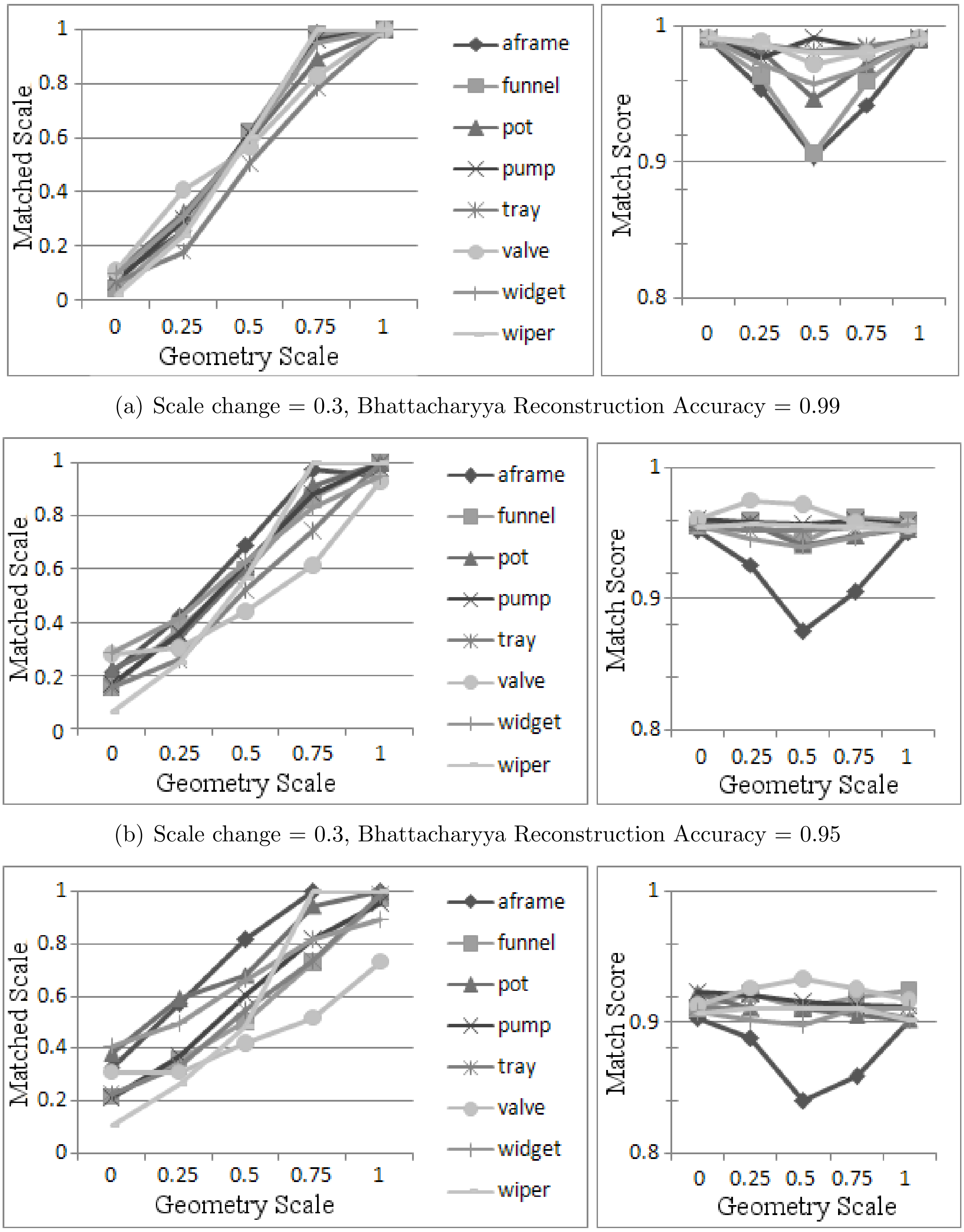


(a) Scale change = 0.4 (see Figure 5.12), Bhattacharyya Reconstruction Accuracy = 0.99

(b) Scale change = 0.4, Bhattacharyya Reconstruction Accuracy = 0.95

(c) Scale change = 0.4, Bhattacharyya Reconstruction Accuracy = 0.9

Figure 5.13: *With reference to Figure 5.12, the graphs presented across the next 4 pages serve to validate the proposed metric for PGH-based scale space recognition of objects' projected edge features. Each 'Scale change' parameter represents the proportion of the original object's size through which recognition is sampled in each direction across the triangular manifold formed at the indicated level of precision. If the proposed representation is valid at the sampled parameters, the graph to the left hand side should pass linearly from the bottom left to the top right. The maintained accuracy of the interpolated scale samples should also remain constant at the prescribed level as indicated in the graphs on the right hand side.*

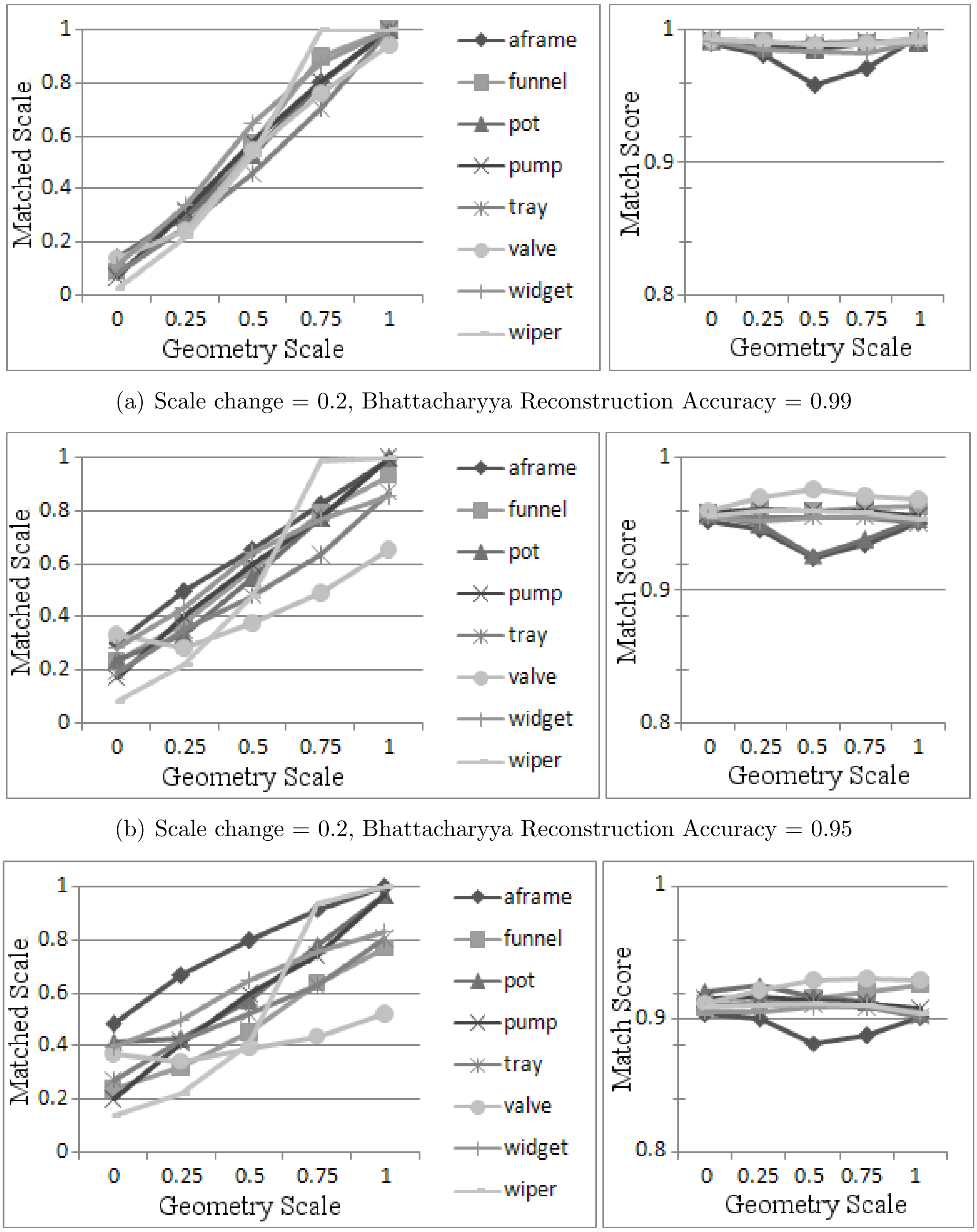


(a) Scale change = 0.3, Bhattacharyya Reconstruction Accuracy = 0.99

(b) Scale change = 0.3, Bhattacharyya Reconstruction Accuracy = 0.95

(c) Scale change = 0.3, Bhattacharyya Reconstruction Accuracy = 0.9

Figure 5.14: *Figure 5.6.1 continued.*

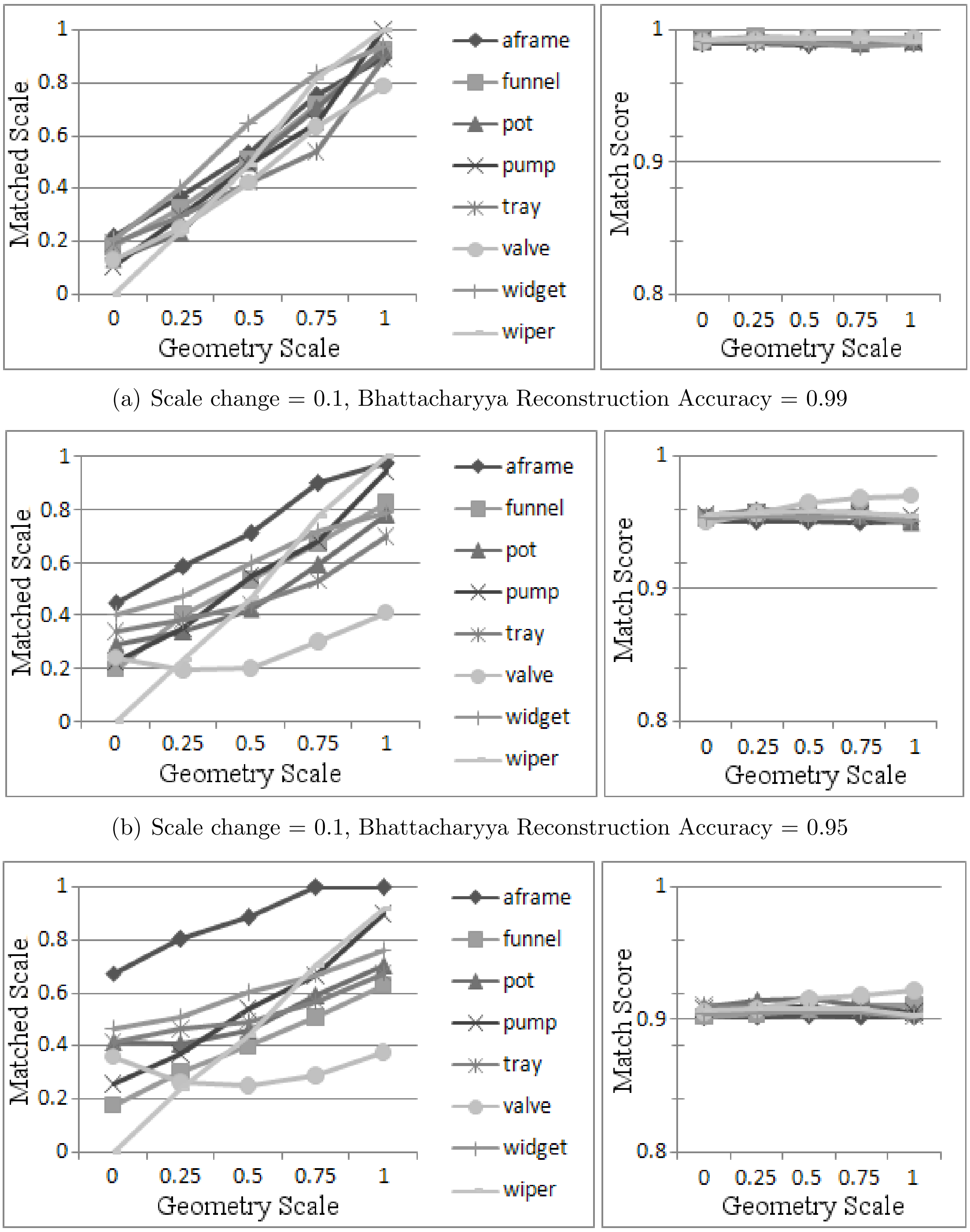


(a) Scale change = 0.2, Bhattacharyya Reconstruction Accuracy = 0.99

(b) Scale change = 0.2, Bhattacharyya Reconstruction Accuracy = 0.95

(c) Scale change = 0.2, Bhattacharyya Reconstruction Accuracy = 0.9

Figure 5.15: *Figure 5.6.1 continued.*



(a) Scale change = 0.1, Bhattacharyya Reconstruction Accuracy = 0.99

(b) Scale change = 0.1, Bhattacharyya Reconstruction Accuracy = 0.95

(c) Scale change = 0.1, Bhattacharyya Reconstruction Accuracy = 0.9

Figure 5.16: *Figure 5.6.1 continued.*

The graphs presented in Figures 5.13-16 indicate that the proposed representation for recognising 3D objects through scaled regions of the view-sphere is indeed valid, with scale generally increasing monotonically as required. The accuracy required and permissible size of the scale change is however shown to vary across the objects examined. The graphs indicate that the accuracy of the representation is most stable at the smallest sampled scale change, i.e 10%. Some objects however appear much less affected by scale and may therefore be sampled across wider regions, potentially supporting optimised learning. The majority of the sampled objects, for instance, appear relatively unaffected by doubling the scale change to 0.2 in each direction when precision is maintained at 0.99. The precision required also varies across objects, meaning that less precision may be used to recognise certain objects. Although desirable to extend the scope of each PGH-sampled view-sphere region as far as possible, a certain level of accuracy will however be required to ensure that recognition results are not confused and so that localisation is accurate enough to support the final parameter optimisation alignment procedure.

One of the most noticeable results derived from the data distributions is that for the valve object. For less accurate examples (e.g. 0.9), the valve's precision can be seen to increase over the sampled minimum, suggesting that a better match was found across the distorted structure tolerated within the linear model, i.e. a false-positive. In these cases, the linearity of the scale assumption is seen to be invalidated. This is presumably because of the complicated repeated structure exhibited by the object, meaning that line entries are more likely to be blurred together and aligned with each other in the noise field of the distorted linear space. The valve is only good at the most accurate sampled level of precision, although interestingly, it is almost as good with twice the minimum scale change, which could effectively be used as compensation. Furthermore, the wiper object is shown to be very unreliable at broader scales, but almost perfectly behaved across all sampled accuracy levels for the smallest scale change.

The data suggests that a scale change of 10 to 20% is generally tolerable by the proposed system, according with the 15% empirically sampled across a broader object test set in associated practical work. Because the tolerable degree of scale is relatively constrained (on account of the non-linearity of the 1D scale manifolds), for an object to be recognised across wide fields of vision (with fixed perspective), the presented operation would require sampling in intervals through scale space. Although the data plots are not perfectly linear and consistent, it must be borne in mind that we are dealing with an approximation task, restricted by the scope of linear methods.

Potentially, where there is advantage to be gained in using less accurate representations to save on storage and processing costs, any instances where the scale is shown to be linear but displaced out of synch could be appended with scale factors that would force alignment from 0 to 1. The subsequent alignment optimisation procedure is otherwise expected to compensate for some degree of imprecision in the initial prediction of model scale and pose. Because the examples given are based around 2 raw scaled PGHs, there is also the possibility that features may come into or go out of view across the 2 samples, which may interfere with the process, especially for wider scale samples where there is more opportunity for this to happen.

In conclusion, the presented graphs show that the proposed representation needs to be set at the most precise level with the smallest sampled scale change (10%) in order to reliably cater for all the object samples. However, some objects appear far more amenable to processing at wider scale separations with less precision. It seems that an adaptive learning strategy should therefore be implemented to learn objects' features' representations at the precision required to enable as wide scale sampling as possible. Processing would otherwise be effectively wasted. It is however desirable to impose a minimum accuracy level across the system, so that relative match scores can be meaningfully assessed for recognition and so that any subsequent object localisation processes remain well informed. A feedback loop could therefore be built into the learning process, so that the optimal precision and scale parameters were autonomously set by the system for each sampled reference region over a prescribed minimum level of precision.

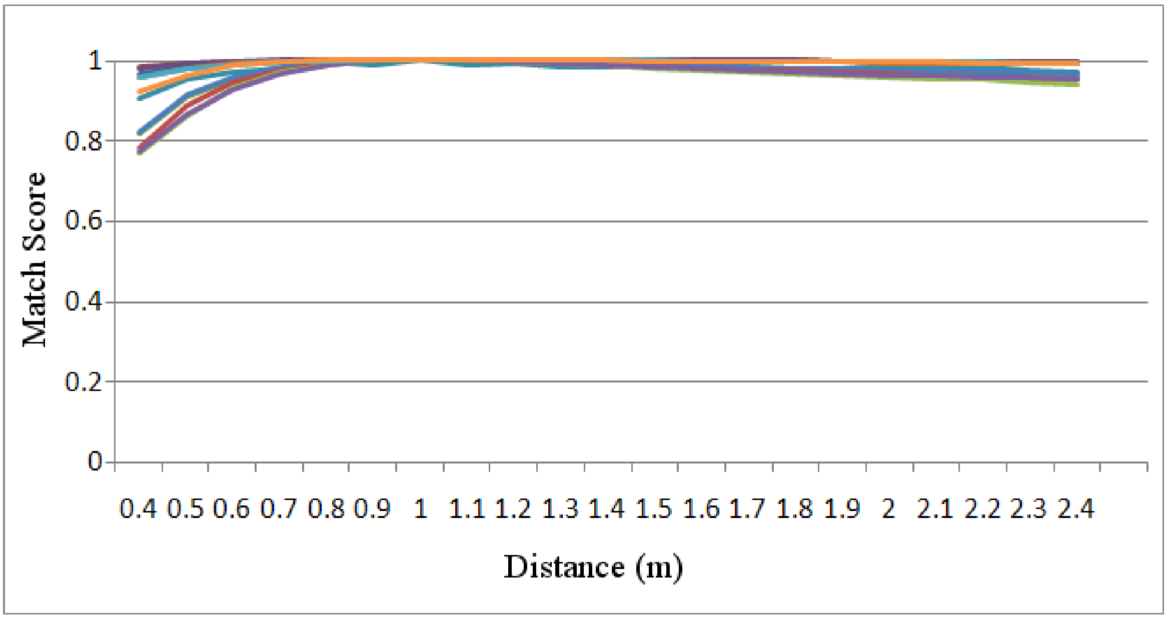
5.7 Perspective

For the scale space experiments in the previous section and all other figures presented so far, the wireframe models have been sampled with perspective distortion approximating that which would be observed if the vision system was presented with equivalent tangible objects at arms' reach, as would be required for traditional interactive learning. This is approximated with a pin-hole camera projection model by setting the virtual camera origin to a certain distance along the camera's z axis from the objects' (centred) view-sphere origin. In relation to the experimental camera rig positioned around 1.0 m from the object test-bed and the camera's fixed 1.0 m focal distance, perspective distortion has been set to that observed 1.0 m away from the view-sphere origin.

As an object moves away from the camera, any effects of perspective distortion will diminish, with projected geometry tending towards an orthogonal projection model at infinity (Figure 5.18). Conversely, as an object moves closer to the camera, perspective effects will become more pronounced, especially for larger objects. If a computer vision recognition system needs to be capable of reliably recognising objects irrespective of relative distance then further analysis is required to model any such distortion effects. Figure 5.17 illustrates Bhattacharyya match scores between a selection of objects' representative views' features' PGHs progressively distorted with perspective projection (away from the initial sample at 1.0 m perspective distance) relative to distance from the 3D object centroid under a pin-hole camera projection model. Figure 5.18 similarly shows how perspective distortion tends to an orthogonal projection at distance.

Figure 5.17 shows that a fixed perspective scale of 1.0 m is sufficient to support recognition across a reasonably wide range of object features and perspective scales with reasonable precision. Although the graphs show that some objects are more affected by perspective than others, because many prototypical object features relate to more planar geometry (e.g. on the side of an object), a cross section of deep-3D structural features was sampled from a range of objects' representative wireframes. Typically, a random sample set would coincide more with the upper, more stable features in the graph that are less affected by such pronounced perspective projection effects. Some objects' views' features are however liable to be critically affected by the effects of perspective distortion, especially at close quarters (Figure 5.19). Further modelling of perspective is therefore required in order to be able to assuredly recognise any such object part irrespective of distance from the viewer (within sensible ranges).

The effects on the PGH representation from 3D edge projected perspective distortion are similar to those induced though localised scale and view-point variation and can therefore be similarly locally linearly modelled. Because we sample scale between 2 points, for fixed size tangible objects, the perspective could potentially be built into the scale model. Each scaled sample could be presented under the perspective effects observable for that object at that distance from the camera. One problem here is that the proposed projected shape recognition system is to use a single scale model so that



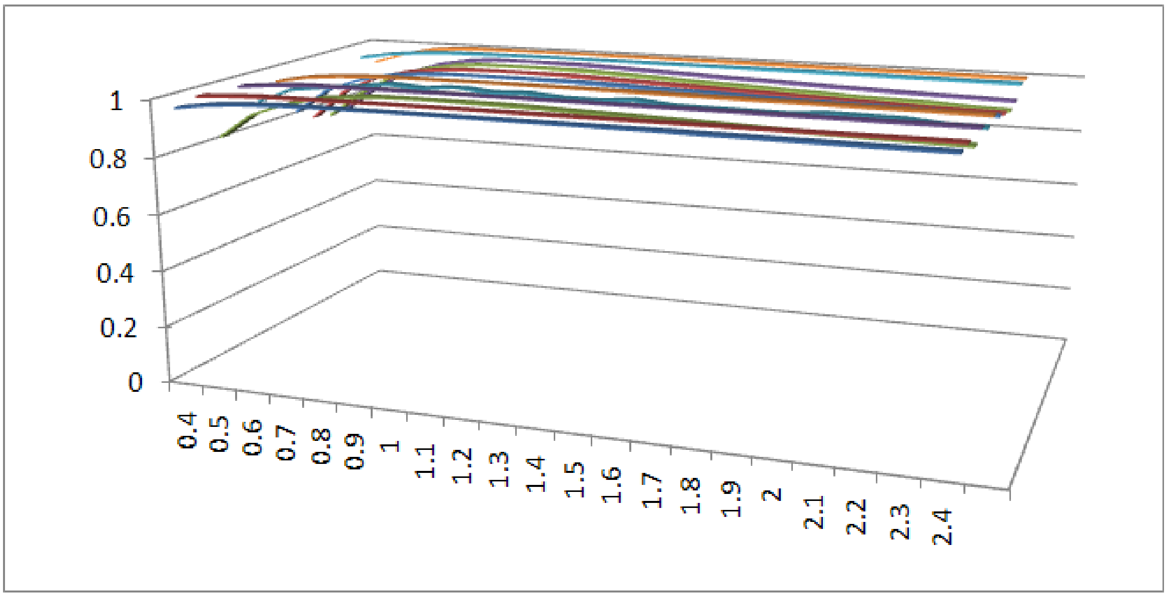
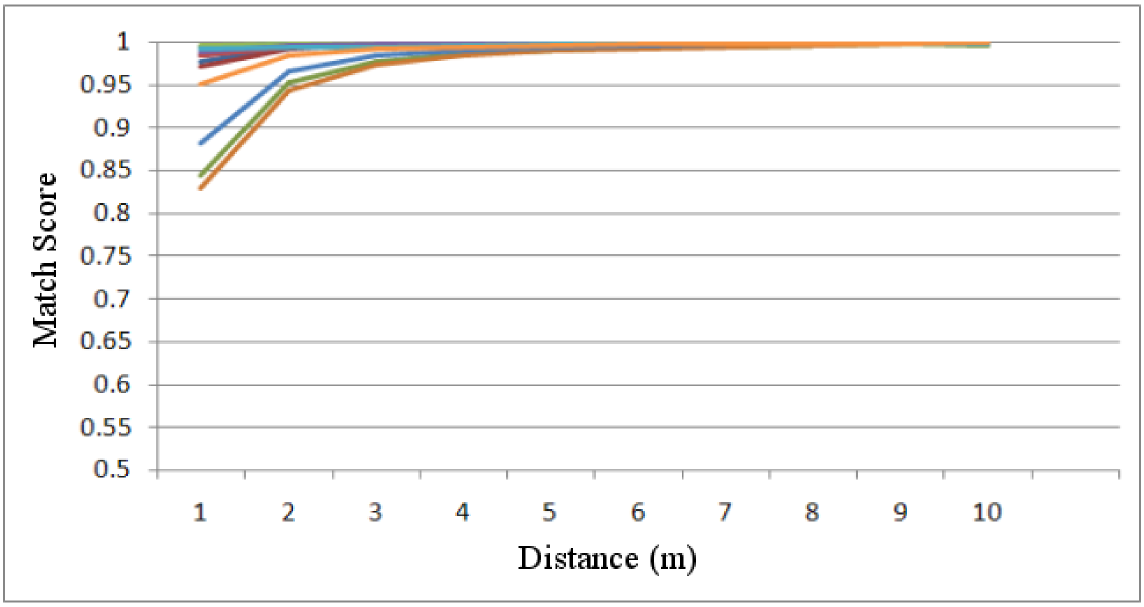


Figure 5.17: *The charts above detail the relative Bhattacharyya match scores (0 to 1) for PGHs sampled from a range of 3D objects' representative features with perspective effects set at distances ranging from 0.4 to 2.4 metres as compared to the same PGHs sampled 1.0 m away (as learned). The data is shifted into 3D in the lower chart, to descriptively separate the cluster of features' responses at the upper levels.*



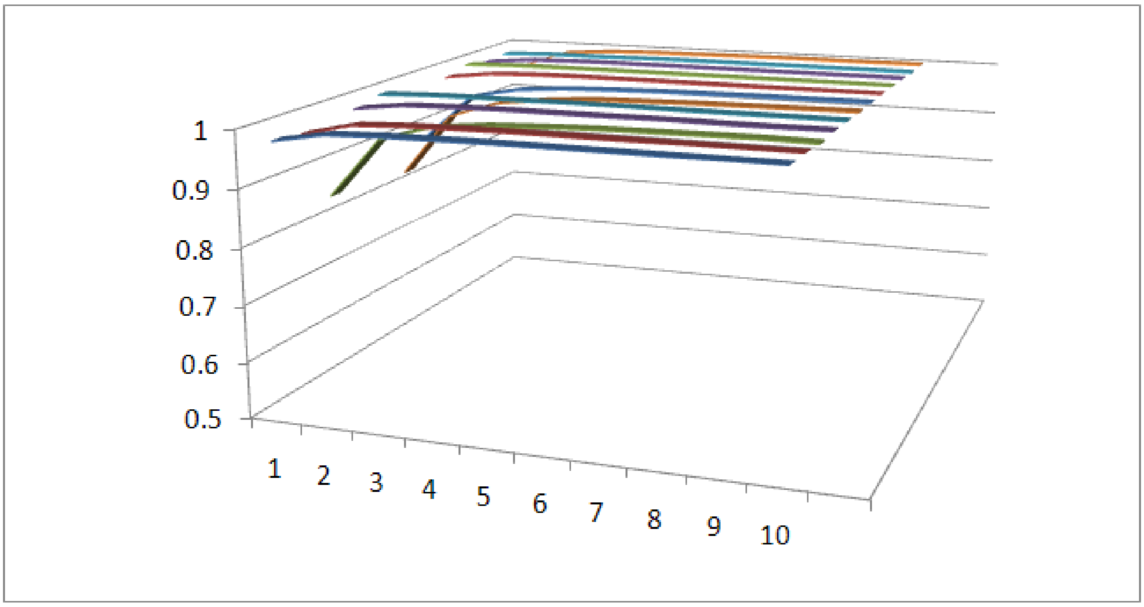


Figure 5.18: *Further to the data presented in Figure 5.17, scaled Bhattacharyya match scores (0.5 to 1.0) are presented for a range of objects with perspective effects ranging through those observed from 1 to 10 m, as compared to the same geometry under orthogonal projection. The data can be seen to converge to the orthogonal representation at distance.*

the image data can instead be scaled to match without having to store all the scaled sample representations. Because the image data is 2D and unknown to begin with, it is impossible to predict how different image parts may distort as their unknown 3D parent features are viewed at their respective distances. To parameterise perspective, as an acute vision system should be able, supplementary learned perspective models would thus be required, although only for those features that needed it. This might involve, for instance, 10% of the object feature samples having 1 or 2 supplemental perspective view-sphere representations, so that the effects of perspective could be linearly interpolated between them, as required. Objects may otherwise exist and need to be recognised at different physical scales (e.g. a car and a toy car), so that perspective effects need to be potentially learned for a range of object scales relative to the camera. This suggests that an independent model of perspective projection is required for generic shape recognition. Because of the high processing costs associated with implementing a single perspective scale model of 3D projected shape recognition, current research is focused on the more stable local regions of scale space, e.g. a scale range of 0.7 to 2.0 metres in Figure 5.17 where all features maintain a high match score relative to the learned scale at 1.0 metre.

There may also be a subtle difference in the shape of an object under perspective depending on which point on the object the viewer is looking directly at relative to scale. Each model is learned with the object's centroid centred on the z axis. A series of tests were conducted by changing the view-point to pass through various extremal projected edge features and examining any effects on the match score. For the objects examined in this chapter, such effects are negligible, with a depletion of around 1% in the worst case. As indicated in [98], a prospective vision system could employ a spherical sensor (equivalent to the human retina) to remove any distorted projected effects observed across the sensor away from the centre of view for wider fields of vision.

5.8 Complexity Analysis: Nearest Neighbour versus Triangulated PGH View-Sphere Mapping

To summarise the advantage to be gained in using triangulation instead of nearest neighbour matching, an estimate of the gain in the area of the view-sphere covered by a single PGH can be sampled in accordance with the experiments supporting

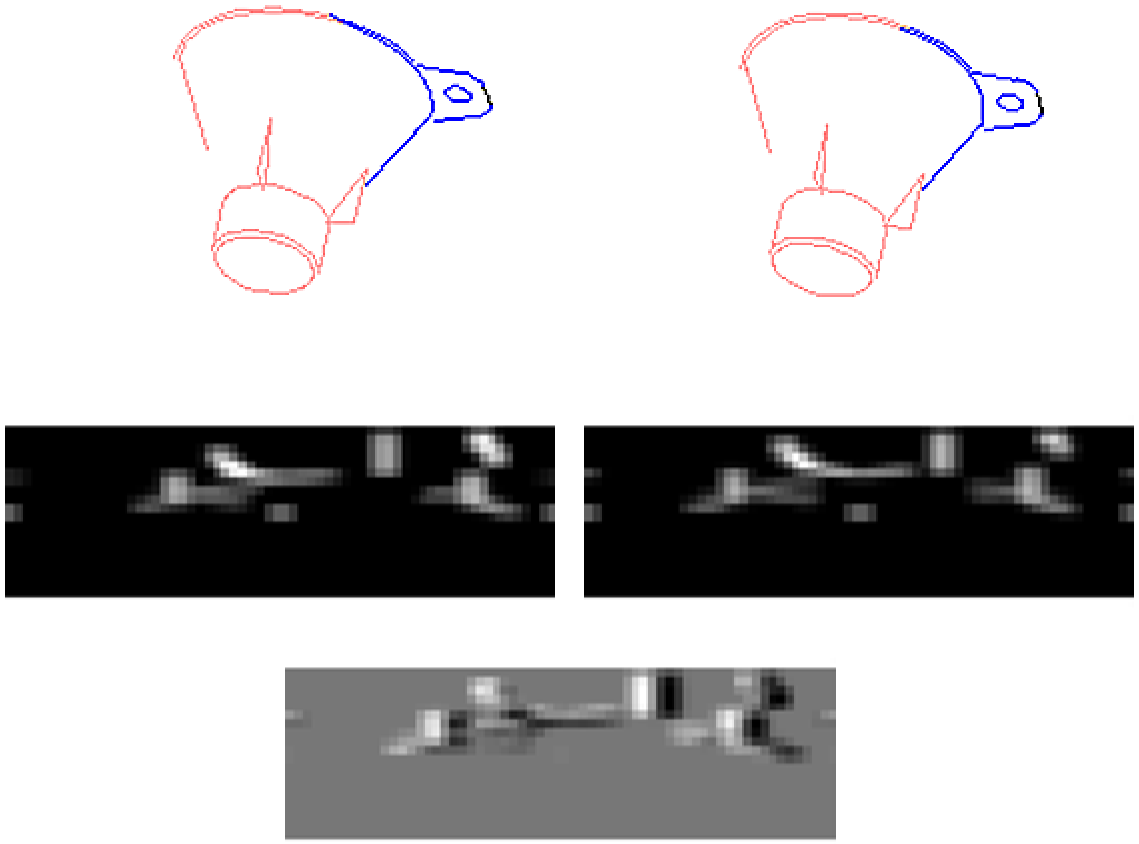


Figure 5.19: *The worst observed match score in the presented perspective experiments related to the funnel object imaged above at approximately the same pose with the rightmost line feature selected and the features in blue included in the sampled PGH. The images to the left relate to an orthogonal projection of the presented object and the ones to the right to a perspective projection with the virtual camera positioned 0.5 metres away from the 3D object centroid. Although the wireframes and PGHs look very similar as presented, for this projection, the Bhattacharyya match score between the 2 PGHs is just 0.67 (out of 1.0). The lower, lighter PGH imaged above was formed by subtracting the 2 presented PGHs. If the PGHs are identical, the difference image will be completely grey with 0.0 registered in every bin. The black and white regions of the difference PGH represent opposing structure in each overlapped PGH. Because much of the structure is adjacently displaced, the Bhattacharyya (overlap) score will be low or zero for these regions (See Figure 5.7). Problems are exacerbated for simple structures, such as the blue shaded wireframe regions above, because if there is otherwise additional stable structure across the PGH, the contribution of the mismatched elements to the match score will be down-weighted with global PGH normalisation. Most features will be much less affected by such pronounced shifts in perspective.*

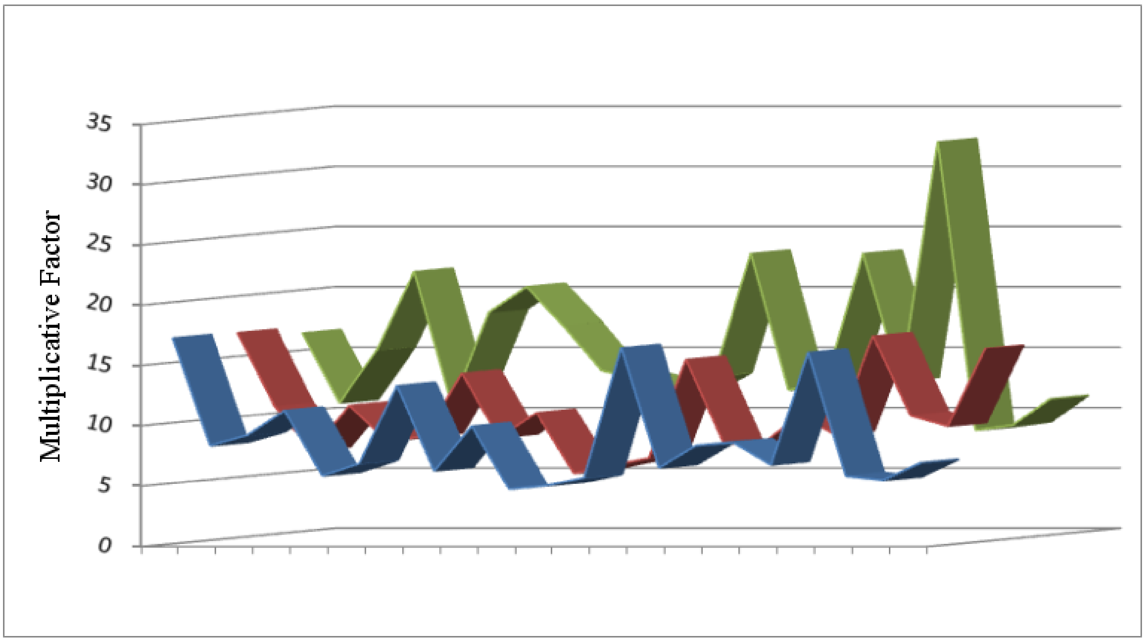
Figures 5.10-11 (see Section 5.5). For these experiments, an equilateral triangular reference region is grown about a sample view-point with fixed feature geometry, so that the specified minimum Bhattacharyya match score (0.9, 0.95 or 0.99 in this case) is maintained across the planar triangular sample points. Although the triangle is not rotated in the image plane so that it could perhaps be further extended relative to the specific terrain of the underlying shape manifold, the same tracks of reference points are sampled between the 2 methods (interpolative and nearest neighbour) so that the ratio of utility should be stable.

In order to sample the area of an equilateral triangular patch spanning the 3D view-sphere, the regular Euclidean formula for the area of an equilateral triangle can be substituted with angular triangle-edge lengths (relative to separation of vertices about the view-sphere origin).

A = Area of an equilateral triangle = $\frac{r^2 * \sqrt{3}}{4}$
r = side length (substituted here with angular vertex separation about the view-sphere origin)

For the nearest neighbour matching strategy, each PGH will be solely accountable for each triangular region. If using interpolative triangulation, each PGH will be shared by 5 or 6 triangles in a continuous web (see Figure 6.1), so that the contribution to the area of the view-sphere covered requires moderation. Because each PGH shares a third of the area of each of 5 or 6 triangles, the area covered by a single PGH will be 5/3 or twice the area of the triangular patch. This assumes a continuous triangulation of the view-sphere; otherwise the scope of each PGH around the edge of the region will be wasted. Considering the distribution of triangulated pentagonal and hexagonal reference regions sampled across the view-sphere (Figure 6.1) and the restricted nature of sample regions, a PGH utility scale factor of 1.5 has been sampled as an approximate guide for the presented experiments. While this figure would be closer to 1.0 for simple views with 5 or 6 learned triangular reference regions, many view regions may contain many triangles. It is otherwise impossible to predict exactly how many triangles may make up a view for a random object.

There approximately 7 times more PGH access-based operations (e.g. PGH normalisation, subtraction, addition and multiplication) involved with the proposed scheme for scale-based triangulation as compared to the number required for interpolating a single reference node through scale space. This includes, for instance, procedures to ensure the sample is within the specified triangular region and interpolating across



blue = 0.9 (minimum Bhattacharyya reconstruction score across the triangular region),

sample mean (blue) = 8.9

red = 0.95, mean = 10.6

green = 0.99, mean = 16.9

Figure 5.20: *Each coloured strip in the chart above is sampled with 3 prototypical edge feature PGHs from sample views of each of the 8 objects presented in Figures 5.10 and 5.11, as linearly connected in 24 points along the horizontal axis. The vertical axis represents the proportional increase in the area of the view-sphere covered by a single PGH using the proposed linear interpolation procedure instead of a simple nearest neighbour matching strategy at the prescribed levels of precision (0.9, 0.95 and 0.99 minimum Bhattacharyya match scores in the examples given).*

the outer edge instead if not. The mean values given in Figure 5.20 suggest that an increase of utility per unit area of the view-sphere of around 1000% is achievable if using interpolation. Because of the processing costs involved, this figure is reduced to around 40%. Any such advantages in using PGH triangulation may be further leveraged because fewer nodes would be required for recognition meaning that less memory is required and less time is required to read in the data. The advantage gained is shown to be more pronounced for more accurate regions, such as those sampled with a 0.99 minimum Bhattacharyya match score, as shaded in green, where this factor may rise to over 140%.

5.9 Conclusions

This chapter started by reviewing the nature of 3D object representative shape manifolds in PGH vector space. Virtual wireframe models have been adopted for use to allow precise control over all viewing parameters in support of analysis. Experiments showed that such manifolds can be locally approximated with isotropic planar regions, allowing for 3D objects' continuous ranges of projected appearance to be approximated by piecewise triangulation.

An interpolative scheme for the recognition of scaled objects across their piecewise triangulated manifolds has been presented. The supporting idea is that only a single representation of view-deformation around the view-sphere is required, with the image sampled geometry being scaled to match. Analysis of data distributions corresponding to the proposed mathematical framework supports the validity of the methodology, indicating that all sampled objects may be accordingly modelled though scale space for recognition. All object features examined were amenable to processing with the most accurate sampled manifolds for a scale change of 20%. This figure is indicative of the non-linearity of the 1D scale manifolds and means that objects' scale-based representations must be linearly sampled in intervals. The precision required and tolerable degree of scale change has however been shown to be object-view specific, meaning that feedback learning procedures could be implemented to optimally model the parameters of any such representations.

The effects of perspective on the validity of the proposed methodology for view and edge-based 3D object recognition have also been examined. To replicate machine-based interactive object learning, virtual object models have been constructed with perspective approximating that observable from 1.0 metre away. Analysis has shown that although such fixed perspective may be adequate to reliably recognise objects over short distances, some object features are more critically affected by perspective distortion, especially when viewed close up. In order to reliably recognise similar objects at different physical sizes irrespective of perspective effects, as required for generic 3D shape recognition, it is proposed that an independent model of perspective distortion is required. This can be instantiated as any number of complementary view-sphere representations, to enable linear interpolation to be conducted between them as required. For many object feature sets, perspective effects are however almost negligible, meaning that only a single model is required for recognition for all but extremely close up perspectives. Current research is directed around a single model of perspective through stable regions of view-space, such as 0.7 to 2.0 metres away (see Figure 5.17).

Finally, the advantages to be gained in using the proposed triangulation procedure for scale space recognition have been assessed in relation to the costs involved in performing nearest neighbour matching. An estimate of the relative utility of the two procedures was sampled as the relative proportion of the view-sphere covered by a single PGH in each case. Using interpolation, an increase in area of around 1000 to 1500% was realisable, with effects being more pronounced for the more accurate representation. However, further analysis indicated that approximately 7 times more PGH access-based operation were required for the triangulation procedure, suggesting that a 40 to 110% increase in utility was instead realisable through the different representations of precision. The advantages of interpolation are further reinforced because less PGHs are required in memory and corresponding object localisation is better informed.

Chapter 6

A Framework for Learning, Recognising and Localising 3D Objects using Pairwise Geometric Histograms (PGHs)

6.1 Introduction

Given the evidence presented in the previous chapter supporting application of a triangulated shape manifold representation for 3D object recognition, the issue remains of how best to triangulate objects' features' representative view-spheres in order for the prospective vision system to autonomously learn a specified object's range of projected appearance. Methods for accomplishing this task are considered in the following section. Subsequent sections of this chapter discuss issues associated with practical implementation of the proposed ideas in support of 3D model matching and object recognition, including the supporting stage of edge detection. An optimisation scheme for efficiently matching complex triangulated view manifolds is also presented. Finally, the methodology employed to localise recognised objects in 3D scenes is presented.

6.2 View-Sphere Modelling

The proposed process supporting 3D object recognition involves finding the best possible match score for a sample PGH across a triangular reference region delineated by 3 reference PGHs. Triangular regions are used for convenience as the simplest interpolative basis supporting polygonisation of any prospective shape manifolds. Reference triangles can in effect be grown about specified object viewpoints so that the worst possible interpolated match score across the triangle, as estimated at critical reference points at the triangle centre and mid-way along each edge, is kept above a prescribed threshold (e.g. a 0.95 Bhattacharyya score). More continuous mappings of objects' view-spheres can be acquired by iteratively growing connected triangular reference regions about the initial sample triangle, or more straightforwardly, by iteratively geometrically triangulating a predefined spherical mapping such as a dodecahedron or an icosahedron (Figure 6.1). The 20 vertices of a dodecahedron exist as standard Cartesian coordinates and an icosahedron is formed as a rectified dodecahedron.

In current work, 3D objects' representative wireframes are constructed from 3D edge features with view-point dependency files detailing which features are visible at a sample of representative viewpoints around the view-sphere. This partitions the view-sphere into an aspect graph type representation, describing regions of the view-sphere for which certain features are mutually visible. The nearest view to a specified point on the view-sphere is taken to indicate which features are visible from that projection. In order to maintain a relatively faithful representation of feature visibility throughout 3D view-space for reasonably complex objects, a relatively high number of reference nodes may be required. In this work, the view-sphere is aligned with a triangulated icosadehdral template, with a view being assigned to each of the 42 shared triangular vertices. A Voronoi tessellation of the view-sphere is then taken about these 42 reference points forming a mesh of approximately equally sized hexagonal and pentagonal view reference regions which can each be extracted as sets of 5 or 6 approximately equilateral triangles as the basis of triangulation. The geometrically triangulated icosahedron can otherwise be iteratively subdivided (in parts) to give more sample view-points if required to represent more intricately structured objects. Alternatively, simpler platonic mappings such as dodecahedra could be substituted giving 32 and 122 views instead of 42 and 162 for the first 2 orders of sub-division, as indicated in Figure 6.1.

Because PGHs are specified for individual features, a 3D object will have a (partial) view-sphere for each edge feature sampled for recognition. Given a set of views around the sphere, such as those in Figure 6.1, for each pentagonal or hexagonal view, each constituent triangle can be learned for the specified feature at the prescribed minimum level of precision. This is performed by sampling a PGH at each corner of the triangular region and ensuring that all points within the region can be reconstructed at the given level of precision (see 5.9). If the level of precision cannot be maintained across the region, the triangle is iteratively subdivided until the required precision is attained. More specifically, by positioning a new triangle within the first (upside-down), with the vertices at the mid-points of the edges, each approximately equilateral triangle can be divided into 4 similar smaller ones. The match score criterion is then applied independently to each, allowing for arbitrarily complex structure to be continuously modelled. A minimum vertex separation can be set so that computation is not wasted in modelling any potentially awkward regions of view-space that would in effect be unsuitable for recognition. Learned view-sphere manifolds are presented for 2 random object features in Figure 6.3.

While the proposed scheme for modelling object view regions may have limitations with regard to the fixed nature of the template and its ineffectiveness for detailing specific feature view boundaries, current research is more tuned with establishing the general merits of the PGH representation for view-based 3D object recognition. Because the proposed view-based recognition system is intended to be robust to missing features, as required for occluded object recognition, the system should be able to operate effectively without perfectly complete and faithful object representations. Potentially, any division of view-space could be represented in the proposed manner. The boundaries of an individual feature's visibility around the view-sphere could for instance be polygonised in accordance with aspect graph theory and the enclosed region could be efficiently filled via a Delauney triangulation procedure. The same iterative learning scheme could then be applied across those triangles. This would allow for a more natural partitioning of view-space taking better account of self-occlusion. After all, PGHs are matched on an individual basis, albeit lots of them sequentially. A finer view-sphere sampling could also be used, such as the 162 view icosahedron derivative shown in Figure 6.1, with the potential for any similar regions to be merged. Learning processes could otherwise be implemented to optimise the positions of any modelled view-boundaries so as to more efficiently fit the underlying data.

By iteratively sub-dividing a fixed global template, uniform sub-sampling of triangu-

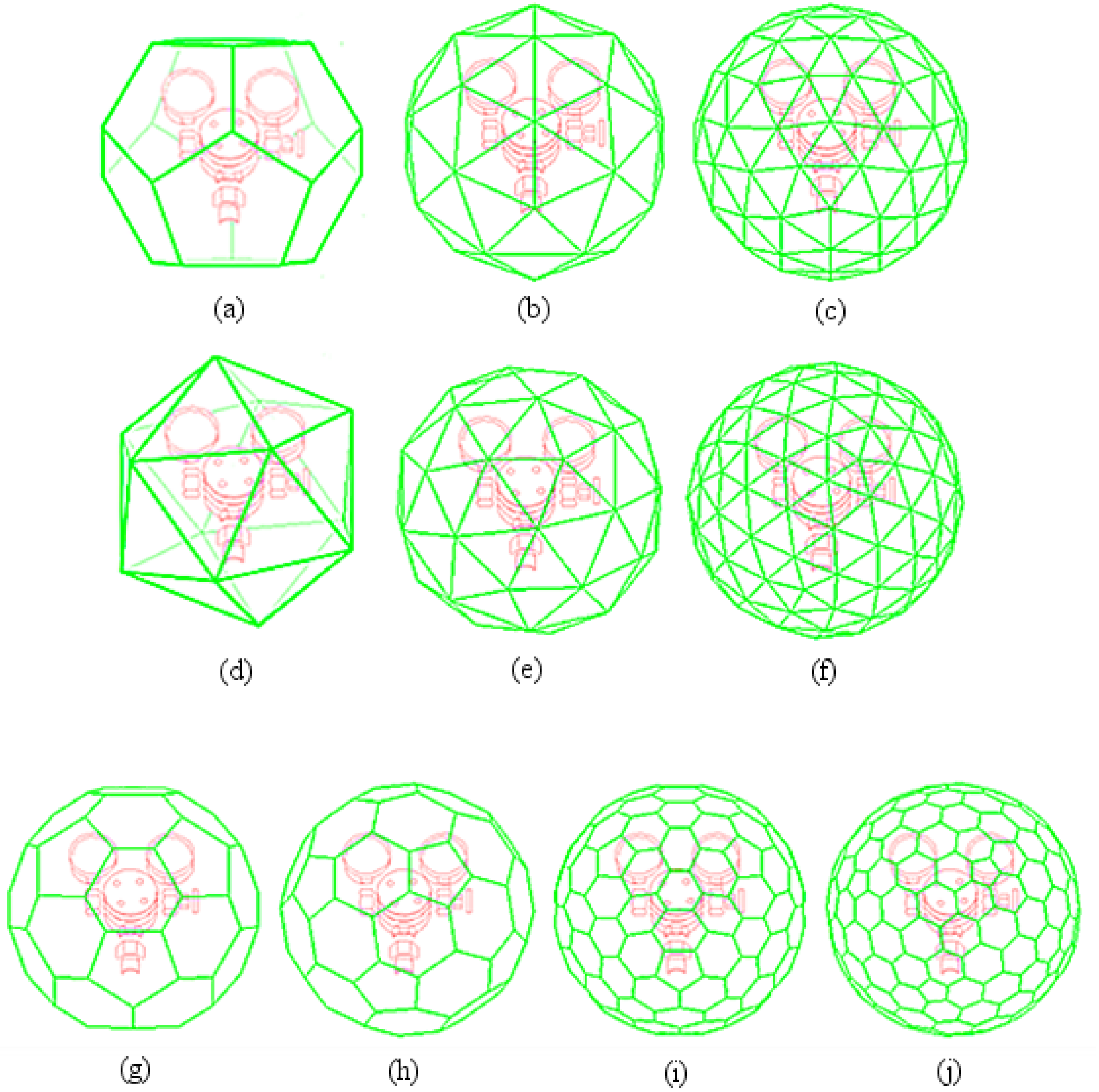


Figure 6.1: *Objects' view-spheres can be modelled about encompassing platonic solid constructs such as dodecahedrons (a) and icosahedrons (d), which can be iteratively subdivided with approximately equilateral triangles with each vertex being normalised to the sphere (b)(c)(e)(f). The spheres in the bottom row of the diagram are Voronoi diagrams representing the scope of each centred reference view (vertex) around the visible half of the view-sphere for the 4 spherical triangulations in the top right images (b)(c)(e)(f). There are 32 (g) and 122 (i) hexagonal or pentagonal views for the dodecahedral view-spheres and 42 (h) and 162 (j) for the corresponding icosahedral view-spheres. Figure 6.2 indicates how each view of the spheres in the bottom row may be populated by learned PGHs.*

lar regions may mean that certain triangular regions may be over-sampled, in that they may have a much higher minimum reconstruction error than that specified as the global minimum, i.e. the triangular regions could potentially be extended. The triangulation may therefore be sub-optimal in terms of number of required nodes, although representational accuracy would be increased. Techniques for iteratively growing sets of connected triangular reference regions have otherwise been investigated. Triangles could for instance be grown around an initial reference triangle in a closed spiral formation. The advantages of such techniques may however be diminished when modelling within specific regions (i.e. aspects), where the potential of such methods may be wasted filling in gaps around the edges of such regions. The potential for sharp contrast in the scope of any adjacent neighbouring regions also complicates any such scheme with the potential for the representation to fold in on itself. The previously outlined 'filling in' routines are otherwise more straightforward to implement and are still fit for the purpose, with associated aspects of adaptive optimisation being set aside for future research.

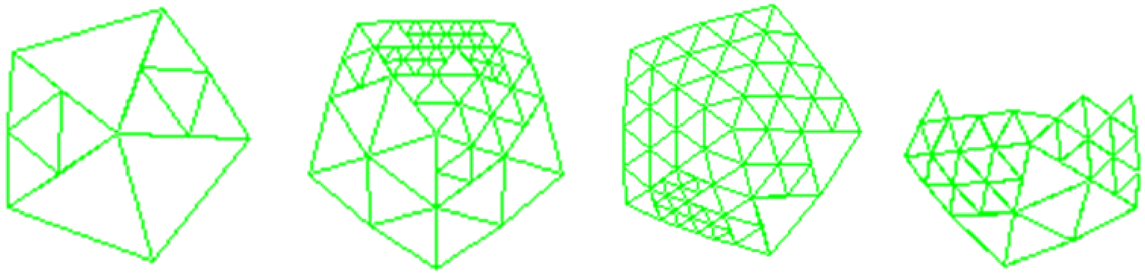


Figure 6.2: *The 4 triangulated pentagons pictured above represent learned regions of the view-sphere, as depicted as view regions in the bottom row of Figure 6.1, for 4 random object view features. Each vertex represents a reference PGH. The outer edges of the pentagonal regions are approximately 24 degrees wide, with the minimum Bhattacharyya reconstruction score being sampled as 0.95 across each example. The diagrams indicate the variance in complexity observed across typical objects' features' localised shape manifolds. The broken mesh is indicative of how self-occlusion may be modelled across a learned manifold for elliptical segments (See Chapter 4).*

Symmetric objects pose further modelling considerations in that they may not require full view-sphere analysis. The funnel object pictured in the previous section, for instance, is bi-symmetric, meaning that only one half need ever be learned, with the other being inferred as an opposite image. The triangulated icosahedron has 3

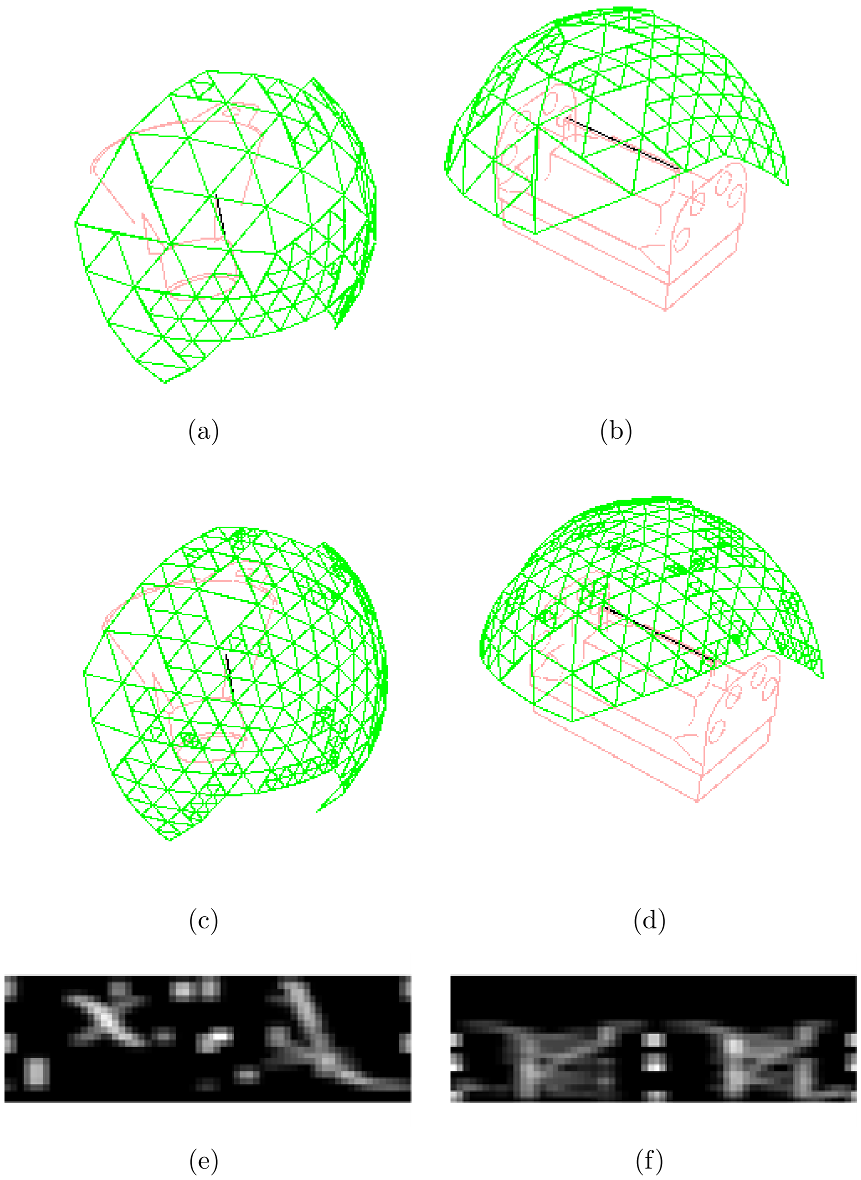


Figure 6.3: *The images above are screen shots of learned triangulated regions of view-spheres sampled for the funnel and wiper objects with PGHs sampled for the line features shaded in black (e.g. see (e) and (f) respectively). Images (a) and (b) represent the manifolds as learned to maintain a 0.9 minimum Bhattacharyya reconstruction error and those in (c) and (d) for 0.95. Such manifolds are learned for a range of features around the view-sphere, which are then matched to sampled image data to find best correspondence using the Bhattacharyya match quality metric.*

symmetric axes so that it could be split in half for such a purpose. By way of a more extreme example, an object such as a traditional wooden pepper grinder would only require a single arc of interpolative views from the top to the bottom of the view-sphere, with any other point on the view-sphere being equivalent to the corresponding point on the 1D arc.

The proposed method of learning is applicable to any 3D object with an edge-defined shape, with the possibility of optimised adaptation for symmetric objects. Although manually composed view dependency files have been used to determine feature visibility for virtual models in this work, the techniques should be extendible to deal with tangible 3D objects given appropriate robotic handling apparatus. In the first place, the 3D wireframe model could be generated and learned by stereo analysis. If stereo sensors were not available, the same processes could be inferred from a single camera with controlled orientation about the object. View-dependency files could then be learned in a similar manner with a feedback loop ensuring their validity. Because interpolation is only strictly valid across fixed sets of features, in accordance with aspect graph theory [9], such view dependency files naturally partition the object into fixed feature sets that are amenable to interpolation. The learned feature sets could therefore be sampled as views, with each being triangulated in accordance with the proposed learning methodology. Further analysis would however be required for view-based feature modelling, where stereo reconstruction is unstable because each camera will observe a different, although possibly well-correlated, 3D surface boundary.

6.2.1 Summary of Proposed Method for Learning a 3D Object's Range of Projected Appearance

For a given 3D wireframe object, a specified number (**N**) of the longest linear edge features to be learned for each sampled view region and a minimum Bhattacharyya reconstruction match score (**Bhatt_min** (e.g. 0.9)) to be maintained around the view-sphere:

- Using the virtual 3D camera and television model in the TINA software
    - fit a bounding sphere to each 3D wireframe

  - position the virtual model camera 1 metre from the virtual sphere centroid along the connecting line of sight to approximate perspective effects for a range of localised viewpoints (see Section 5.7)
  - ensure all objects are uniformly scaled in 2D so that the projected bounding sphere of the broadest object is contained within the 256 by 256 pixel model television screen

- Partition the object-encompassing view-sphere into 42 approximately uniform views (see Figure 6.1 (h)) (represented as triangulated pentagons or hexagons, i.e. with 5 or 6 triangles)
  - although 42 views are used for each object in current research, the method is applicable for any triangulated partition of the view-sphere

- For **N** of the longest projected linear edge features from each of 42 connected view regions

  (**i**) for each view triangle
  - for each vertex
    - ∗ sample image projected geometry for 3D edge features
      - · convert curved edge features to line segments
        - maintain a pre-set projected curve linearisation factor
      - · add (view-based) occluding boundaries
    - ∗ sample a PGH for a 50 pixel distance in each direction perpendicular to the supporting linear reference feature
      - · only sample features that are visible between all 3 triangulated viewpoints
      - · PGHs are currently 20 (±10) (5 pixel wide rows) by 64 ($\frac{\pi}{32}$ radian wide columns) (1280) dimensional
  - form a triangulated PGH representation space (see Equation 5.2)
    - ∗ Sample the worst case Bhattacharyya reconstruction error across the triangular view region (**Bhatt**)
      - · 4 sample points (see Figure 5.9)
        - mean triangular vertex (i.e. viewpoint)
        - mean linear vertex along each triangle edge
  - if (**Bhatt** < **Bhatt_min**)

- ∗ subdivide the connecting triangle into 4 similar smaller ones
  - – disregard original triangle
  - – disregard triangles with vertex separation below tolerance
  - – repeat the process (**i**) for each triangle

As detailed in the previous chapter, objects' appearances are learned at a fixed scale, with the sampled image edge geometry being scaled (in intervals) to match. Because objects' edge-based projected appearances may be substantially different through scale extrema, a limited number of such representations may be required for truly generic object recognition.

Although wireframe models are currently used as the basis of learning, given robotic handling and visual inspection equipment, the same processes could be effected for tangible objects.

6.3 PGH Matching Considerations

Assuming that 3D objects' edge-based projected appearances can now be continuously modelled, further consideration is required regarding matching any such representation to image data. In previous associated research regarding 2D object recognition, each scene line's PGH was matched to each object's PGH with any well supported object match hypotheses being Hough transformed for match parameterisation. The extension of object recognition into 3D poses similar requirements, with some form of accumulation of coinciding object view features required to evoke recognition. A Hough transform could be implemented for each view of each object to this effect, so that analysis of an image would result in a set of votes for each view of each learned object. Any well supported views could then be accepted as recognition hypotheses to be verified by the proposed wireframe alignment and validation procedures (see Chapter 4). The task of object recognition is not however typically this clear cut. Any individual PGH match hypotheses are liable to be affected by scene clutter and occlusion and other imaging artefacts, meaning that a valid match may be buried in the noise field amongst other coincidental feature matches. If there is a clear cut best match from the object database to an image sampled line, this may be cast as a single vote. If there is no definite best match, then a certain percentage of features over a minimum prescribed match score can each be voted for with their contribution down-weighted by the number of features comprising the set. To account for occluded object features more reliably, the proportion of votes cast can be increased. Although it is otherwise desirable to limit the number of votes made so as to not obscure any valid match responses buried in the noise field, previous research has identified that for larger object libraries, any interfering coincidental feature matches will be more diluted in the larger representation space, allowing any valid matches to be more readily identified.

One potential problem with a view-based voting mechanism for 3D object recognition arises when an object needs to be recognised at a view boundary. Because of the potential for noise on any PGH-based recognition inferences, such an occurrence may mean that votes are cast across either side of the view boundary for different features, meaning that the vote would be effectively halved (albeit twice). Overlapping accumulator bins could instead be constructed to alleviate such issues, or cross-bin assessments could be made. Alternatively, in current research, each vote is accumulated as the best sub-triangle interpolated matching view point across the view region and a sample of the best matches is taken as a depth first search seed list. A Hough transform type parameter specifying the tolerable degree of misorientation between the interpolated matched view-points is then used for each match on the seed list to accumulate other coinciding match hypotheses. Any repeated lists are then discounted and the remainder are ranked as match lists in terms of the sum of their corresponding weights. As well as grouping feature-based object match hypotheses together on the basis of recognised object view-point correspondence, the tests can be bolstered with scale, centroid and image-orientation constraints to ensure the validity of any coinciding feature sets. View-point and scale are utilised for model matching in current work, with the precision of any such grouping parameters being relative to the precision used to model the representative shape manifolds.

One obvious concern with such a view-based recognition scheme is the amount of data that requires analysis, especially if high levels of precision are required to model 3D objects' manifolds. While the brain may be well suited to such recognition tasks with its massively parallel architecture, in using a standard computer with a sequential processing stream, we are restricted to examining each view of every learned object in series. Even with the speed of modern day processors this may render any such view-based recognition scheme intractable for real-time recognition across large object databases. This raises the question of whether any optimisation schemes may be utilised to ease the computational burden.

6.4 Optimised Triangulated View Manifold Matching

As an object is (semi-locally) rotated away from a sample view-point, the corresponding Bhattacharyya match scores should fall off monotonically. Because of this, for valid matches, the match scores sampled across the aspect manifold (i.e. for pentagonal or hexagonal views) should equivalently descend around the optimal solution. This means that a gradient descent type optimisation routine may be implemented to search through any potential view regions without having to examine every single triangular sub-region.

With reference to the step by step examples presented in Figure 6.4, an image line sample is initially matched across each triangle connected at the centre of the specified view region and the triangle with the highest match score is referenced. Any triangle connected (at a vertex) to the winning one that has not been searched before is then matched and the triangle with the highest overall match score is updated. This process is repeated until no improvement in the optimal match score can be made. The highest match score and corresponding interpolated viewpoint (see Chapter 5) for the winning triangle are returned as the solution.

Although this operation could be made much more efficient by only matching individual nodes (without interpolation) across the view region, practical assessment showed that the raw PGH nodes could not be reliably used for this purpose. Interpolative match scores may be substantially different from those realisable through direct PGH matching with the raw reference nodes.

Although the proposed optimisation scheme may be liable to distraction from local minima across very wide partitions of the view-sphere, the method is very well suited to the proposed view-sampling method where view-spheres are divided into 42 approximately equally sized regions. For complex view regions where the valid triangular region is central to the view, this could make the difference between searching a handful of triangular regions instead of a hundred, to exactly the same effect. Because the method has every chance to escape from any local minima across the (semi-localised) view region, the scheme proves to be very robust regardless of the complexity and irregularity of the view region. Even if the method misses the very best match possible across the manifold, it is likely that it will still find a very close

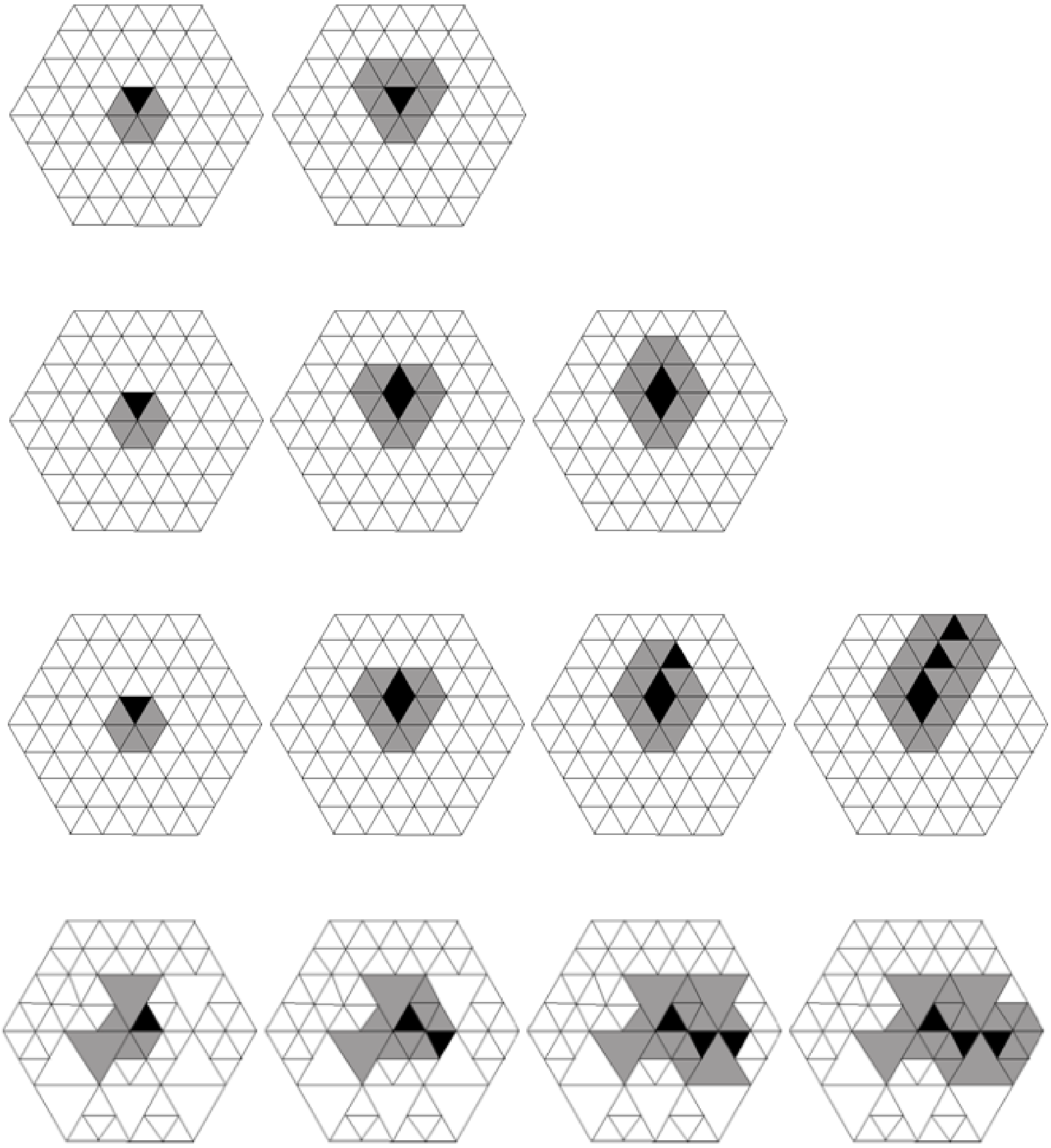


Figure 6.4: *Each row of triangulated hexagons represents a left to right transition of the steps involved in finding the best match across the hexagonal view region using the proposed next-best connected triangle search strategy. The best matching triangular region is initially selected about the grid's origin as indicated in the diagrams to the left. The grey shaded triangles are ones that have been searched at each step and the black shaded ones represent the triangles with the best interpolated match scores. At each subsequent step, any new triangles sharing a vertex with the previous best matching triangle are matched. If no improvement can be gained by moving to an adjacent triangle then search is concluded with the last selected triangular region with the corresponding match score and interpolated viewpoint being returned as the solution. As evidenced in the figures to the right of each row, only the shaded triangles need to be searched so that processing is not wasted exhaustively interpolating across all of the hexagonal web's triangles.*

match with no adverse effect on subsequent object recognition.

Another possible recognition optimisation strategy is that of coarse to fine search. This would involve an initial matching stage using low resolution PGHs so that any unlikely features can be efficiently disregarded from consideration before the computationally expensive, full resolution PGH matching processes are implemented. This is a possible avenue of future research.

The proposed mathematics for scale space recognition assumes that PGHs are normalised. There is a subtle issue here in that if an image sampled PGH is corrupted by entries arising from irrelevant scene clutter, upon normalisation, the contribution of the valid edge features will be down-weighted, thus possibly impairing recognition. Because recognition is to be assessed over multiple features, this should not make much of a difference to the outcome of the recognition task. If it were found to be a problem, normalisation could instead be performed only across histogram bins corresponding to those which are non-zero in the learned model sample. Most of the interference from the clutter could therefore be removed from consideration.

6.5 Edge Detection

The Canny edge detector [14] is the most widely used edge detector in the computer vision community. The Canny edge detector's applicability is due to the way in which it produces well localised and relatively unbroken edge segments. This is achieved by the process of non-maximal suppression, which suppresses all but the most prominent ridges of any gradient regions, coupled with hysterisis thresholding. The latter process involves defining two threshold values, where only edge pixels with a value above the highest threshold are initially selected. Any remaining edge pixels with values above the lower threshold are then additionally selected if they extend the edge segments initially determined. The locations of any edge segments can be practically interpolated to sub-pixel precision (empirical observation suggests that this is accurate to approximately 0.1 of a pixel). As with the previously described TINA 3DMM system, an implementation of the Canny edge detector is to be employed by the proposed view-based object recognition system.

Although the objects in the current test dataset are all amenable to step edge detection, such approaches to feature detection are not well suited for the analysis of thin line or ridge features in images for which an alternate methodology is required. A more generic approach to edge feature detection, capable of detecting both step edges and thin ridges, would be to use a variance calculation for a specified pixel defined over a local region of N pixels as the basis of the existing Canny extended-feature detection routines. I.e.

$$V \;=\; \frac{1}{N}\sum_{i}^{N}(I_i \;-\; \hat{I})^2 \qquad where \qquad \hat{I} \;=\; \frac{1}{N}\sum_{i}^{N} I_i \tag{6.1}$$

Where I_i is the image grey-level. This can be applied over a weighted region (e.g. Gaussian) in order to improve the noise characteristics and achieve rotation invariance.

$$W \;=\; \sum G(r_i)(I_i \;-\; \hat{I})^2 \qquad where \qquad \hat{I} \;=\; \sum G(r_i) I_i \tag{6.2}$$

where $G(r_i)$ is a radial weighting. This can be rewritten in convolution form as

$$W \;=\; G \otimes (I^2) \;-\; (G \otimes I)^2 \tag{6.3}$$

Image points with high values on the basis of this measure are those that have locally large grey-level variation, corresponding equally well to step edges and thin line structures.

Although the variance-based method proves robust as a general feature detector with more potential power than a standard step edge detector, as discussed, virtually all the features comprising the appearances of the 15 3D objects in the test dataset are step edges. Without any opportunity to detect the features for which it was implemented, the variance-based operator instead proved too responsive in other terms, being far more likely to detect any spurious edge features across noisy (e.g. corroded or textured) surfaces, thus confusing any edge maps and breaking the continuity of certain extended step edge features. Furthermore, the variance operator proves more computationally expensive. Considering that current research is more focused on establishing the merits of the PGH representation for view-sphere modelling, the original step edge detector is maintained for current research analysis on account of its speed and generic reliability for detecting unbroken extended step edge features.

Image edges are currently only extracted at a single scale. While other authors have attempted to base object recognition on multiple synchronous-scale edge analysis, the edge features that we are most interested in are those that can be reliably extracted in high resolution. Broader sampled (more blurred) edge features by definition carry less useful information for both recognition and subsequent localisation.

6.6 3D Object Localisation

Given a potential match between a sampled image edge line and an object feature's triangulated shape manifold, the remaining task is that of 3D object localisation. Although some approaches to object recognition [68][71] limit themselves to simply indicating whether a known object is present in an image, or region thereof, many vision-based applications require subsequent interaction with observed objects, meaning that the relative position and 3D orientation of the object is required. Furthermore, in order to reliably verify the validity of any hypothesised model matches, direct correspondence between model and observed features may be required, which can only be inferred against an oriented projection of the hypothesised 3D model in correspondence with the original image data.

As detailed in Chapter 4, the proposed view-based recognition system is supported by a robust projected wireframe alignment optimisation sub-system, which is capable of accurately aligning an objects' representative (2.5D*) edge model whilst optimising any projection parameters given just an approximate initial indication of the projected pose of the model. Although accurate initial estimates of model match parameters may be helpful and may reduce any optimisation processing costs, only an approximate indication (e.g. up to a 10 degree misalignment) of 3D object pose is typically required from the outset.

The standard methodology for 3D object localisation is to infer the projection parameters from the particulars of sets of recognised features in the image. Given 3 corresponding projected points, e.g. corner features, it is possible to infer the likely 3D pose and scale of an imaged object [24]. Because of the high likelihood of incomplete line features being present, this is not generally possible, meaning that the transformation needs to be ascertained from 3 extended line features. There

*2.5D refers to a 3D wireframe model representing only the features that are visible throughout a localised region of view-space, e.g. the features visible between two cameras in a stereo rig.

is evidently no established solution to this specific problem in the computer vision literature.

Although PGHs are designed to be employed in clustered feature sets for recognition, correspondence with a single PGH may be enough to support 3D object localisation given a detailed enough view-sphere mapping (as is typically observed). Although such assumptions would break down if using a sparsely sampled nearest neighbour matching strategy, the previous sub-chapters have identified the potential for continuously modelling connected regions of the view-sphere via triangulated interpolation. By matching a sample PGH to any such region, we are able to directly extract the (intra-triangle) parameters of the best match across the region, so that the corresponding viewpoint can be inferred to a degree of precision moderated by the initial interpolation precision parameter.

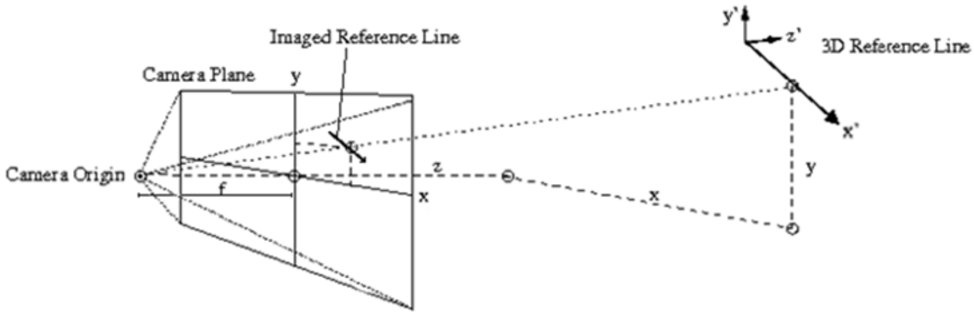


Figure 6.5: *The diagram above shows a 3D reference line projected onto an inversed pinhole camera projection plane. The immediate problem is one of inferring the x, y and z components of the camera-based world model for the 3D reference line's mid-point. The corresponding 3D object geometry is then based about the situated 3D reference line.*

A projective transformation from the recognised 3D model line, in its own coordinate system, to the camera's 3D coordinate system is required (see Figure 6.5). This transformation can be broken down into a rotation (and minor translation) component (see Figure 6.6), serving to align the 2 line features and a major translation component (see Figure 6.7), identifying the relative coordinates of the centre of the 3D reference line. Because the directed PGH format is being used, a matched image line will also indicate which way round the 3D model geometry lies relative to the projected line, so that the rotation of the object can be unambiguously solved (i.e. against its 180 degree rotated counterpart).

Although 3D model lines can be straightforwardly matched, complications arise when

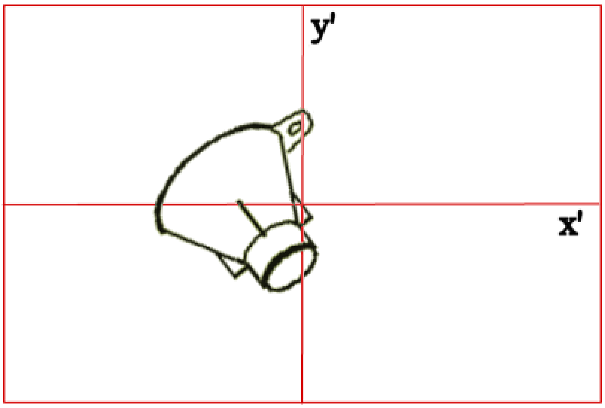
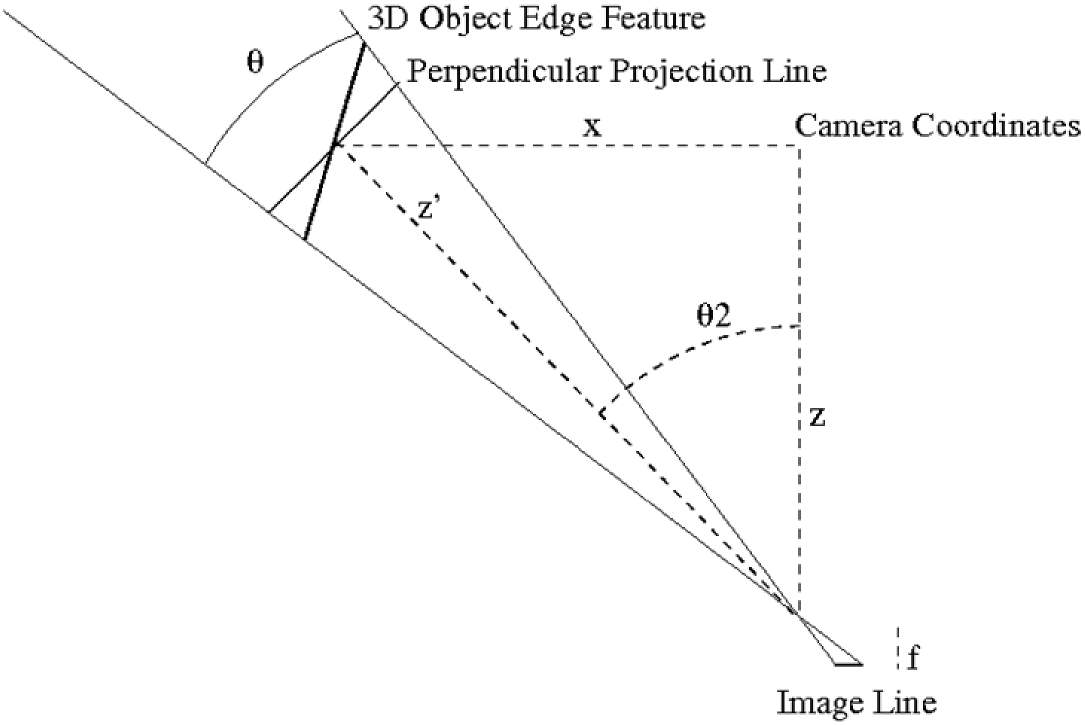


Figure 6.6: *In order to infer the relative 3D location of a recognised 3D object, the system essentially transforms the representative 3D wireframe model view (z') to the recognised viewpoint, centres the recognised 3D line feature in the projected reference frame (x'=0, y'=0, z'=0), rotates the object in the image plane (about the z' axis) to co-orient the recognised (directed) projected model feature (as exemplified above) with the corresponding image edge feature and the then transforms the coordinates of the recognised 3D model by the 3D vector (x, y, z) defined in Figure 6.7.*

dealing with elliptical features and occluding boundaries. The problem with elliptical features is that extracted linear samples require anchoring to specific points on curves and the linear extension may change with viewpoint. Anchor points are therefore specified along curved 3D features for any associated PGH mappings. If a curve fragment is then hypothesised as a match at a certain view-point, the corresponding projected line can be grown out about the reference point and the 3D end-point parameters of the line can be sampled along the curve. Similarly, view-point-dependent occluding boundaries have no fixed 3D parameters. Instead, for a specified projection, the 3D parameters of any occluding lines can be sampled as the points of connection with their supporting 3D elliptical features.

The first step of the localisation process is to infer the relative angular displacement between the 3D model's recognised feature and its 2D image projection. The 3D model first needs to be rotated so as to be viewed from the recognised interpolated viewpoint. The model then needs to be translated so that the recognised 3D line feature is centred at the origin of the model's coordinate system, allowing the rotation between the orthogonally projected 3D edge feature and the corresponding image line to be calculated about the z axis. These 3 steps are accumulated as the rotation transformation for the camera's projection model.

Because the recognised scene line may only be a fragment of the associated projected



Perpendicular Projection Line = view-cone centered orthogonal projection of 3D Object Edge Feature

f = camera focal length

θ = angle between Image Line endpoints in x-axis camera coordinates and the camera origin

d = width of Perpendicular Projection Line in x, z

$z' = (d / 2.0) / \tan(\theta / 2.0)$ = length of hypotenuse of x, z

$\theta 2$ = angle between (z & z') Image Line mid-pt in x, camera origin and camera origin shifted by f

$\theta 3$ = same as $\theta 2$ in alternate y, z coordinate frame

$z = z' . \cos(\theta 2)$, $x = z . \tan(\theta 2)$, $y = z . \tan(\theta 3)$

Figure 6.7: *The diagram above shows x and z projections of the assumed pin-hole camera projection model indicating the geometry required to ascertain the x and z camera frame coordinates of the mid-point of the 3D Object Edge Feature. The y component follows in a similar manner, with the y axis replacing the x axis in the diagram. The 3D distance to the centre of the Perpendicular Projection Line sampled orthogonally to z' is first calculated to indicate the depth of the mid-point of the 3D Object Edge Feature and the corresponding location in the x and z plane. The x, y and z coordinates of the 3D reference line follow.*

model line, the parameters of the full projected model line can be inferred from a projection of the 3D object model at the recognised view-point and scale relative to the model scale at which the PGH was learned. As indicated in Figure 6.7, trigonometry can then be used to infer the x, y and z camera coordinates of the centre of the model line in the field of vision, thus forming the translation component of the transformation for the camera's projection model. The 'Image Line' in Figure 6.7 represents the full model line projected from the 2D-aligned view-point-matched 3D wireframe.

The proposed approach to 3D object localisation matches the 3D model around the centre of the reference line feature. This may present a problem if only a fraction of the edge feature is visible in the image. Although the post recognition alignment optimisation process may well alleviate any such issues, the system can infer whether the detected image line feature is fully visible by virtue of projected length. If the feature is detected at a smaller scale than that expected, localisation can be re-sampled at intervals along the adjusted length of the hypothesised feature. The optimisation method (simplex) could prospectively be adapted to explicitly account for any such localisation ambiguity. Otherwise, a second non-collinear recognised edge feature could be used to constrain the localisation along the length of the first feature. This would however introduce the risk that an erroneous feature match would invalidate the method, while it is convenient to only require a single matching feature to be found to support recognition and localisation. The dual (non-collinear) feature localisation constraint was incidentally more critical to the localisation techniques associated with 2D PGH-based object recognition (see Chapter 3) because the process had no provisions for localisation optimisation unlike the proposed 3D methodology. The reported model matching experiments (Chapter 7) prove that the proposed single feature-based localisation method performs adequately well.

6.6.1 Summary of the Proposed 3D Object Localisation Method

- Get transformation matrix between 3D model (in its own coordinate system) and its recognised 3D scene location (in the camera's 'world' reference frame)
  - Get transformation Rotation component (See Figure 6.6)
    - rotate recognised viewpoint to model television view point (z' axis)

        - shift matched 3D line feature mid-point to origin of reference frame (0,0,0)
        - rotate 3D model about view direction (z') to co-orient matched directed model line with recognised (directed) image line
    - Get transformation Location component
        - i.e. the relative location of the centre of the recognised 3D line feature
        - see Figure 6.7: x, y, z
- Optimise the transformation parameters using the methodology outlined in Chapter 4 (4.8)

6.7 Conclusions

This short chapter has introduced a scheme for partitioning 3D objects' view-spheres into sets of representative views that can be used as the basis of automated learning. Learning proceeds as an iterative sampling of each view's constituent triangulated regions, maintaining a minimum reconstruction error in accordance with the mathematics outlined in the previous chapter. An efficient gradient descent type optimisation scheme for analysing objects' triangulated view manifolds has also been presented along with the theory underpinning the proposed method for localising 3D objects via a single matched line feature.

Chapter 7

3D Object Wireframe Model Matching: Stereo versus Mono

7.1 Introduction

This chapter analyses the applicability of the Pairwise Geometric Histogram (PGH) representation for the task of 3D model matching. 3D model matching is a sub-task of 3D object recognition, concerned with localising instances of specified 3D objects in scenes in primary support of robotic interaction tasks such as bin-picking and assembly. The efficacy of the Pairwise Geometric Histogram (PGH) representation for 3D model matching is a prerequisite for application to the more demanding tasks of object recognition and scene interpretation, which are discussed in the following chapter.

Stereo vision has been adopted for 3D model matching in previous associated research [79]. Stereo vision is invaluable for depth perception in human vision and can similarly be harnessed by any prospective robotic vision systems. Stereo is potentially especially useful for 3D object localisation because the relative 3D positions of an object's features are directly realised. However, despite the advantages to be gained in being able to directly interpret visual data in 3D, stereo perception has a number of limitations for the model matching and object recognition tasks.

Having detailed the operation and performance of the proposed view-based 3D object model matching system, a comparative study of the pre-existing TINA stereo-based

3D Model Matcher (3DMM) is presented. The relative merits of the competing methodologies are practically assessed in terms of model matching performance across a test set of 15 full-view-sphere-sampled 3D objects.

7.2 The TINA View-Based 3D Model Matcher (VB3DMM)

The proposed strategy for learning and subsequently recognising 3D objects' edge feature-based PGHs is detailed in Chapter 6. In summary, each object's wireframe is currently manually modelled with 42 approximately equally-spaced views detailing feature visibility around the view-sphere. Given such a view-based 3D object representation, each object's appearance is automatically encoded (i.e. learned) in terms of a set of triangulated feature-specific PGHs. In order to make the scheme more tractable, for each of the 42 views, a reduced number of 'key' features are sampled. In the current context, the 12 longest features from each view are used as the basis of model matching with a 50 pixel projected perpendicular distance being used to sample surrounding features relative to the supporting 256 by 256 pixel virtual model television. For the presented experiments, each PGH encoded shape manifold is learned to maintain a minimum 90% reconstruction accuracy level so as to limit memory and processing requirements while maintaining a reasonably high degree of representational precision.

Given a learned object model, model matching proceeds by sampling selections of linear image edges in their local edge contexts with PGHs and searching the learned manifolds for best correspondence. The process of scale-based PGH analysis is detailed in Chapter 5. In brief, each image edge sample is scaled between upper and lower bounds, so that the learned triangulated manifold can be further interpolated between these bounds to achieve local-range scale invariance. Contiguous scale samples offer invariance over extended ranges of scale, assuming that the representative model is stable through such deformation. In order to limit temporal processing costs, a single uniform scale sample equating to a 15% shift in projected scale is used for current experimentation.

In accordance with selection of the longest model lines for each object view, sets of the longest image edge lines are preferentially selected for model matching. Model matching proceeds by sequentially matching each image line to the set of learned manifolds. Because the process of reading in learned manifolds can be computationally demanding, the system currently samples 8 image edge lines at once, meaning that the data access overhead can be reduced by a factor of 8. Although 8 image lines may be ample for matching many objects, other objects may be predominantly composed of elliptical features with very many projected linear edge segments, meaning that more image line features may be required in order to achieve correspondence. Similar problems arise with interference from scene clutter and shadow features. The next longest set of 8 edge features can therefore be iteratively sampled until all image evidence has been assessed or a well supported match has been found. For the presented experiments, a maximum of 3 iterations has been performed, accounting for the longest 24 linear edge features in each image. In order to reduce the effects of any scene interference and to obtain a good cross section of features, a process of sampling the longest image lines from each of 8 uniform orientation bins has been investigated. This mechanism however proves ineffectual for objects such as the desk-tidy, for which all the longest lines are collinear and for which distracting correlated reflections and specular highlights coexist.

Although many objects' features may be relatively stable through changes in viewpoint, some features' PGH-based shape manifolds may be highly complex, especially if high levels of precision are maintained. In order to avoid exhaustive sampling of sub-manifold triangular reference regions, a process of optimised connected triangle 'hill-climbing' is adopted to reduce search costs for complex manifolds (see Chapter 6).

Because objects' inferred image edge representations may be highly incomplete due to environmental illumination conditions, for many objects, correspondence between image and model sampled PGHs will be far from perfect. Such problems are exacerbated with coincidental matches between similar features and amongst any distracting image detail arising from shadows, specular highlights (and other reflections) or scene clutter. Although a reasonable PGH-based Bhattacharyya match score may not therefore be sufficient for unambiguous object detection, correspondence regarding detected object orientation and scale across multiple image features may be used as a more reliable basis of object detection. In the current implementation, as will now be detailed, a selection of the highest scoring PGH match hypotheses are therefore extracted and clusters of concurrent votes are formulated as ranked match lists.

As discussed, the 8 longest image edge lines are iteratively sampled as the basis of model matching. A match list is then constructed for each of the 8 image edge features against the 12 longest (linear) model features from each model view. Initially, for each sampled view (i.e. currently 42), each feature scoring over a prescribed Bhattacharyya match limit (currently 0.5) is added in ranked order to each line's match list. Analysis of the distributions of valid and invalid match scores following experimentation across the object test data set should otherwise allow for any such parameters to be moderated to more useful effect. Once each of the 8 sample edge feature's individual match lists is constructed, any features with a match score less than a certain value (currently 95%) of the best match are discarded along with any sub-optimal entries for repeated matches to the same model feature. The remaining features are then weighted by the number of entries left on the match list. The rationale here is that a single distinctly matched feature will be treated as a single vote, whereas, if a number of features are confused, their contribution to the object matching task should be down-weighted, in this case with their vote being normalised by the number of competing entries.

Having formulated a potential model match list for each sampled image edge feature, the match references are pooled into a global image match reference list. As detailed in Chapter 6, any individual feature-based match hypotheses agreeing on the pose and scale of the detected object are then grouped together into global object match lists, which are then ranked by the sum of the individual weighted votes. Ideally, a minimum level of support in terms of the number of entries comprising a match list and the combined weight should then be required before verification procedures are executed. The process of model matching can however be quite slow and there is the potential for interference from irrelevant scene clutter and occluded object features. In the current implementation, each entry on the top 4 match lists, or any other with a weight equivalent to the 4th entry, is therefore verified (as explained below) after each iteration of 8 image lines, regardless of list length or weight. The process is stopped when a verification score sufficient to reliably indicate a valid model match is attained.

In order to infer the 3D pose and location of a hypothesised 3D object, only a single PGH is required. As explained in Chapter 6, the best interpolated position across a learned shape manifold will indicate the approximate view-sphere pose of the object and the image plane orientation of the object can be unambiguously inferred from the 2D orientation of the image line along with directed PGH-based correspondence.

Given an initial estimate of the 3D pose and location of the object relative to the camera frame, the quality of match between the projected model features and the underlying image evidence can be assessed by virtue of the quantitative probabilistic metrics defined in Chapter 4. Optimisation of these terms allows for the inferred pose and location of the object to be optimised with the simplex algorithm [36] and for an assessment of the quality of match at the optimised solution to be sampled as a probabilistic validation measure. The validity of the edge correspondence metrics and associated thresholds are reviewed in section 7.3.

7.2.1 Summary of the Proposed View-Based 3D Model Matching Routine

- Learn a 3D object's projected edge-based appearance around the view-sphere according to the PGH sampling methodology outlined in Subsection 6.2.1
- Match the N (e.g. 24 (in sets of 8)) longest scale interval sampled linear image edge feature-based PGHs to the M (e.g. 12) longest linear object edge feature-based PGHs from each of the views (e.g. 42) populating the view-sphere (see Chapter 6)
    - scale the sampled image edge geometry in consecutive sets of intervals (e.g. 15% of a calibrated base scale (depending on precision and object (see Section 5.6))) for a desired range of semi-constrained scale (e.g. for 1 octave)
        - use a ±50 pixel perpendicular sample range for each line feature's PGH in the (currently) 576 by 432 pixel images (see Figure 3.2)
        - keep the best match across each feature's range of scale intervals
    - use the next best connected triangle gradient descent method outlined in Section 6.3
    - only keep matching features over a minimum Bhattacharyya match score (e.g. 0.5)
- For each image edge feature, keep the best matching object view feature and any other features within tolerance (e.g. 95%) of the best Bhattacharyya match score

- – add entries to a global match list ranked by Bhattacharyya match score

- Formulate a match seed list as the best X (e.g. 10) matches from the global match list
  - – remove any duplicated seed list entries agreeing on the view-point, scale and image-plane-orientation[*] of the model match hypothesis (ref. preset parameters, e.g. (respectively) $\frac{\pi}{10}$, $\frac{scale\_interval}{6}$ and $\frac{\pi}{6}$)

- For each entry on the match seed list, go through the global match list and gather any other feature matches agreeing on the pose, scale and image-plane-orientation of the hypothesised object
  - – formulate a potential match list
    - ∗ weight as a sum of the individual feature match scores[†]
    - ∗ add to parent match list (forming a list of model match hypotheses)

- Verify each match list on the parent match list in weight ranked order
  - – currently, only top 4 match lists verified
    - ∗ possibility to ensure minimum match list length maintained (e.g. 3)
  - – localise each entry on each model match list in (individual) weight ranked order (see Subsection 6.6.1)
    - ∗ verify each localised model match (see Chapter 4)
      - · edge location only metric used for localisation and verification (with a 1% confidence limit on each sample point for verification)
      - · possibility to stop verification processing when a sufficient verification score is achieved (see Section 7.3 (e.g. 90% for the edge location only metric with a 1% confidence limit))
  - – return the model match with the highest verification score as the solution

[*] The image plane orientation constraint was retrospectively added to the model matching routine having been found to be beneficial for the more demanding object recognition task (see Chapter 8).

[†] Match list scores were originally formulated as sums of votes, where features' votes were scaled from 0 to 1 according to the number of confused matching features. The match score weighted metric was retrospectively fitted to the matching routine during subsequent object recognition oriented research (see Section 8.2). If enough features can be matched, model match hypothesis scores would ideally account for the observed proportion of total edge feature length making up the hypothesised object's projected appearance.

∗ multiple objects of the same shape may be matched using a supplementary constraint on the detected image position of each object (e.g. a projected centroid) (only a single instance of each object is currently available). I.e. a match score can be formulated for each detected object instance, rather than for each object for each image

7.3 Quantitative Projected Edge Feature Verification Review

Chapter 4 defined a quantitative statistical framework for the verification of sampled projected edge feature points in terms of correspondence with edge location and orientation. Now that an associated model matching system is available, it is possible to evaluate the utility of the proposed approach to verification in terms of discrimination between valid and invalid instances of model matches. For these experiments, 2 random images of each of the 15 objects in the test dataset have been accurately model matched with valid model views and the proportion of sampled edge feature points passing a specified hypothesis test have been sampled. For each of the 30 test images, 3 other random objects from the set of 15 have been optimally matched to the image edge data with the same criteria. The experiments have been repeated with and without the additional constraint of edge orientation to allow any advantages of using edge orientation information to be observed in terms of valid/ invalid class separability.

For the verification experiments conducted in Chapter 4, a 1% probability has been used to validate feature points along extended edge features. This level has been adopted in order to allow virtually any suspected edge feature point to be validated, as may be required in unfavourable illumination conditions. However, despite accounting for image noise, such a weak constraint is also liable to occasionally verifying irrelevant edge feature points, such as ones observed across texture in the cloth background used for experimentation. Higher confidence limits may therefore be better suited to the competing model match verification task. The current batch of experiments have therefore been repeated with a range of confidence limits, i.e. 1, 15, 30 and 45%. If a confidence limit other than 1% is better suited to the task of differential model match verification, the observed distributions will be better separated at that level, allowing for more reliable and discriminative verification to be inferred.

The 8 histograms representing the outlined experiments are presented in Figure 7.1. With initial regard to the utility of the additional edge orientation constraint, it is evident that there is no distinct advantage to be gained in this regard. Although the distributions appear to be stretched out along the x-axis (i.e. the proportion of feature points passing the test (0.0 to 1.0)) with a slight negative (i.e. unfavourable) skew to the invalid distribution, there is little in the way of enhanced class separability. Indeed, the clearest-cut class separation is maintained for the edge location only distribution (at a 1% confidence limit), with a very well-defined peak for valid model matches. ROC curves are presented for the data in Figure 7.2, confirming that optimal class separability is maintained for the edge location only metric. So while the negative skew of the invalid match sample distributions may be advantageous in down-weighting any obviously wrong coincidental matches (which would be useful if dealing with occluded objects), the clearest cut decision boundary regarding valid and invalid model matches is repeatedly maintained by the edge location only hypothesis tests.

In terms of the strength of the hypothesis test used to validate positive edge feature point samples, there appears to be no advantage (i.e. enhanced class separability) to be gained in using a confidence limit other than 1%. This is confirmed with regard to the ROC curves in Figure 7.2. The relative stability of the distributions across the 4 samples otherwise indicates generic uniformity in this regard, suggesting that distributions sampled with any intermediate confidence limits will appear similarly distributed. This homogeneity reaffirms the statistical integrity of the proposed metric.

In conclusion, there appears to be no advantage to be gained in using edge orientation information as an additional constraint to the projected model localisation and verification task. In practical terms, both metrics perform almost indistinguishably well for the given task. However, it was noted that on occasion, in cases where the initial prediction of model location and pose was sub-optimal, the edge location only metric was slightly more robust in terms of accurate projected model alignment. This is thought to be because the location only scheme is more likely to 'latch on' to any misaligned features during optimisation. Also, the aforementioned equal variance domain constraint (see Chapter 4 (4.7)) means that any weak edge features, coincidentally at the correct orientation, may be over-favoured and may thus distract the

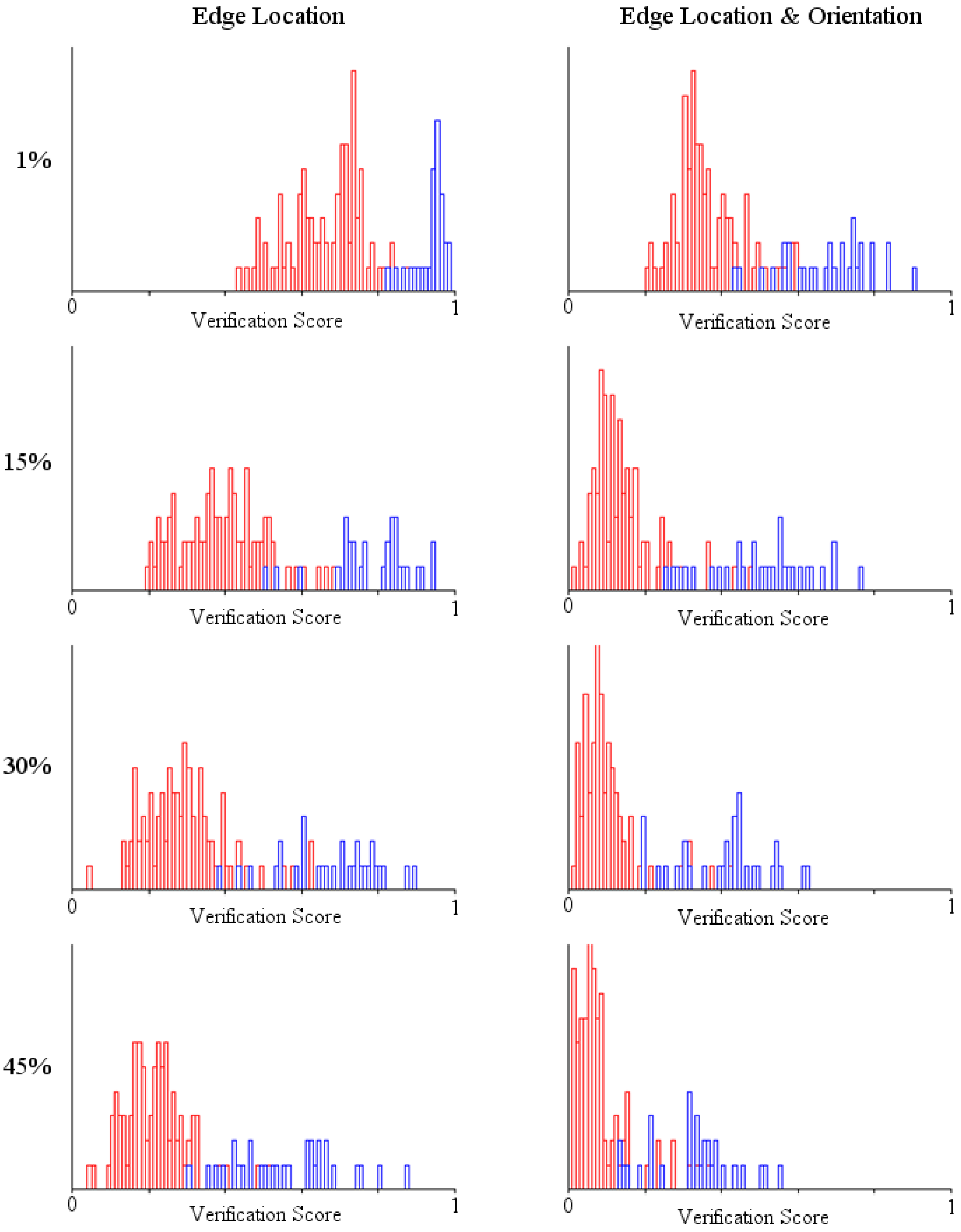
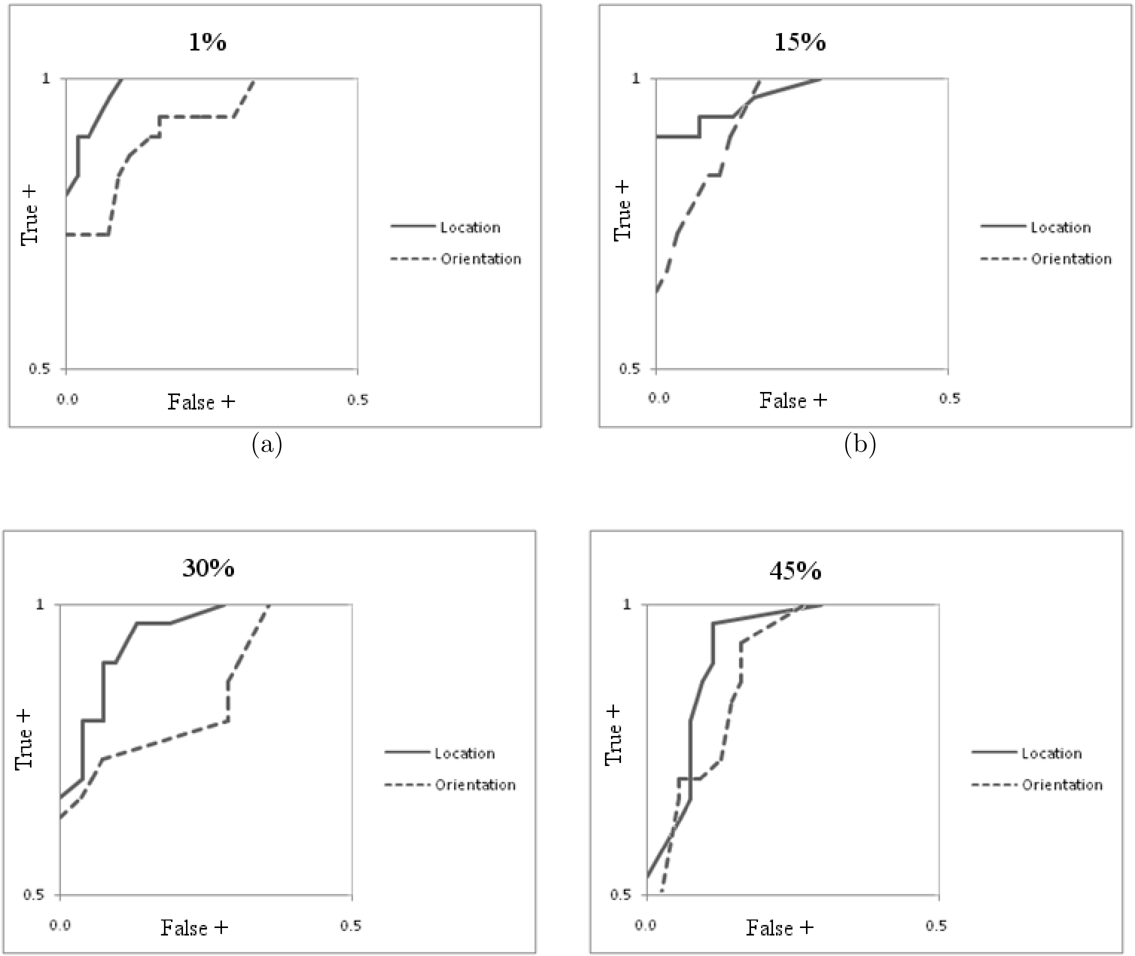


Figure 7.1: *Each entry in the histograms above represents the proportion (0 to 1 along the x-axis) of sampled edge feature points of a projected object model passing a statistical hypothesis test for projected edge verification (see Chapter 4 (4.9)). The* ***confidence level*** *of the test for the top row is* ***1%****, followed by* ***15%****,* ***30%*** *and* ***45%*** *for subsequent rows. The entries in blue represent valid instances of well-aligned view-based object models and those in red represent optimally aligned instances of random invalid object models. The left column of histograms represents hypothesis testing of edge location and the right column of combined edge location and corresponding orientation. The graphs indicate that there is no distinct representational advantage to using an additional edge orientation constraint or to using a confidence level other than 1%.*



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

Figure 7.2: *The four diagrams presented above are* ***ROC curves*** *corresponding to the histograms in Figure 7.1 (true positive rate along the y-axis and false-positive rate along the x-axis). Each diagram represents both edge location and combined edge location and orientation for* ***1%*** *(a),* ***15%*** *(b),* ***30%*** *(c) and* ***45%*** *(d)* ***confidence limits****. The (jagged) curves are sampled between the overlapping regions of the positive and negative distributions in each case. Optimal classification can clearly be observed for the edge only distribution (solid line) at a 1% confidence limit (a) in terms of the minimum area between the curve and the top left corner of the graph.*

alignment in awkward situations. It is suggested that the use of an edge orientation constraint may be more beneficial if dealing with extremely depleted image edge evidence or for distance-based approaches to feature correspondence. The bottom-line is that requiring a reasonably high degree of underlying feature correspondence along an extended edge feature requires any supporting edge evidence to be aligned, i.e. co-oriented, with the feature, thus obviating the need for a dual representation. Considering the enhanced class separability offered by the edge location only term, the simpler model accounting for just edge location is herein adopted as the basis of object localisation and verification processing.

Finally, a 1% confidence limit is shown to be most suitable for the verification hypothesis test in terms of edge location. This test limit ensures that all potential image edge evidence is accounted for. Having established the most appropriate basis for verification, the histograms in Figure 7.1 may also be used to set a positive match verification level. The relevant histogram (at the top left) suggests that if at least 90% of a projected model's features are validated at these levels, the match hypothesis can be almost certainly verified as being valid. Extended model search can therefore typically be cut-short when such levels of correspondence are experienced. However, associated experimentation highlighted an occasional problem with very sparse edge profiles for certain object views. In these circumstances, the simple templates may be well matched to features from other objects giving very high match scores for incorrect, coincidental model match solutions. Any obviously simple edge profiles should therefore be treated as being less reliable than any more complex counterparts. Reliable verification of very simple object views may require account of other visual information such as texture (i.e. material properties) or surface shading (if available). Such considerations will likely also benefit verification of heavily occluded objects, for which the proposed scheme will otherwise suffer from confusion problems in the randomly matched object noise field.

7.3.1 Quantitative Projected Edge Feature Verification Summary

- There is evidently no advantage to be gained in using an additional edge orientation constraint for projected edge feature localisation and verification under the proposed framework

- A 1% confidence limit on the lower bound of the edge strength hypothesis test sampled for each (sub-) pixel projected reference point is evidently best suited to discriminating valid and invalid model match hypotheses

7.4 View-Based Model Matching

7.4.1 Methods

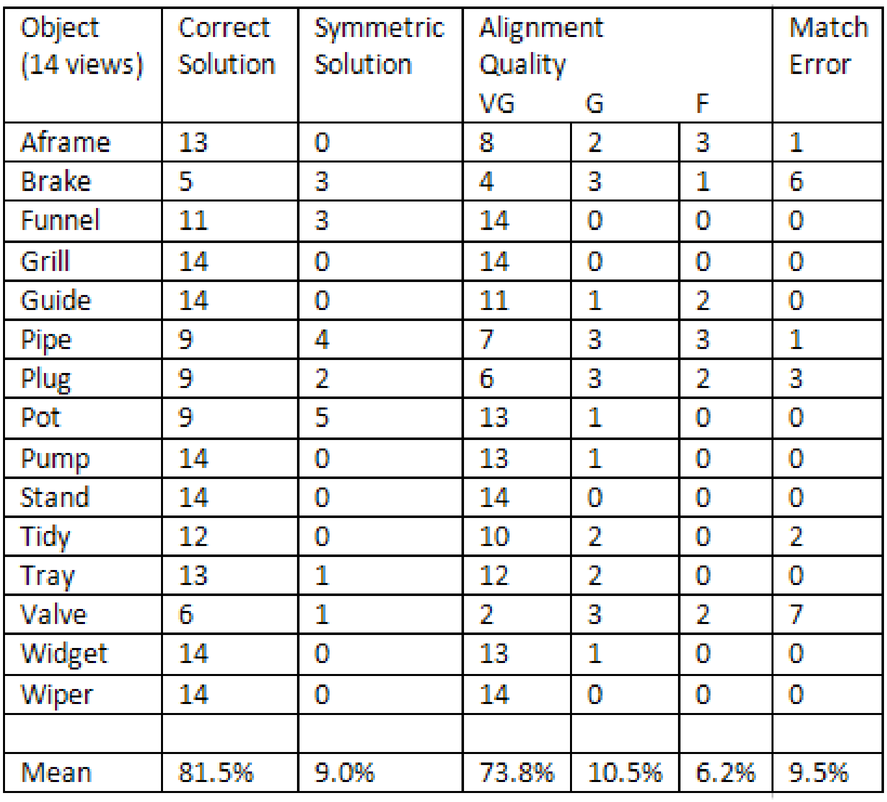
To evaluate the performance of the proposed view-based model matching system, a selection of 15 3D objects covering a diverse range of shapes and materials has been obtained. Because robotic handling and inspection equipment is unavailable for this project, for convenience, a 3D wireframe model has been manually constructed for each object. View-based feature visibility is approximated with reference to manually composed visibility files, which are currently uniformly sampled with 42 views around the view-sphere. Using these constructs as surrogates for the corresponding tangible objects, each object's range of PGH-based projected appearance has been automatically learned in accordance with the proposed methodology (see Chapter 6 (6.2.1)).

In order to approximate full view-sphere tests, for these experiments, each of the 15 test objects has been imaged in stereo by itself against a plain cloth background from 14 view-points approximating the 8 corners and 6 faces of an encompassing virtual cube. The right stereo image has then been sampled for each view for these experiments. The model matching system has then been presented with each image (see Section 7.2.1) and the performance has been observed in terms of whether or not the object has been detected at the correct pose along with an indication of the quality of model alignment. Alignment quality has been manually graded as very good (VG), good (G) or fair (F) with any poorly aligned model matches being classified as match errors. In line with the example parameters indicated in the model matching algorithm presented in Section 7.2.1, the longest 12 linear edge features (relative to a central viewpoint) from each of 42 views are matched to the longest 24 linear image edge features. A larger set of image edge features is sampled to account for potential interference from background artefacts (e.g. shadows) or extended line fragmentation. The results of these experiments are tabulated in Figure 7.3 and a selection of representative object images, corresponding Canny edge maps and overlaid model matches are presented in Figures 7.4-6.

7.4.2 Results & Discussion

The results presented in Table (Figure) 7.3 show that the proposed system performs well in the model matching task. Out of the 210 presented images, the system correctly identifies the presence and pose of the imaged object over 80% of the time. A further 9% of the time, the system alternatively identifies a symmetric view of the object. For instance, the coffee pot object is highly symmetric about a number of axes and for some is almost indistinguishable in terms of correspondence across the projected edge template. Because the search stops when a sufficient match verification score is met (0.9 in this case (see Section 7.3)), it may be that the correct solution is missed in these instances. It is proposed that recognition of highly symmetric objects, such as the funnel and pot, be followed by top-down pose inference processes, where correspondence with any distinctive view-disambiguating features is made. Otherwise, the system can be run until all image edge evidence has been examined without stopping the search prematurely when a suspected valid model match has been found.

With regard to the 9.5% of object views that were not matched, 7% of these emanated from the brake and valve objects. From the outset, the brake object was never going to work from the rear because there are very few edge-features that could be modelled. The rear of the object is effectively a free-form surface, for which only the outer circular rim and a handful of small central holes could be modelled. In selecting the longest 12 linear edge features from each view, all that was therefore modelled by the system was segments of the outer rim, which are almost totally uninformative. Similarly, the modelled side profile of the brake object was insufficient for recognition. The problem with the valve object was that there are very few prominent edge features and large amounts of interference from specular highlights and corroded surfaces as exemplified in Figure 7.7. In these circumstances it is difficult to select a feature that actually relates to physical structure in the first place and then any valid match scores are degraded on account of contextual normalisation of the image sampled PGHs. Furthermore, the multitude of spurious image edge elements means that many model PGHs will attain reasonably good match scores making it very difficult to differentiate any valid match hypotheses. One possible solution to this sort of problem would be to only model the most distinct and reliably detectable object



Object (14 views)	Correct Solution	Symmetric Solution	Alignment Quality			Match Error
			VG	G	F	
Aframe	13	0	8	2	3	1
Brake	5	3	4	3	1	6
Funnel	11	3	14	0	0	0
Grill	14	0	14	0	0	0
Guide	14	0	11	1	2	0
Pipe	9	4	7	3	3	1
Plug	9	2	6	3	2	3
Pot	9	5	13	1	0	0
Pump	14	0	13	1	0	0
Stand	14	0	14	0	0	0
Tidy	12	0	10	2	0	2
Tray	13	1	12	2	0	0
Valve	6	1	2	3	2	7
Widget	14	0	13	1	0	0
Wiper	14	0	14	0	0	0
Mean	81.5%	9.0%	73.8%	10.5%	6.2%	9.5%

Figure 7.3: *The table above presents the results of the view-based model matching experiments conducted across 14 views of each of the 15 3D objects in the test data set. Aside from the number of correct solutions automatically returned by the system, symmetric solutions refer to instances where the object has been located and validated at a symmetric pose. For correct or symmetric solutions, the quality of match is graded as being very-good (VG), good (G) or fair (F). The remaining instances where the system has failed to identify the object and its pose are classified as 'Match Errors'.*

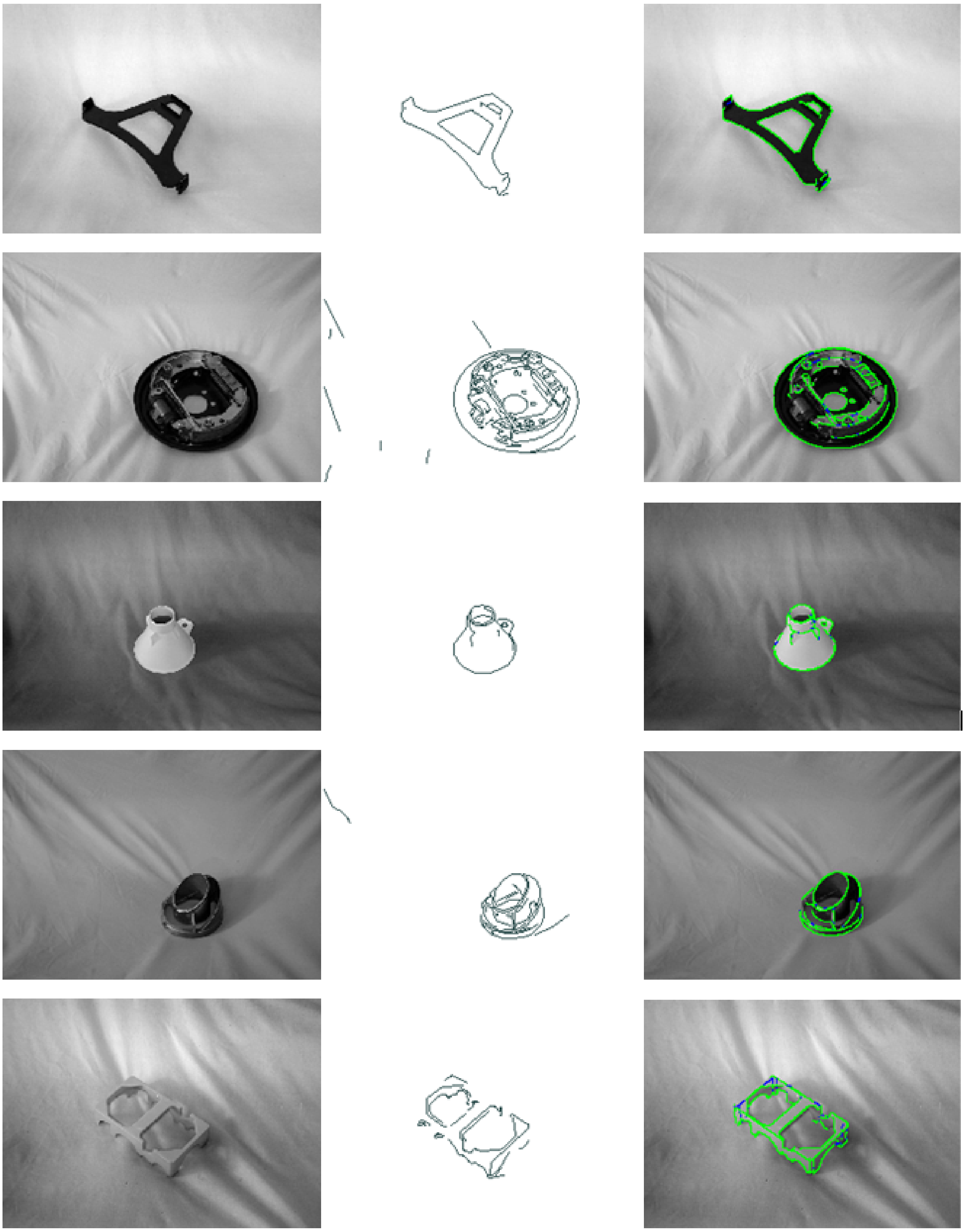


Figure 7.4: *The first of the 14 images sampled for each of the 15 test objects is presented across the following pages along with an adjacent Canny edge map. The accompanying rightmost images indicate optimised projections of the view-based 3D wireframe edge feature models (see Figure 4.5) correctly matched to the underlying data using the proposed view-based model matching system. Projected sample points passing a 1% hypothesis test for edge location are shaded in green and those failing in blue. The solution returns the precise 3D location and orientation of the imaged tangible objects in the robotic system's camera-based world reference frame.*

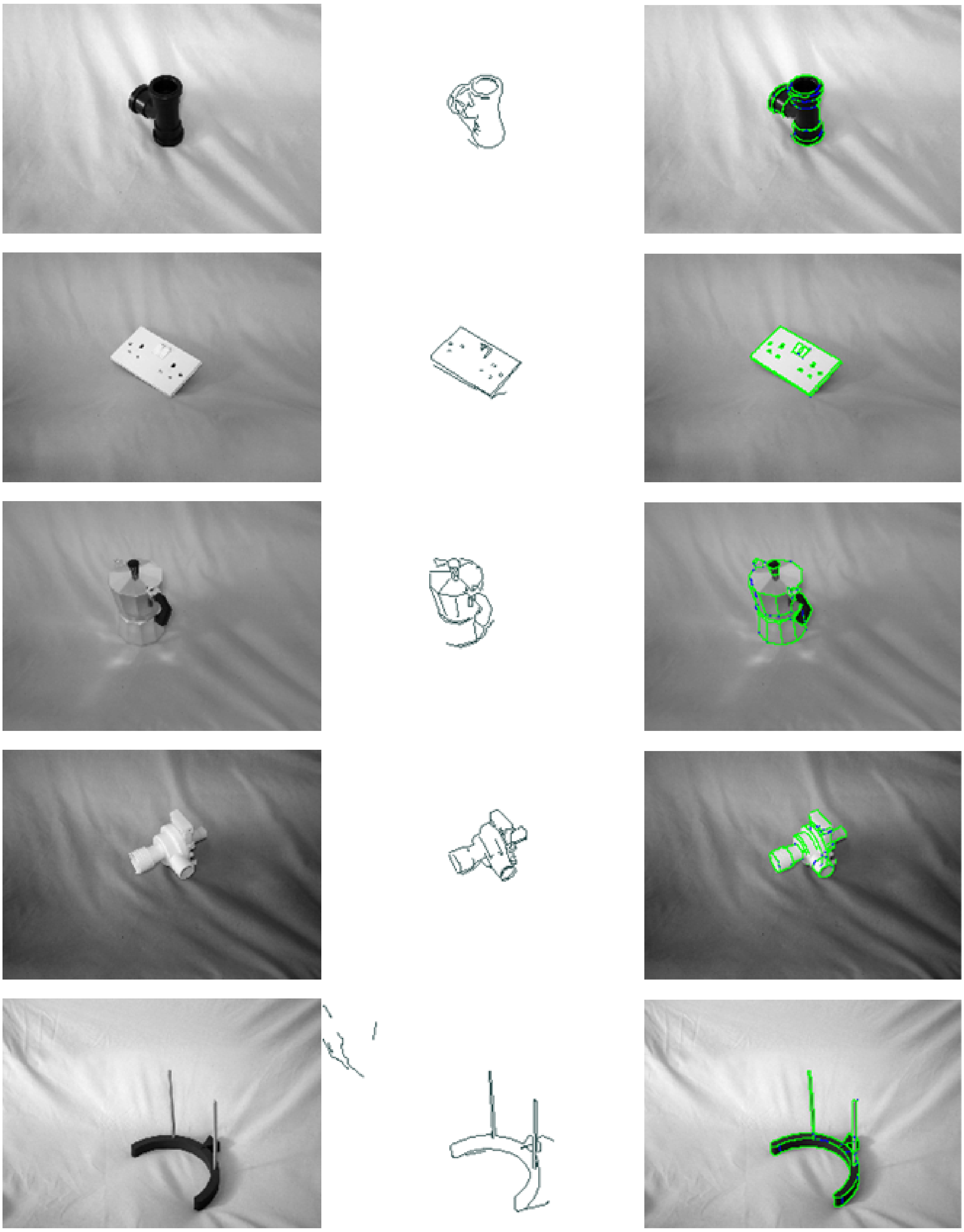


Figure 7.5: *Figure 7.4 continued.*

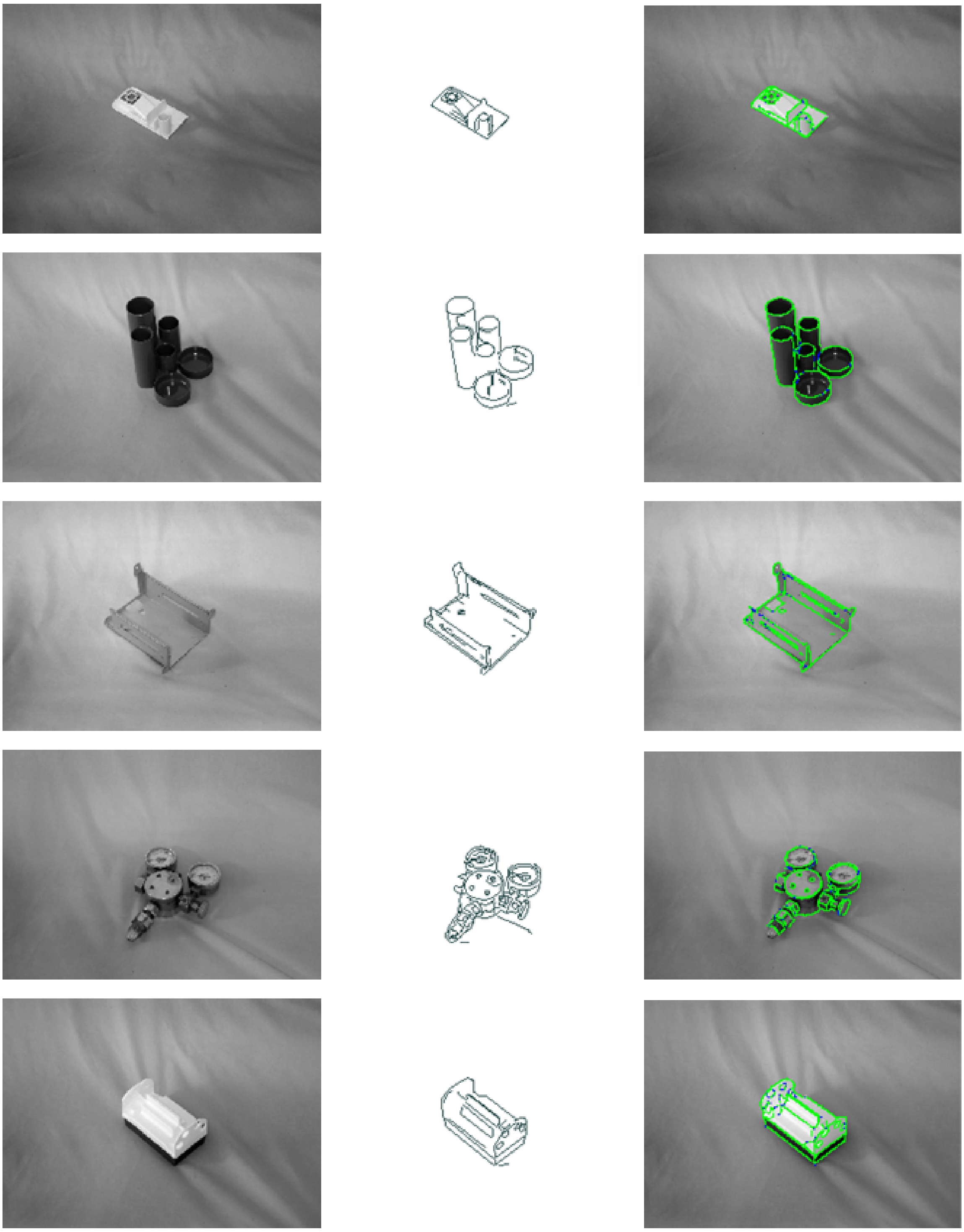


Figure 7.6: *Figure 7.4 continued.*

features such as boundary features and also to renormalise any image sampled PGHs to the bins occupied by the stored reference PGHs. This latter option should allow for any distracting image elements such as specular highlights or clutter to be effectively ignored, thus making recognition much more robust. If using a sparser set of defining features, it may also be beneficial to extract edge maps at lower scales, so that only the most significant edge features are sampled.

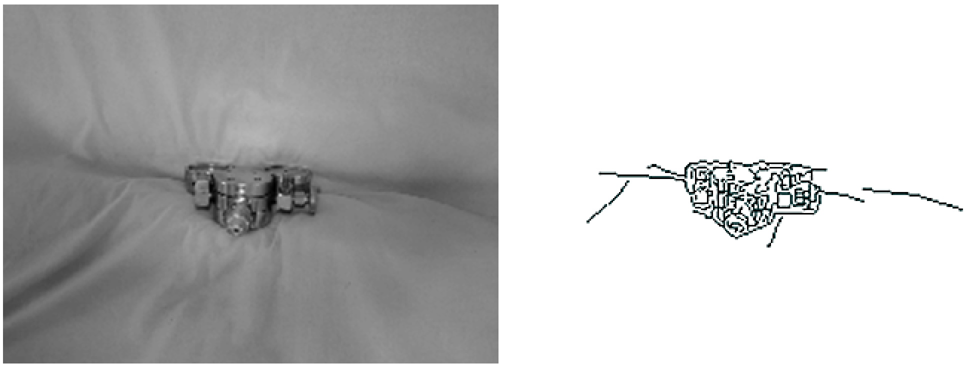


Figure 7.7: *Particular model matching problems were found for the reflective and corroded valve object, pictured above, for which the sampled Canny edge maps, as indicated, were typically too noisy to support reliable recognition.*

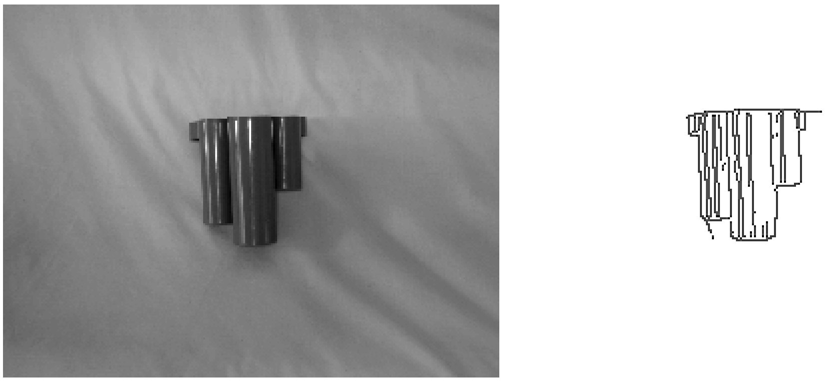


Figure 7.8: *The pictured (desk-) tidy object was prone to disruptive interference from specular highlights, as indicated in the corresponding Canny edge map.*

Model matching of the valve object may however have been better if longer was spent running the model matcher. For these experiments, a maximum of 3 iterations (i.e. 24 image features) were performed for each object view, with it typically taking just under a minute for each plus any verification processing overheads. For most other objects, recognition was achieved with a single iteration, although the possibility of arbitrary interference from background edge features meant that there was little point quoting this figure for each match. Similarly, most correct matches emanated from the best supported match lists, but because so few image features were sampled amongst various spurious image edge features, direct quotation of these figures may be misleading in cases where recognition was supported by a single valid image feature.

Similar problems to those observed for the valve object were observed for the pipe and tidy objects. For the pipe object, although every single feature was modelled, only a few were ever detected by the image edge detection software because of a lack of contrast. The logarithmic sensitivity to illumination exhibited by the human retina would otherwise presumably improve the scope of any such methods. In contrast, the glossy nature of the plastic surface of the pipe meant that for some views a number of distracting specular highlights appeared. The culmination of these effects was that the object's projected edge profile rarely matched that which was expected. The same was true for certain views of the desk-tidy object, which was particularly prone to specular interference (see Figure 7.8). The simple nature of these shapes combined with high levels of specularity meant that many well supported edge feature matches arose for erroneous image edge structures. In such circumstances, the correct solution will often have been identified, but it will effectively be swamped in the noise field. Short of loosening the constraints on selection of suspected valid model matches, which would lead to a massive overhead in verification costs, a strategy of only modelling the most distinct object features such as boundary features and normalising any image edge PGHs to the positive entries of the stored sample PGHs again seems to be a sensible strategy.

The only observed case where an error arose due to inadequacies of the supporting model alignment software was for a view of the tidy object. Using a slightly higher minimum precision level across the learned manifold would otherwise likely improve the accuracy of the initial prediction of model pose, which would make it more likely that the correct pose was subsequently optimally aligned. The only other object with significant model matching problems was the plug object. The main problem here was with the rear of the object, where there were very many edge features which although modelled were not reliably detected as edge features. Retrospectively, a much simpler wireframe model representing only those edge features that are regularly observed would have been much more effective for model matching. Similarly, the simple wireframe model was inadequate as viewed from the side.

Aside from the instances where the system fails, there are around 10 objects for which the system works almost invariably exceptionally well. As long as some reasonably prominent and distinctive edge structure is present, the system is evidently able to efficiently decipher the correct solution regardless of 3D view-point. For many objects, much fewer than the 12 lines currently sampled for each view may also be used as the basis of model matching, thus speeding up the process significantly.

7.5 Bhattacharyya Match Score Distribution Analysis

Following on from the success of the view-based model matching experiments across the object data set, on each occasion where an object was correctly matched to an image, the Bhattacharyya match scores for a sample of edge feature matches leading to a valid solution were recorded in order to analyse the distribution of such data. Accompanying 400 positive match scores sampled in this manner, a negative distribution was formulated by sampling the best match score for each of the longest 15 linear edge features sampled for each of 42 views for 3 random objects for 3 random linear edge features for a random imaged instance of each of the 15 objects in the test data set. The positive and negative distributions are presented overlaid in Figure 7.9. As can be seen, the two distributions are well separated, thus supporting the applicability of the representation for the model matching and object recognition tasks. Although there is a small, almost negligible degree of overlap between the two distributions, this would be expected with any representation in cases of shared appearance between elements of certain object views. The histograms suggest that for unoccluded objects, any image features with Bhattacharyya match scores less than 0.7 should be discounted from further processing.

7.6 The TINA Stereo-Vision-Based 3D Model Matcher (3DMM)

The TINA stereo model matching computer vision system (3DMM) was originally developed in the late 1980s [19], offering a solution to various vision guided robotics

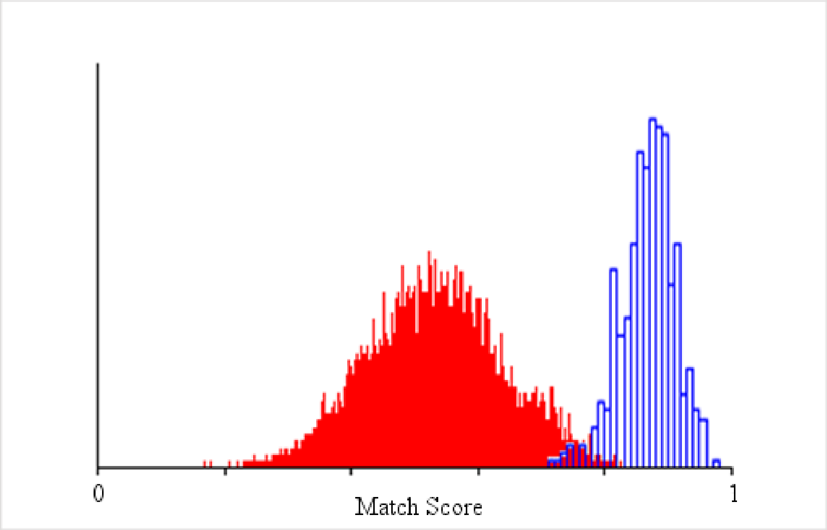


Figure 7.9: *The histogram above represents valid (blue) and invalid (red) PGH-based Bhattacharyya match score distributions sampled for corresponding edge features.*

applications. The system has since been refined in various ways, as described in document [79] (2001), which details the previous state of the art (see Figure 7.10) . A number of more subtle updates to the system's functionality have however been made within the TINA group during the course of this project (see TINA [1] Change Log: 91 and 97). In essence, the system serves to locate known 3D objects in scenes with sufficient accuracy to facilitate robotic eye-hand-object interaction. As with the proposed 3D object recognition system, the 3DMM is concerned with the analysis of edge information. The system is also therefore currently applicable for rigid objects, such as industrial components, whose shapes are defined by edge features.

The 3DMM operates by forming an edge-based depth map of a scene from a pair of camera calibrated stereo images. A calibration tile is used for calibration in current research, although the proposed methods support ongoing calibration via matching of accurately modelled 3D objects. An implementation of the Canny edge detector [14] is used to extract images' significant edge features before a 'stretch correlation' algorithm is applied for stereo reconstruction. The stretch correlation approach to stereo reconstruction is detailed in a previous associated publication [43]. Given an edge-based depth map representing a scene, an algorithm is employed to convert any such edge features to geometrical primitives such as lines and ellipses. Objects' representative wireframe feature models are (currently) manually constructed from these same geometrical features, enabling 3D edge-feature-based object models to be matched to 3D edge-feature-based scene reconstructions.

Object model to scene matching is performed by searching for sets of three-dimensionally distributed edge features with the expected geometric configurations. More specifically, tables are initially created for both model and scene geometries, detailing the relative 3D distances and orientations of each feature within each set. Because representative object models may potentially incorporate hundreds of features, matching is initially based on reduced sets of features (e.g. 10 at a time), so as to avoid combinatorial search problems. Prior to this work, each 3D wireframe was supplemented with manually composed key feature sets that were deemed likely to be visible together through local ranges of view-point. A further reduced set of the most important features is specified as a focus feature set.

Model matching proceeds by finding any potential matches for each specified focus feature with the scene geometry, thus forming a potential match seed list. For each entry on the seed list, an exhaustive search is made between the remaining features in the (reduced-) model and (full-) scene geometry sets to find any other coinciding features. This is performed by checking whether each potential feature pairing is of a compatible type and if so analysing whether the pairwise 3D geometrical relationship between the features accords with each entry already on the match list with reference to information embedded in the pre-defined match tables and associated thresholds. The initial match lists may contain many matches for each focus feature. A subsequent stage of processing then extracts any unique combinations of features. For each such feature set with a prescribed minimum number of features (e.g. 3), a further search is conducted over the remaining features in the full wireframe model, with each match list being extended accordingly. Each matching pair of features is weighted by length, with any cumulative match lists being ranked by the sum of such weights. Although this description indicates the essence of the processing involved in the 3DMM, the system is very complex and cannot be described in detail in such a short space. The interested reader is referred to the source code [1] for a more precise account of the processes involved.

Upon matching a model to a scene, the process is completed with a ‘closed loop validation stage’ which iteratively tests the generated hypothesis of object location and orientation against the original image data whilst optimising the calibration of any camera parameters. This process is equivalent to that used by the view-based model matching system and is crucial in order to maintain calibration consistency in a working system whilst eliminating problems with a lack of consistency in the coordinates generated by some forms of camera.

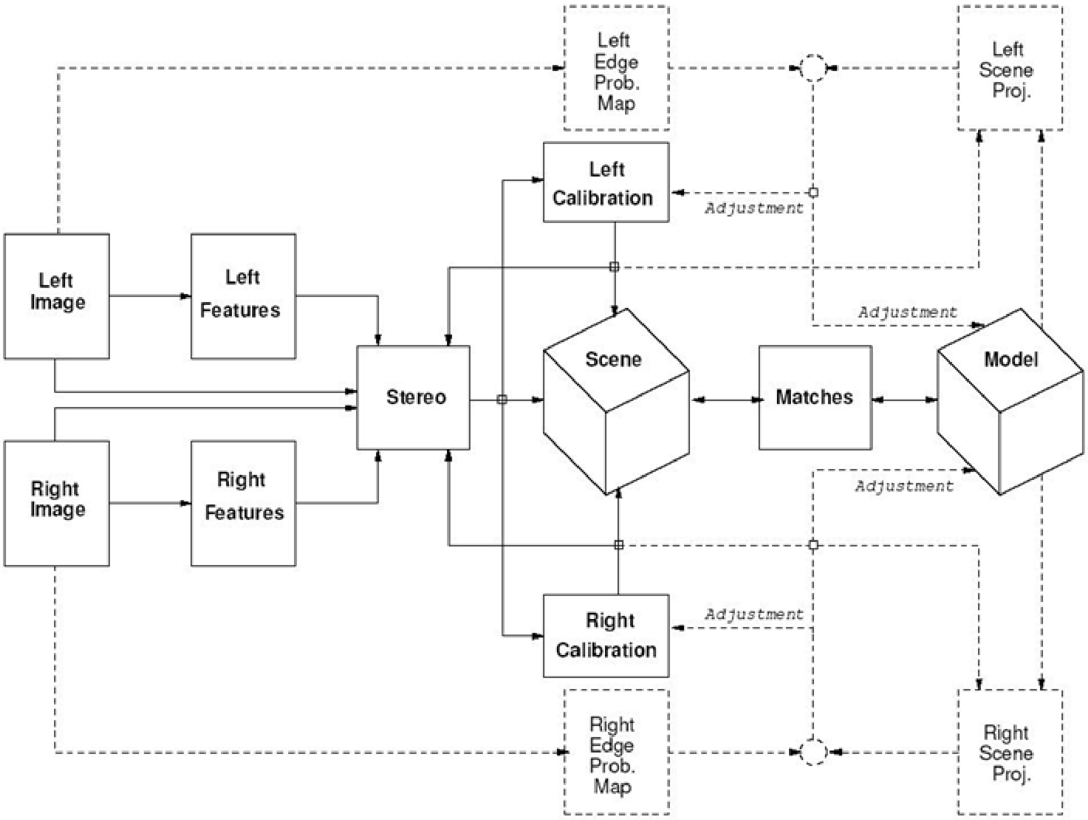


Figure 7.10: *The above diagram indicates the flow of information throughout the 3DMM.*

7.7 Amendments to the 3DMM

A number of amendments have been made to the stereo vision-based 3DMM during the course of this project. The most significant development is the incorporation of a hypothesis verification stage which is used to rank any competing match hypotheses. Previously, any resulting match lists were ordered in terms of a sum weighting of the lengths of the matched edge features. This metric does not account for the quality of the match between features, but instead simply accepts features as being matched or not. Because of the abstraction and potential for geometrical distortion introduced to the representation through the medium of stereo, this is perhaps a prerequisite of any such approach. This however means that match lists may not be reliably ordered, with a high probability of an incorrect match being classified as the best, thus limiting the utility of the approach for comparative object recognition. To assess the quality of any match hypotheses, it is instead proposed that verification be based directly on the underlying image evidence under a projection of the view-specific features of a matched wireframe. This model accords exactly with that outlined for use with the view-based localisation and verification procedures as outlined in Chapter 4.

The nature of stereo vision-based 3D model matching precludes use of view-point dependent features such as occluding boundaries. Although corresponding view-dependent features in each stereo camera frame may be highly correlated, because each camera senses a shifted physical edge region, stereo reconstructions cannot be reliably made. Furthermore, even if a corresponding extended edge feature could be reasonably reliably inferred, there is no corresponding model edge feature to be matched to as the single edge would only effectively be an infinitesimal partition of a continuous surface region. Although many objects may still have enough fixed edge features to facilitate stereo model matching, the post-match process of verification has no such restriction and may be supported with view-point dependent feature visibility constraints. Occluding boundaries have therefore been modelled in current research to support more informed model match verification, as detailed in Chapter 4. This shortcoming of the stereo model matching process otherwise supports application of alternate view-based methodologies, such as the proposed PGH-based one.

In previous work, stereo feature correspondence search was conducted over manually composed, reduced sets of model features that were deemed likely to co-occur. Further reduced sets of the most reliable focus' features were specified, upon which to base the matching procedure. In relation to application of the same wireframe models to the proposed view-based recognition system, objects' view-spheres are divided into 42 approximately uniform regions, with a set of mutually visible features being assigned to each view region (see Chapter 6). These sets of view-specific features can therefore be sampled as the basis of model matching. Because certain object views may still contain too many features to make search tractable in a reasonable amount of time, a further reduction process is required. To this end, it is proposed that the longest features in each view set be preferentially selected as the basis of model matching. This process could ultimately be supplemented by a learning feedback loop, whereby any features that are consistently reliably or uniquely detected through repeated exposure of an object are favoured. Features exhibiting high Fisher information scores for localisation could also be preferentially selected in this regard. Although the longest features comprising some objects' shapes may be unevenly distributed, short of using all features, as long as a reasonable number of features are used this rule should be sufficient to support generic model matching.

To further enhance the practicality of the 3DMM, a number of constraints have also been imposed on any resulting match list entries to avoid any unnecessary or duplicated verification procedures. Typically, especially for cluttered environments or objects with repeated structures, there may be very many competing entries on each match list, meaning that image-based verification across the list may be very expensive. In the current implementation, each hypothesis is only kept if it represents a unique transformation within predefined location and orientation constraints. Furthermore, because the matching process is now based upon view-point specific sets of features, match hypotheses are rejected if the hypothesised view-point significantly differs from that relating to the sampled set of features. Although the previous match quality metric has proven unreliable for the task, it still may be used as a rough guide, so that any obvious relatively deficient match lists can be cheaply disregarded.

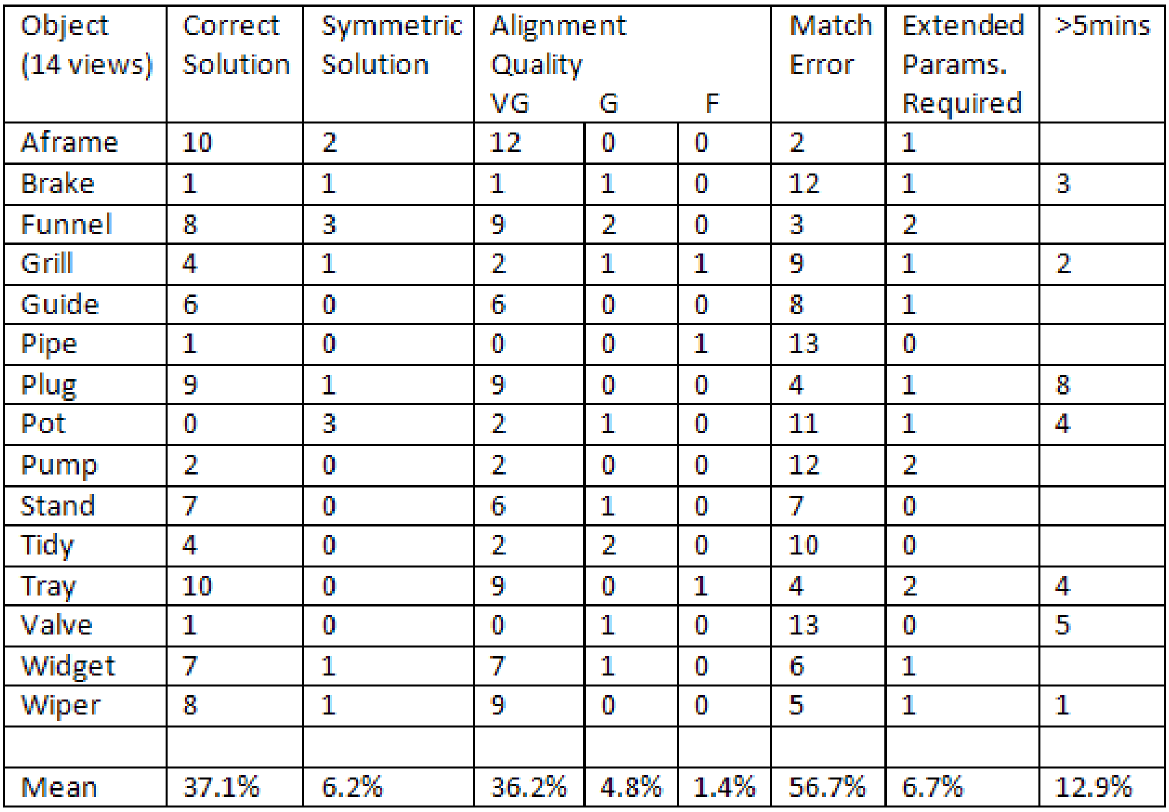
7.8 Stereo-Based Model Matching

7.8.1 Methods

To assess the performance of the stereo-based 3DMM, the same set of 15 3D objects as utilised for the view-based experiments have been used. In this case, the stereo images are used for each of the 14 sample views instead of the single right-hand images. Specifically, the stereo cameras are positioned approximately 0.2 metres apart, with the stereo rig positioned approximately 1.0 metre away from the focal point of the supporting test bed. Otherwise, the experiments are exactly the same as for the view-based model matcher; the aim being to assess the utility of the referenced stereo model matching system.

In terms of system parameterisation, the model matcher is set to work with the same 42 views as used for the view-based system. For each view, a selection of the longest features is used as the basis of model matching. Specifically, the system is set to initially use the longest eight features from each view, with the longest 5 being sampled as focus features. The minimum prescribed match list length is set at 3. In cases where these settings proved ineffective for model matching, the matching process was repeated with the focus feature set size being incrementally increased up to the longest 12 features until matching was successful.

7.8.2 Results & Discussion



Object (14 views)	Correct Solution	Symmetric Solution	Alignment Quality			Match Error	Extended Params. Required	>5mins
			VG	G	F			
Aframe	10	2	12	0	0	2	1	
Brake	1	1	1	1	0	12	1	3
Funnel	8	3	9	2	0	3	2	
Grill	4	1	2	1	1	9	1	2
Guide	6	0	6	0	0	8	1	
Pipe	1	0	0	0	1	13	0	
Plug	9	1	9	0	0	4	1	8
Pot	0	3	2	1	0	11	1	4
Pump	2	0	2	0	0	12	2	
Stand	7	0	6	1	0	7	0	
Tidy	4	0	2	2	0	10	0	
Tray	10	0	9	0	1	4	2	4
Valve	1	0	0	1	0	13	0	5
Widget	7	1	7	1	0	6	1	
Wiper	8	1	9	0	0	5	1	1
Mean	37.1%	6.2%	36.2%	4.8%	1.4%	56.7%	6.7%	12.9%

Figure 7.11: *The table above indicates the success of the TINA stereo-based model matching system across 14 views of each of 15 objects. The format is similar to that in Table (Figure) 7.3, except that extra table entries are included to indicate occasions where the match parameters (focus and cliche feature group sizes) required extension and also where the system took over 5 minutes to operate in instances where matching was affected by combinatorial search problems.*

The results of the stereo model matching task presented in Table (Figure) 7.11 indicate significant shortcomings with the approach. As indicated, the system fails altogether nearly 60% of the time. On other occasions, the system can get trapped in combinatorial searches, taking, in some cases well over 10 minutes to come to a solution. This is not to say that the system is completely useless; in cases where the object geometry is reasonably sparse and well defined, the system is able to very quickly arrive at an accurate solution. Furthermore, the stereo method solves for scale, whereas the view-based methods require analysis across multiple scale intervals, with only a single scale interval being sampled as the basis of the presented experiments. Aside from potential combinatorial search issues, perhaps the main problem with such approaches is generation of quality 3D data in the first place.

As observed for the single view-based analysis in the previous section, the object detection and localisation task is complicated by the invariant imperfect nature of objects' detected edge profiles. While providing enough of a problem for single view-based recognition processes, the problems are compounded when requiring features to be present in both stereo images. Furthermore, because of the separation of the 2 cameras in a stereo rig, it is also possible that features visible from one projection may be self-occluded in the other, rendering any such features unusable for matching. Another recurring problem observed across the test data set relates to horizontal image lines. Because stereo is based upon feature correspondence sampled across epipolar lines, any image lines lying parallel to the epipolar lines will be excluded from processing. Another subtle problem relates to the requirement of a process of lateral feature shifting, as identified in Chapter 4 (4.4). Many structural edge features are slightly curved in nature, meaning that an observed edge feature may shift in relative position depending on the relative illumination. Because stereo samples 2 displaced views, there is the possibility that an observed edge feature (e.g. an extended specular highlight) may be slightly displaced in each view, thus impairing the accuracy of stereo reconstruction for such features.

Perhaps the most obvious problems with stereo vision approaches to model matching and object recognition relate to view-based features. Stereo is inadequate for accurately detecting view-based features because a different physical region will be observed by each separated view-point, thus rendering stereo processing invalid. Furthermore, as mentioned previously, even if the contour of an object could be approximately inferred from stereo cues, this would only represent an infinitesimal partition of an extended surface region raising obvious complications in terms of 3D matching. In order to recognise a view-based feature, a view-based processing paradigm must be used. For many objects, including some from the sampled data set, whose projected appearances comprise significant numbers of view-based features, stereo is therefore an impoverished medium through which to perform model matching and object recognition.

Another minor issue regarding system performance relates to the use of geometrical primitives as fitted to the 3D edge feature geometry as the basis of model matching. Currently, edge features are classified as being linear or elliptical segments, with the final stereo geometry being constructed from these primitives. This raises a potential problem when features are wrongly classified and may thus be falsely discounted from the matching process. A marginal improvement in system performance is likely to arise if instead only a single feature type were used, i.e. curve segments, so that misclassifications would be avoided. In this case, a line feature could simply be modelled as a segment of a very broad curve. The loose constraints imposed on matching 3D wireframe features to those sampled from the error-prone stereo field also renders the system ill-suited as the basis of comparative 3D object recognition. This is because the statistics used for individual feature comparison are not sensitive enough to support comparison of competing match hypotheses with similar numbers of features. Match hypotheses are instead verified by virtue of correspondence with the underlying edge data in accordance with the methodologies outlined in Chapter 4.

Other problems with stereo analysis include drifting problems between sample calibration-tile oriented reference points between sets of images. On a couple of occasions, the low-budget stereo rig can be seen to drift slightly between successive tile-calibrated sets of object images, meaning that the stereo geometry became distorted. It was however observed that such factors would have made little if any difference to the overall model matching results on account of interference from the more significant aforementioned debilitating factors. Prospectively, for active stereo vision systems, where the cameras can move independently, live calibration can be performed via optimisation of matching parameters between independent camera frames for projected 3D model features using the existing routines for localised projected model alignment (Chapter 4).

Although the stereo system can be shown to work very efficiently for sparse, well-defined 3D geometry, testing of the system highlighted significant issues relating to combinatorial search. As discussed, model matching is based around reduced sets of focus features so as to avoid combinatorial search problems amongst large object and scene feature sets. Otherwise, computational requirements expand exponentially in proportion to feature set size, meaning that search soon becomes intractable for moderately sized feature sets. Being forced to limit the number of features required as the basis of model matching however presents problems for scenarios where an object pose may convey many features, of which many are likely to be missing due to unfavourable illumination conditions. In these circumstances, large numbers of focus features may be required as the basis of model matching. Table (Figure) 7.11 indicates a number of instances where extended focus feature and initial-search-group sets were required to support model matching, accompanying other instances where search took over 5 minutes (sometimes over 10) on account of the above-mentioned combinatorial search problems.

The minimum number of features agreeing on an object's pose, so called 'clique' sizes in the current context, also presents problems for model matching. Using a high clique size is useful to exclude any random irrelevant matching feature sets, but in many cases, especially for simpler objects, only a couple or so of object-modelled features may be present in a scene. If a clique size as low as 2 is utilised to account for poorly supported object-scene geometry, it is likely that an inordinate number of false positive matches will arise for other more complicated structures, especially in noisy scenes, leading to computational confusion. It is therefore very difficult, if not impossible, to uniformly parameterise such a model matching system across an object database. The object specific nature of the model matching process does however mean that focus feature type set sizes can be tailored to searches for specific objects, although problems remain for analysis across cluttered, feature rich scenes.

The complexity problems just outlined are further compounded by occurrences of repeated structure across model and scene geometries. In these circumstances, a combinatorial explosion in the number of possible ways an objects' 3D features may align with the scene geometry may arise. The plug and tray objects, for instance, contain many small cliques of features that may be equally well matched against a number of different object parts in various different poses. If a number of object features are parallel and positioned closely together, thus allowing all other matching object features to be similarly matched for each feature, the permutations of the number of ways in which the 3D geometry could be matched again increases exponentially; further exhausting processing.

7.9 Conclusions

This chapter has reviewed the performance of the proposed TINA view-based 3D object model matching system relative to that of the pre-existing, albeit now modified, stereo-based model matching system. The presented experimental results suggest that the view-based model matching system is able to autonomously learn a continuous, partially-scale invariant representation for a presented 3D object and use that representation as the basis of 3D model matching and object localisation. Aside from a number of instances where problems with the basic wireframe models made model matching impossible, the system records near perfect performance across wide numbers of views under essentially arbitrary illumination conditions using just the longest 12 features from each of 42 approximately uniformly sampled view regions.

In contrast to the view-based model matching findings, the full-view-sphere sampled nature of the object test images meant that a number of limitations of the stereo-based model matching process transpired. The system made 6 times as many matching errors in the model matching task as the view-based system, failing nearly 60% of the time, thus proving a poor medium through which to conduct 3D model matching. The main problems with stereo were with the quality of the stereo data, which cannot be reliably inferred for many feature types in arbitrary scenes. 3D model matching routines also suffer from combinatorial search problems, whereas view-based PGHs essentially encode extended regions of geometry, without having to independently search for them first in the depths of the 3D noise field, meaning that search progresses linearly over expanded feature sets. The statistics required as the basis of 3D model feature matching are also inadequate for comparative object recognition, meaning that more exhaustive post-match verification processing is required to compensate (see Chapter 4).

With a working system for view-based model matching in place, the chapter reviewed the validity of the proposed projected edge feature localisation and verification processes. The aim was to quantify the advantages to be gained in using a dual edge location and orientation metric for object localisation optimisation and verification. The associated experiments indicated that there was no distinct advantage to be gained in terms of valid and invalid class separability if using the additional orientation constraint for verification. The concluding reasoning was that if a high level of correspondence is required along an extended edge feature, this will in itself mean that the corresponding edge features are co-aligned, with the extra orientation constraint otherwise contributing to the noise floor. Object localisation also proved to be slightly more robust in awkward situations for the edge location only metric.

Having discounted stereo vision-based processing from the model matching task in favour of the more reliable and powerful view-based methodologies, the remaining issue of object recognition and application of the proposed techniques to scenes containing clutter and occlusion is discussed in the following chapter. The proposed techniques also support object tracking through video streams, for which objects' locations can be iteratively updated in successive frames using the proposed location optimisation functionality following initial object detection. The more expensive object detection routines can then be reserved for surrounding image features or global image analysis if the tracked object is lost.

Chapter 8

3D Object Recognition Using Pairwise Geometric Histograms

8.1 Introduction

Although the development of a fully functional 3D object recognition system has stood as the main motivation for this Ph.D. project, time limitations and work on supporting aspects of the problem have left less time for associated research than would ideally be available. From the outset, this final chapter therefore serves as an introduction to the problem of 3D object recognition. As will be shown, the problem is heavily constrained by the time complexities associated with such view-based processing across large object datasets with standard sequential computing hardware.

Having established the suitability of the Pairwise Geometric Histogram (PGH) representation for the specific-object model matching task, this chapter investigates extension of the proposed techniques for the task of 3D object recognition. Because the model matching process detailed in the previous section is based around construction of well-supported, view-coherent match lists, the task of object recognition is a natural extension to the model matching task. In the first instance, object recognition performance is evaluated with reference to the dataset of singularly presented, unoccluded objects as used for the model matching experiments in the preceding chapter.

Previous associated research [46][44] has proved the viability of the PGH representation for the task of 2D edge-based object recognition through difficult viewing conditions including clutter and occlusion. For this former 2D recognition work, all image edge data is concurrently analysed and objects' location, pose and scale parameters are determined via peak identification in specific-object Hough transform spaces. Because the Hough transform process is computationally expensive, the process is only run for objects receiving a high level of support across a number of PGHs. In attempting to transfer the same techniques to the 3D object recognition task, a problem is however presented in terms of the amount of data that requires analysis with regard to the vast number of 2D views and features required.

8.2 3D Object Recognition using Pairwise Geometric Histograms

In the previous chapter, methodologies were proposed for efficiently recognising and localising imaged instances of specific 3D objects. The method proposed for this purpose was to formulate potential match lists between reduced sets of the longest model and image (linear) edge features agreeing on the pose and scale of the sought object. A subsequent verification stage was utilised to evaluate the validity of any match hypotheses in terms of sampled edge point correspondence across the optimised projected position of the hypothesised object model in the image. In order to recognise an object in an image as being an instance of any learned object, the same process as used for the model matching task can be used with search instead being conducted over all known object models. Any valid instances of matched objects should return better supported match list hypotheses, allowing for the correct solution to be differentially selected from the noise field.

Previously, for specific-object model matching, intra-object match lists were formulated with regard to accord with hypothesised object view-point and scale. Although an additional image-plane orientation constraint may be applied to each match list (0 to 2π), this proved unnecessary for model matching, instead representing a slight computational burden. To enhance the acuity of the proposed object recognition system, the additional constraint of image-plane orientation coherence amongst match list entries is introduced. A 30 degree canonical image orientation constraint is therefore imposed on match list construction on account of any prospective errors in the matching and 3D pose determination procedures. As discussed in Chapter 3 (3.1), the PGH representation is partially invariant to displacement along the reference line to account for possible edge fragmentation. This makes it difficult to impose an absolute location constraint (e.g. a centroid) on any match hypotheses without further assessment of model to scene feature correspondence. For this reason and because only a single instance of each object is currently available, the match hypothesis constraint on image location is excluded from current processing. The remaining constraints used for match list construction are otherwise expected to limit the possibility of multiple spurious instances of the same object pose occurring across such scenes. The analysis of multiple instances of recognised objects in scenes otherwise naturally follows as multiple peak detection across an image.

With further regard to the model matching system, match list entries were treated as single votes, normalised by the number of 'confused' model-feature match entries with a similar match score. The idea here was that any distinct features should contribute more significance to the model matching task and any confused model features should be more indicative of broad-scale interference from irrelevant feature matches in the confused object noise field. This assumes that the objects are not occluded, as missing features will lower any valid match scores into the irrelevant feature match noise field distribution (see Figure 7.9). Although this methodology proved viable for the specific-model matching task, subsequent analysis with regard to the more demanding object recognition task highlighted occasions where correspondence across repeated object structures (e.g. multiple correspondences across a symmetric object such as the (coffee-) pot) would adversely down-weight any such perfectly valid feature match hypotheses. In order to maintain an optimal degree of recognition acuity, this vote weighting mechanism has been abandoned for object recognition.

In basing object recognition on reduced sets of model and image edge features, there is an increased possibility that certain views of other objects may be confused with valid solutions; especially across cluttered image edge feature distributions. It is therefore proposed that each match list entry be weighted by the observed Bhattacharyya score for that feature. Any coincidental, reasonably well supported model match hypotheses should therefore be down-weighted on account of expected reduced levels of support. Ideally, such object match support metrics would be formulated to quantitatively account for the proportion of edge features detected across a hypothesised object view. Although such information is encoded by the weight of un-normalised PGH bin entries, with reference to Chapter 5, normalised PGHs are required in order to interpolate scale between consecutively scaled image geometry samples. In the first instance, since we are dealing with incomplete feature sets, the cumulative normalised match score weighted metric should be sufficient to indicate any well-supported valid object matches.

8.2.1 Summary of the Proposed View-Based 3D Object Recognition and Model Matching Routine

- The 3D model matching routine outlined in Section 7.2 is extended so that each of the 15 objects in the dataset is model matched to the image data with verification being deferred till match lists have been constructed for each object
- The highest ranked match lists are returned as object recognition hypotheses, with the same optimised localisation and verification routines as described for the model matching routine being used to find the best supported object recognition hypothesis (or hypotheses)
- A minimum of 2 features agreeing on the pose and scale of the object are required for match list construction to expedite recognition with the current system

8.3 Single Object Recognition

8.3.1 Methods

Before addressing the more demanding issues of object recognition in cluttered and occluded scenes, the utility of the proposed approach to object recognition is initially assessed by attempting to recognise singularly presented instances of each of the 15 objects in the sample dataset. The aim is to establish that the proposed scheme is adequate for this base-case task before challenging the system with more difficult viewing scenarios. For these experiments, for each of the 15 objects, 2 random, essentially different images were sampled from those used in the previous chapter's experiments and match lists were formulated for each of the 15 objects for each image. Any images for which the model matcher formerly failed were otherwise excluded from selection. As with the model matching experiments, reduced numbers of the longest model and scene features were used to make the scheme more temporally viable on current hardware. Specifically, the longest 12 projected features from each of each model's 42 views were matched against the longest 24 image-sampled lines, with this process taking up to 3 minutes for each object. Although such reduced feature sets may be far from complete for many images and objects, the motivation is that only a few matching features agreeing on the pose, scale and image orientation of an object should be sufficient to support unambiguous object recognition. The bearing of utilising more complete sets of features is discussed later in the chapter.

The distributions of valid and invalid Bhattacharyya match scores presented in Figure 7.9 bear great significance to the task of object recognition and retrospectively to the model matching task. These match score distributions indicate that valid and invalid match distributions are well-separated, with, for instance, no invalid feature matches scoring over 0.85 from a sample of over 28,000 randomly matched features. To this end, if we can be sure that an object to be recognised may only appear in an unoccluded context, as with the test images used for model matching experiments in the previous chapter, only matches with a high Bhattacharyya match score, e.g. over 0.8, may need to be considered. Conversely, setting a minimum tolerable feature match score, for instance, below 0.5 would mean that the majority of random invalid matches are accounted for, thus potentially confusing recognition and exhausting processing faculties. The trade off here is that there is the potential for certain features to be obscured through unfavourable illumination conditions and occlusion, thus degrading any legitimate match scores. Furthermore, interference from scene clutter may saturate any image PGHs, meaning that predicted match scores will devalue between standardly normalised PGHs. In order to assess the bearing of these considerations, the initial single-object based object recognition experiments have been conducted while maintaining match lists made up of entries over 3 sets of Bhattacharyya match score values; 0.6, 0.7 and 0.8. A minimum of 2 features agreeing on the pose and scale of a hypothesised object were required as the basis of match list construction for these experiments on account of the limited numbers of features used.

8.3.2 Results & Discussion

Following analysis of the results of the experiments outlined above, out of the 30 random sampled object images, if using a minimum Bhattacharyya match score of 0.8

as the basis of object recognition, the correct object was identified 29 times (97%), otherwise being closely confused as the 3rd best supported object for a single instance of the plug object; for which there were very few supporting features and for which the object shared geometry with the other 2 objects. As the minimum permissible Bhattacharyya match score was lowered, the recognition results became less-distinct, as other coincidental match lists arose for confused object features. Specifically, when the minimum match score was lowered to 0.7, the correct solution was identified 25 times (83%), being ranked second on 1 occasion, third on 3 others and seventh in the case of the plug object identified in the previous scenario. Finally, for match lists formulated with a 0.6 minimum Bhattacharyya match score, the correct solution was identified 23 times (77%), instead being ranked second twice, third once, fourth thrice and seventh for the aforementioned plug object. In order to elucidate these experimental findings and highlight the utility of the proposed system, object recognition charts are presented in Figures 8.1 through 8.15 indicating the relative match scores for each of the 15 objects in the dataset for a uniformly sampled image of each.

The most distinctive characteristic of the results indicated in Figures 8.1 through 8.15 is that, in general, using the higher minimum match score constraint (0.8) not only maximises the number of occasions on which the recognition system works but also significantly suppresses any possibly confusing interference from other invalid model matches. If using this higher minimum match score as the basis of unoccluded object recognition, in all but exceptional cases, the system can be regarded as being operationally valid.

Even if an object is not distinctly recognised on account of the best match list score, it is evident that valid solutions will still have reasonably relatively high match scores, allowing for the proposed verification routines to be implemented across any competing hypotheses to find the most likely match and recognised object. The verification routines can therefore be used to take account of all pertinent image evidence even if only a few features are used as the basis of model matching and object recognition. Although only a single matching line feature has been shown to be sufficient to support model matching and object localisation, it is reiterated that the completeness property of the PGH representation stems from the constraints imposed by analysis across multiple non-collinear features.

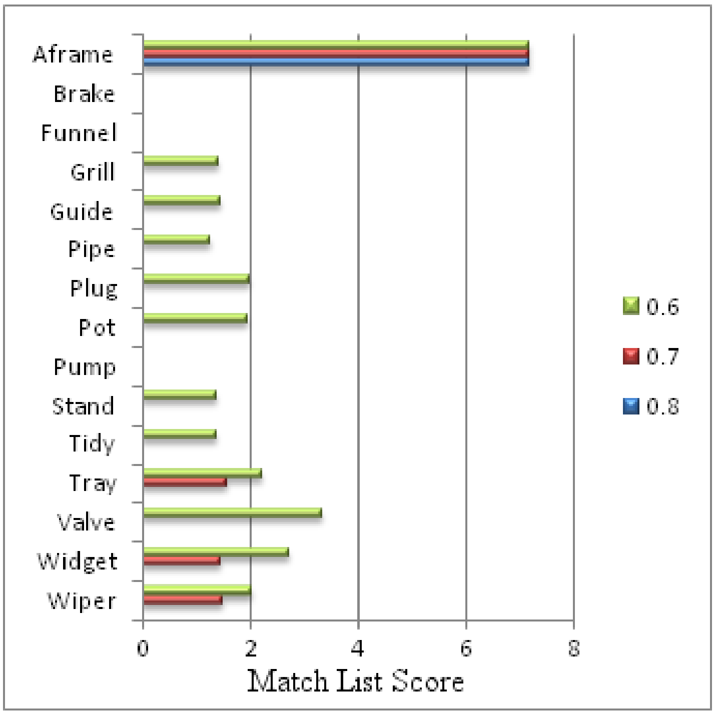


Figure 8.1: *The bar chart above indicates object recognition match list scores obtained for each of the 15 objects in the test database as compared to an image of the Aframe object. Similar charts are presented for an equivalent image of each of the remaining 14 objects in the following figures.* ***Match list scores*** *are formulated as a sum of the Bhattacharyya match scores between the best set of model and image edge features agreeing on the pose and scale of the corresponding object. For these experiments, the longest 24 linear image edge features have been matched to the longest 12 line features from each of 42 object views, as explained in the preceding text. Each object's entry in the chart is threefold (red, green & blue); representing the match scores attained when the match lists were formulated with features above the specified* ***minimum Bhattacharyya match scores*** *(**0.6**, **0.7** and **0.8**). Object recognition is shown to be best supported throughout the following example images with the higher grouping constraint (0.8 (blue)), with which all 15 objects are correctly recognised via a forced best choice decision.*

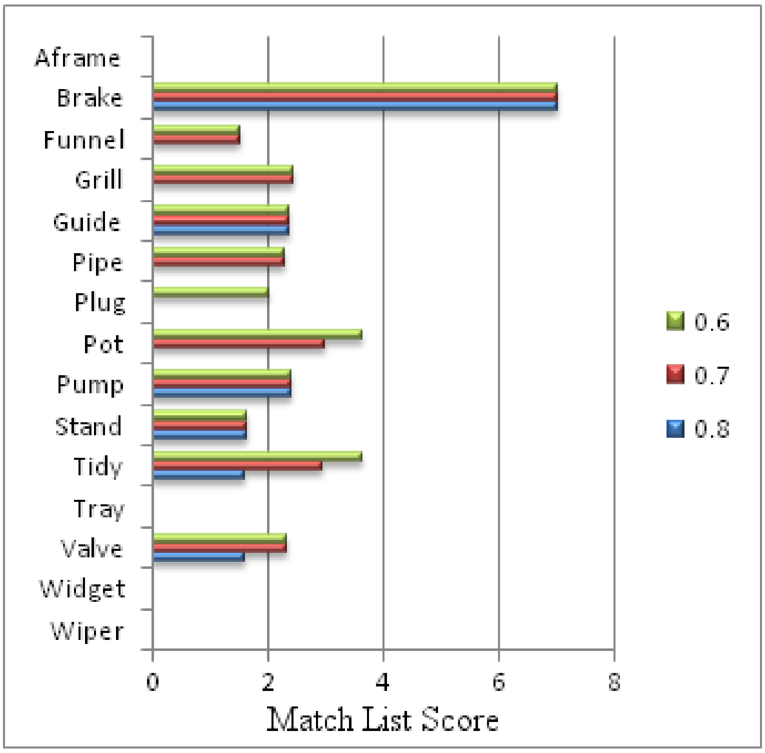


Figure 8.2: *Brake: Object Recognition Match Scores. See Figure 8.1.*

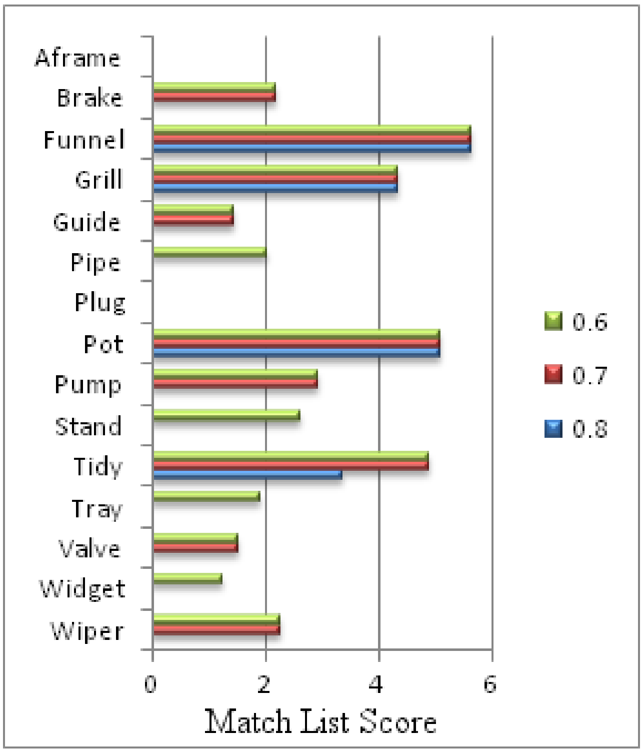


Figure 8.3: *Funnel: Object Recognition Match Scores. See Figure 8.1.*

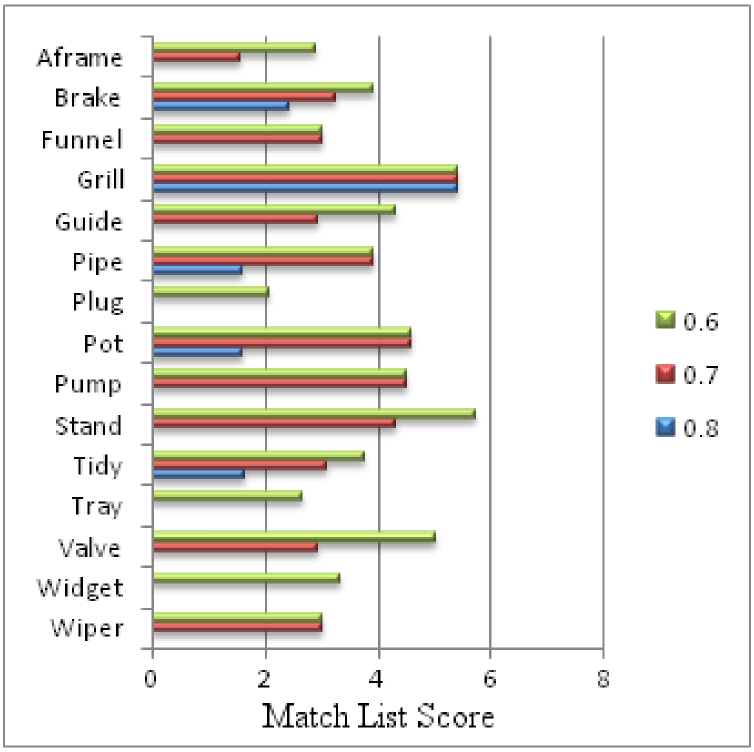


Figure 8.4: *Grill: Object Recognition Match Scores. See Figure 8.1.*

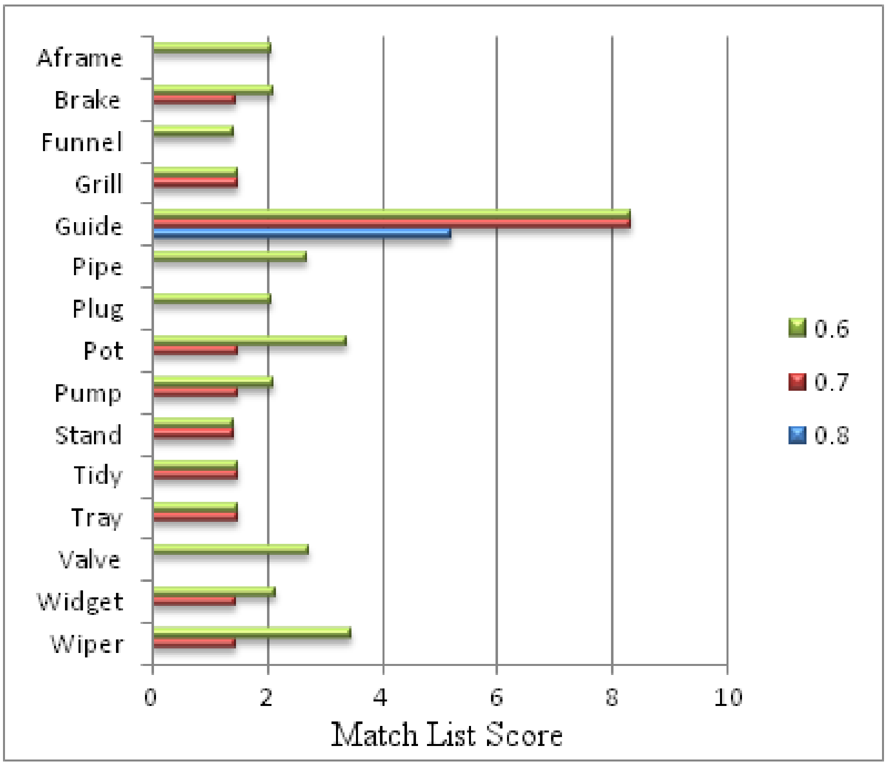


Figure 8.5: *Guide: Object Recognition Match Scores. See Figure 8.1.*

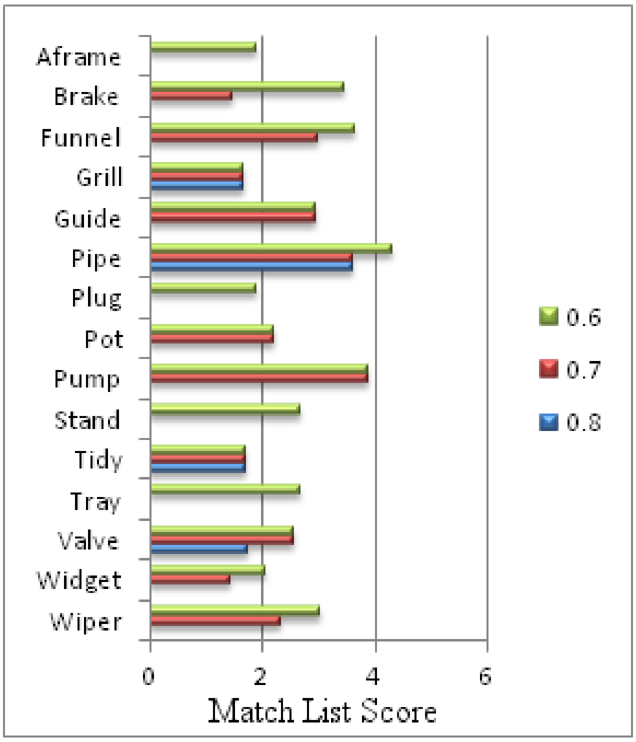


Figure 8.6: *Pipe: Object Recognition Match Scores. See Figure 8.1.*

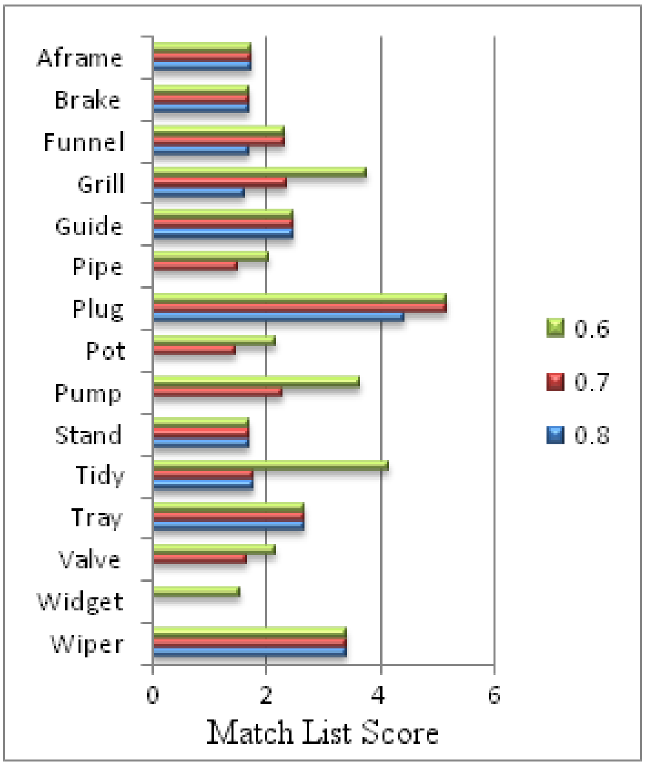


Figure 8.7: *Plug: Object Recognition Match Scores. See Figure 8.1.*

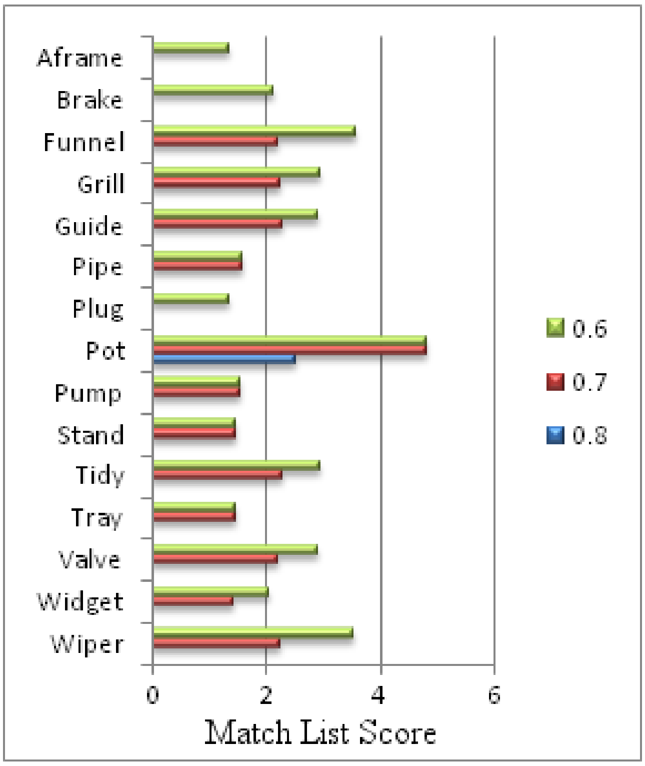


Figure 8.8: *Pot: Object Recognition Match Scores. See Figure 8.1.*

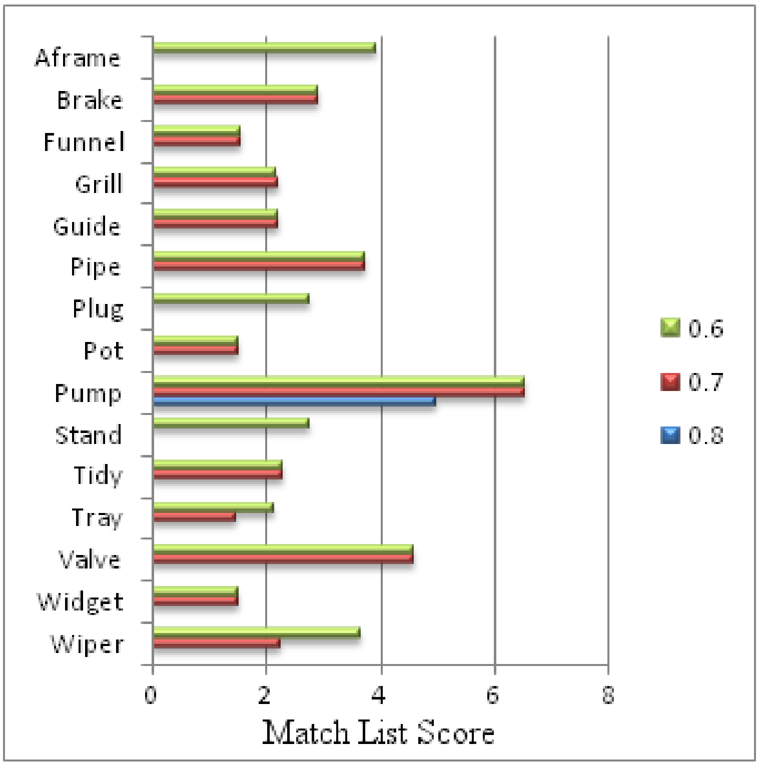


Figure 8.9: *Pump: Object Recognition Match Scores. See Figure 8.1.*

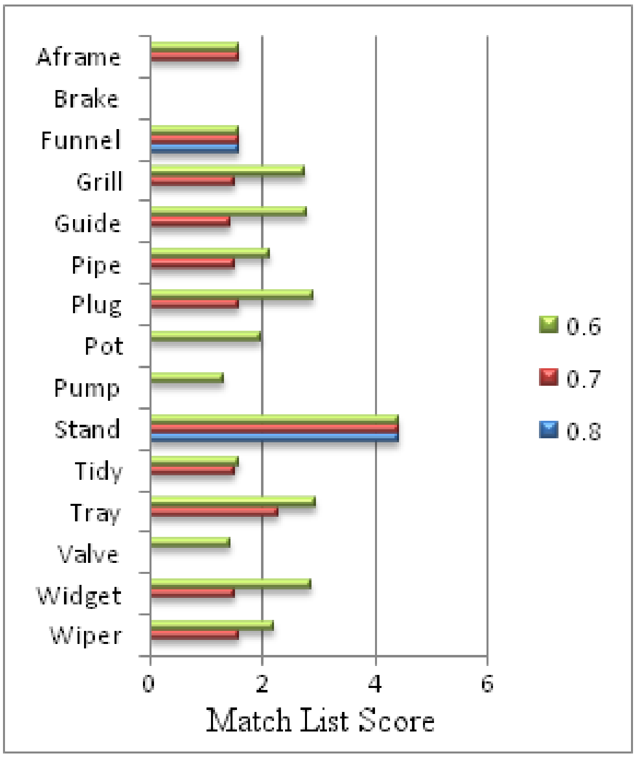


Figure 8.10: *Stand: Object Recognition Match Scores. See Figure 8.1.*

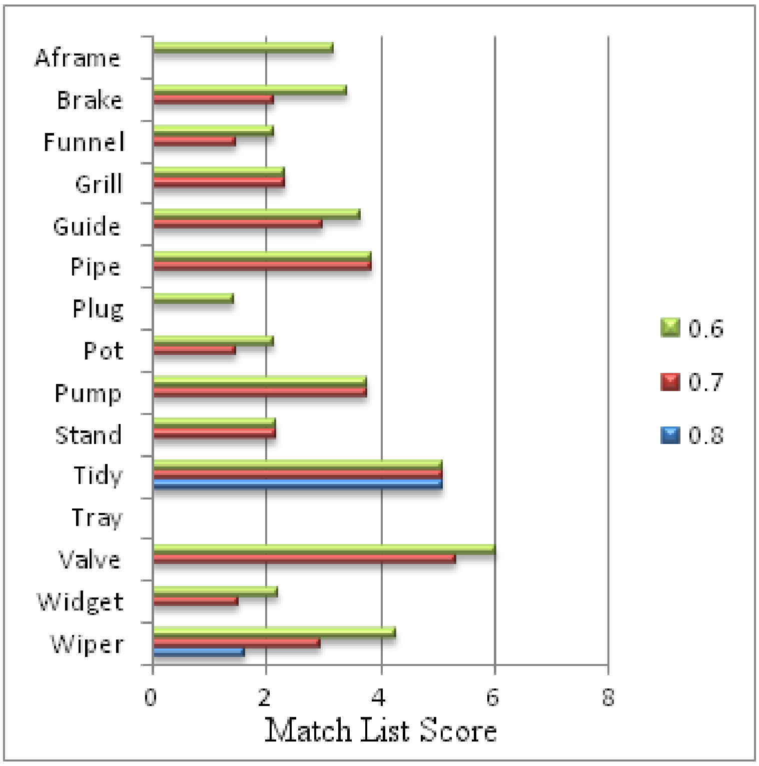


Figure 8.11: *Tidy: Object Recognition Match Scores. See Figure 8.1.*

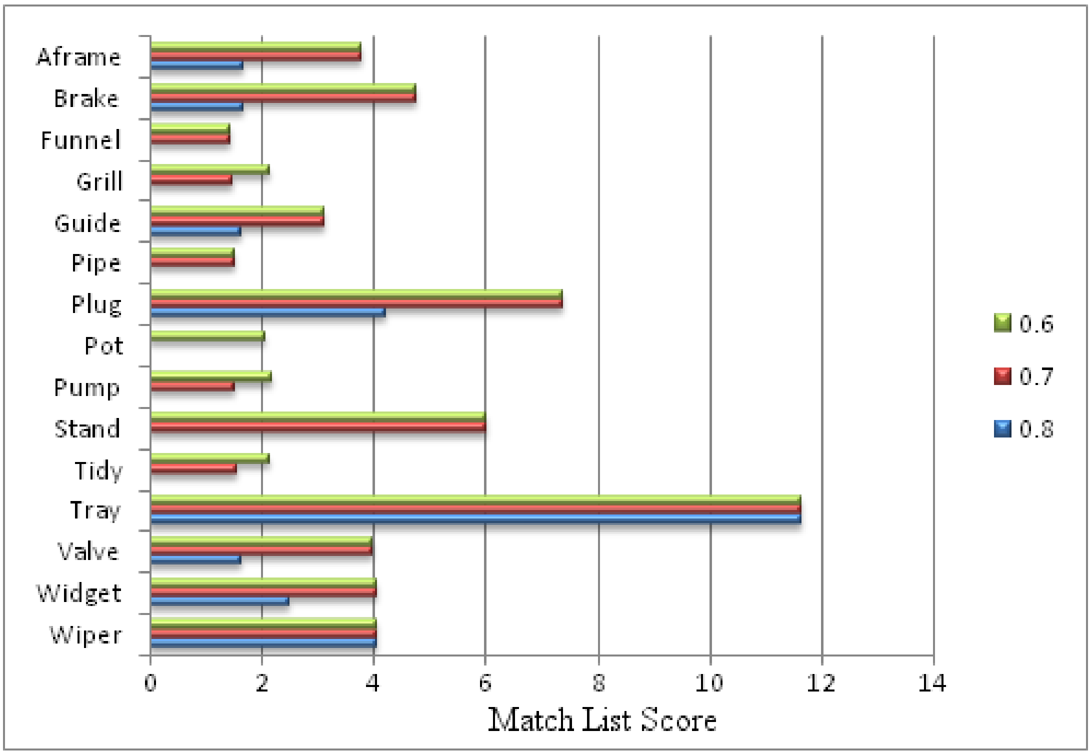


Figure 8.12: *Tray: Object Recognition Match Scores. See Figure 8.1.*

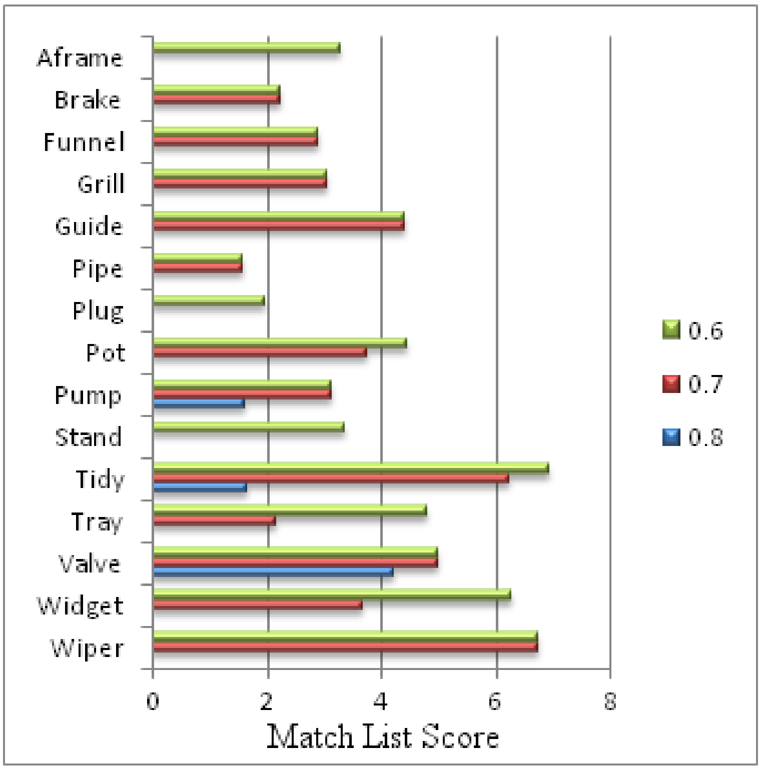


Figure 8.13: *Valve: Object Recognition Match Scores. See Figure 8.1.*

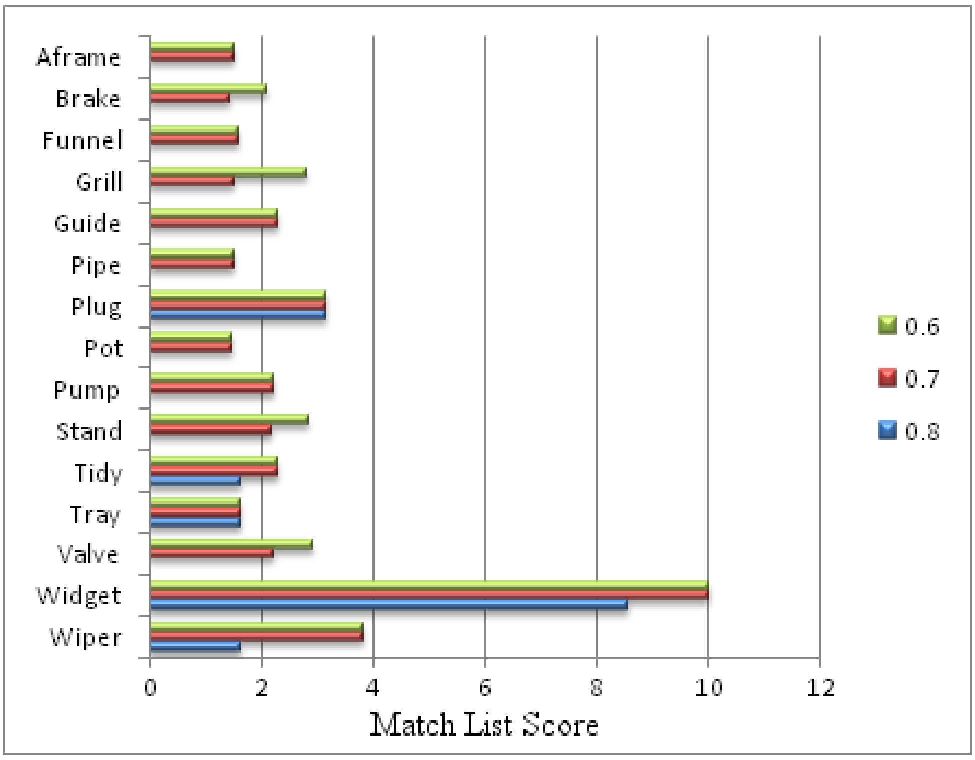


Figure 8.14: *Widget: Object Recognition Match Scores. See Figure 8.1.*

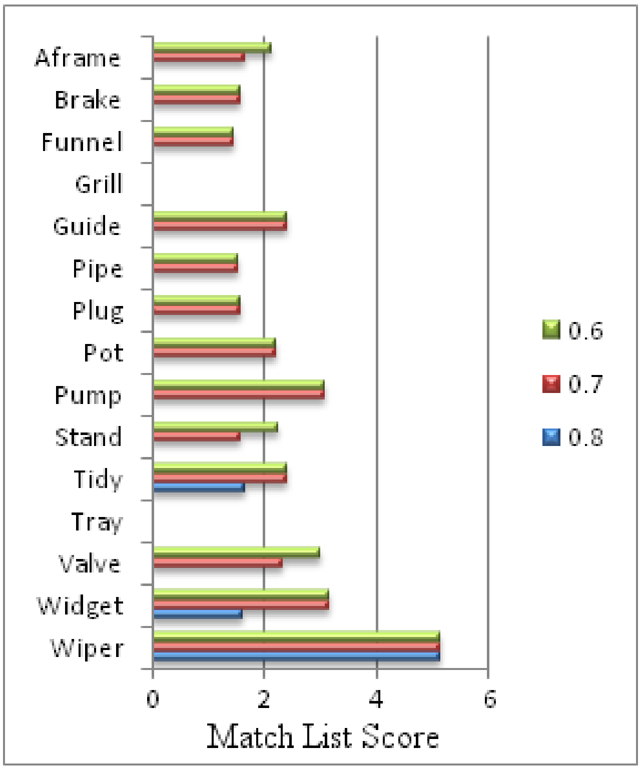


Figure 8.15: *Wiper: Object Recognition Match Scores. See Figure 8.1.*

8.4 3D Object Recognition in Cluttered and Occluded Scenes

The final problem facing development of a PGH-based 3D object recognition system is that of recognition through the everyday factors of occlusion and clutter. As evidenced with example object edge maps in Figure 7.4, the PGH representation is otherwise robust to distortion of objects' edge-based projected appearances through missing features arising from unfavourable environmental illumination conditions. Resilience has also been demonstrated to model matching for objects' edge maps corrupted by mild clutter arising from shadows, background artefacts, corroded surfaces, specular highlights and reflections.

One of the main challenges facing any object recognition system is robustness to occlusion, i.e. being able to recognise objects from highly fragmentary evidence; just as the human visual system is readily able. Previous research on 2D object recognition has proven that the PGH representation is suitable for identifying instances of highly occluded edge-defined objects in images [46][44].The main mechanism supporting this capability is the individual PGH-based modelling of each and every linear edge segment, in its local context, defining an object's projected shape. Any well-supported image-sampled line fragments agreeing on the location, pose and scale of an object are then identified in Hough transform spaces, thus parameterising the particulars of the hypothesised 2D object. Although the modelling of every single linear edge segment making up an object's projected pose may appear an extreme form of processing, especially for highly curved shapes, if we need to be able to recognise an object from only an arbitrary, very restricted partial aspect, such a redundant, exhaustive representation is essential. This issue raises another fundamental matter concerning PGH-based recognition, regarding the intra-object scope of each feature's sampled PGH region.

PGHS are currently learned for wireframe model projections as centred in the development environment's virtual model television (currently spanning 256 pixels). Each feature's assigned directed PGH is then captured over a 50 pixel perpendicular range in each direction. Image-sampled PGHs are constructed in a similar manner, with the image edge geometry being scaled in corresponding consecutive intervals (relative to image size) as the basis of scale-range-invariant object detection. While for some objects' central features, this local-part-based representation means that most of the projected structure may be included in the PGH, for more peripheral features and broader objects, only more local features are included. The motivation here is that by learning part-based representations, the system should be able to recognise fragments of an object in an image with high match scores even if other portions of the object are totally occluded. Otherwise, very sparse correspondence across a global template will be indistinguishable from the majority of random incorrect matches observed between other features, meaning that recognition-based search becomes exhaustive and soon exhausted. Such part-based processing potentially opens the way for system optimisation through shared object features, as presumably characterised by the parallel distributed architecture of the human brain's visual pathways. Furthermore, the reduced scope of each feature's PGH means that much less data is required for storage in memory and far fewer bin comparisons are required for matching, thus making the process of model matching potentially much more efficient.

Ideally then, all, or at least the majority, of an object's projected edge-based appearance should be piecewise modelled with PGHs in order to be able to recognise the object in difficult, highly occluded scenarios where only a very limited random subset of features may be visible. While this may be tractable for 2D object recognition where test objects may comprise, for instance 40 or 50 linear boundary edge segments and a single fixed view, a problem arises for more complex 3D objects, where hundreds of features may need modelling around the view-sphere. While such issues become manageable for massively parallel processing machines such as the human brain, problems become apparent when dealing with conventional sequential processing streams. As the following diagrams indicate (Figure 8.16), some objects may require hundreds of line features for anything like complete and accurate coverage. Since the PGH representation is based upon linear edge segments, the most obvious problems occur for objects composed of curved features.

With further reference to the images presented in Figure 8.16, it can safely be assumed that in order to recognise any such object by virtue of agreement between a number of different features in cases where the object may be highly occluded, at least 100 features would be reasonably required to represent each view. If this many features are required for each object, the number of potential image edge features required for analysis would be this number (minus any expected occluded features) multiplied by the expected number of objects in a scene plus any features arising from irrelevant scene clutter. By way of example, Figure 8.4 presents 5 scene edge images (546 by 432 pixels), each containing a subset of 3 of the 15 objects from the data set, positioned

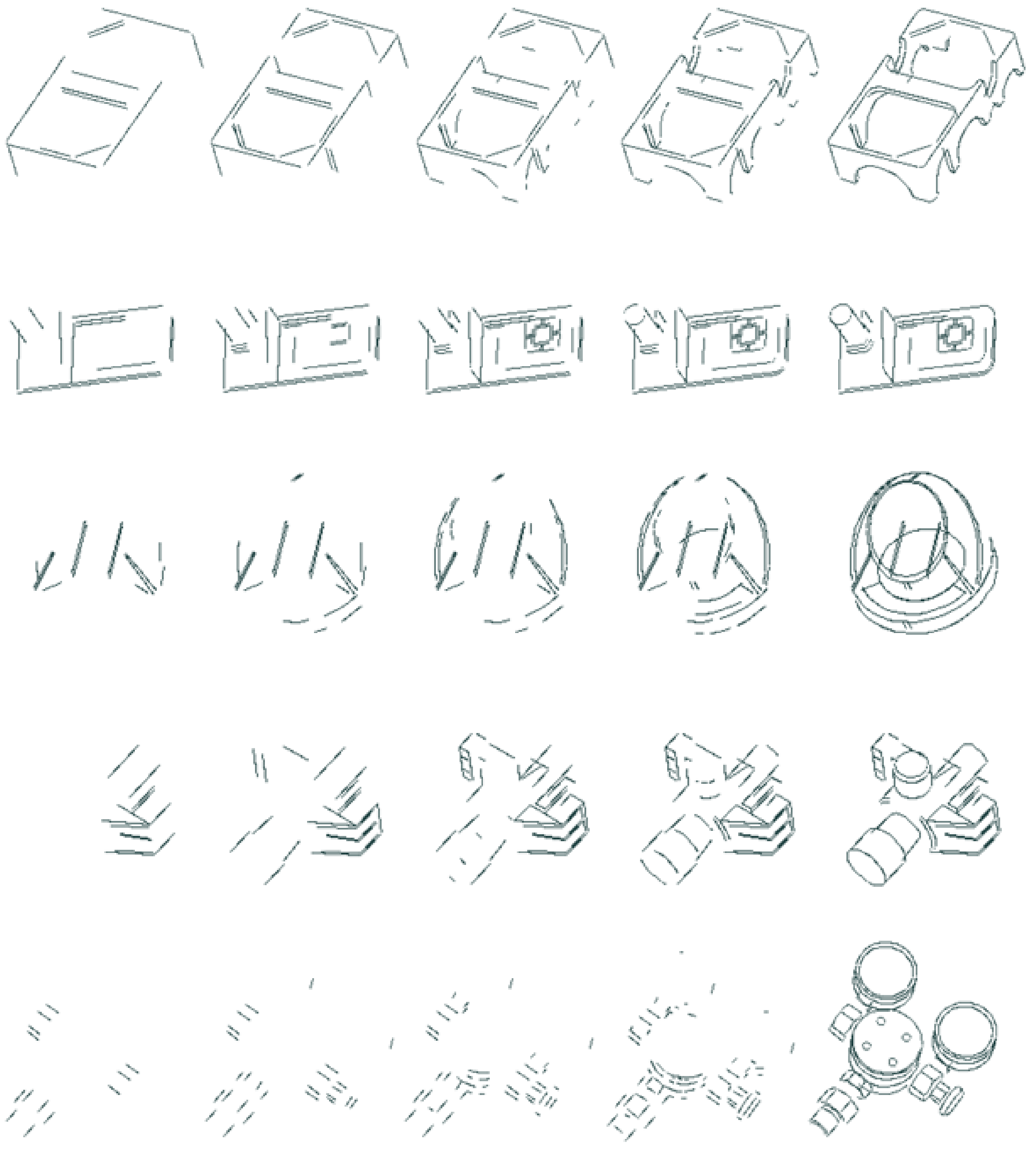


Figure 8.16: *The first four columns of objects in the diagram above represent the specified objects' edge-based appearances as made up from the longest 12, 25, 50 and 100 line segments respectively from left to right. The rightmost column represents each object's full appearance, requiring 375 lines for the (cutting-) guide object, 160 for the widget, 350 for the grill, 275 for the pump and 900 for the valve. The diagrams indicate the computational problems involved with representing PGHs for every object-view-bound linear-edge feature. In the current implementation, PGHs are sampled for edge features within a 50 pixel perpendicular distance (see Figure 3.2) for each reference line relative to a 3D object-encompassing view-sphere - centred in the 256 by 256 virtual model television (see Subsection 6.2.1).*

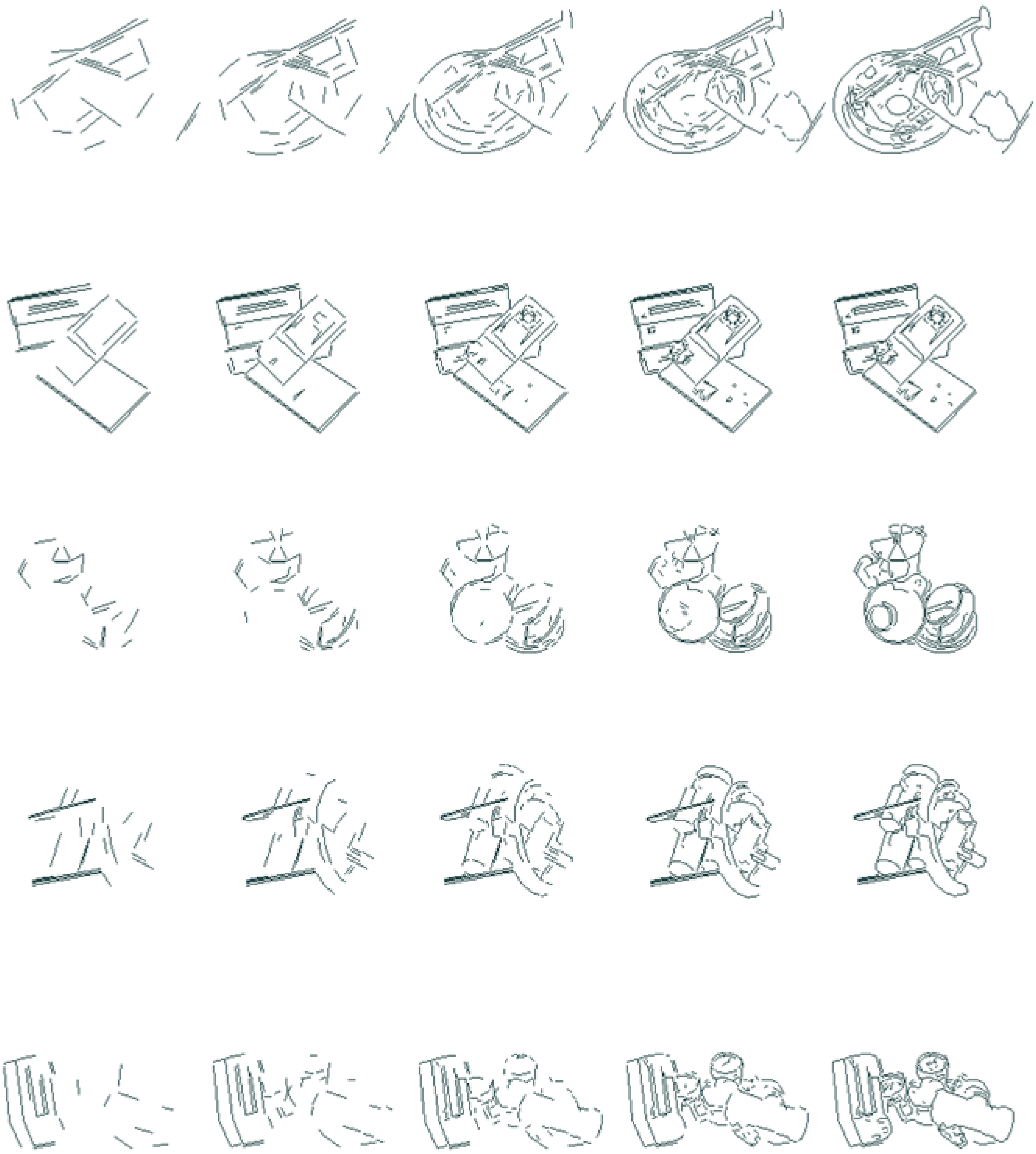


Figure 8.17: *The 5 rows of linear-edge-based images in the figure above represent the amount of scene geometry covered for images containing 3 test objects when varying numbers of the longest linear edge features are sampled as the basis of model matching and object recognition. From left to right, there are 25, 50, 100 and 200 lines used across the first 4 columns. The final column represents all image edge lines over 2 pixels (the original images are scaled to 576 by 432 pixels) in each case, with there being 455, 205, 385, 330 and 440 such features for the respective rows of images from top to bottom. A lack of contrast means that some expected model features are missing altogether. A PGH is formed for each sampled line feature with the 2D edge geometry being scaled in consecutive intervals across a (currently) +− 50 pixel perpendicular distance (see Figure 3.2) as the basis of matching (see Subsection 7.2.1).*

so as to provide instances of partial occlusion and scene clutter. As with the object wireframe examples in Figure 8.16, the images in Figure 8.4 have been reproduced as composed of varying numbers of the longest linear edge features. For relatively small and simple images such as these, it can reasonably be argued that at least 200 linear edge features would be required in order to provide some degree of global inter-object coverage as required for the simplest cases of occlusion.

Using the standard laptop computer that supported this project's research (2009), it takes around 1 minute to process 8 image lines for a single object at a restricted range of scale for just the longest 12 features across each view. If we were instead to require that, e.g., 100 model features were required instead of 12 to help support occluded recognition, this means that it would take around 8 minutes for each object for each 8 image lines. Assuming that many more image lines would then require analysis, e.g. 200 as a starting point, this would mean that to examine a single simple image for a single possibly occluded object (at a restricted range of scale) would take over 3 hours. Considering that there are 15 objects in the current data set, this means that it would take nearly 2 days to run an occluded object recognition experiment on a single primitive image, not to mention the memory overheads required in order to maintain such cumulative cross-referenced processing across so many features. In this regard, the current system proves intractable for generalised 3D object recognition in anything like real-time using a single sequential processing paradigm and a large object database. The remaining issues concern whether the representation is indeed optimal for the 3D edge-based object recognition task, whether edge information alone is sufficient and if so how best to optimise the proposed processing routines.

8.4.1 Methods

Despite the outlined limitations of the proposed approach to unconstrained view-based 3D object recognition with regard to processing requirements on conventional computing machines, focus feature-based mechanisms can serve to more efficiently identify objects via limited sets of features, where some portion of the object is clearly visible and modelled. Indeed, only a single line feature's PGH is typically required to indicate the corresponding image pose and scale of a hypothesised object. Preliminary object recognition experiments were therefore run on the example images depicted in Figure 8.17 using reduced amounts of data in order to see whether the system would still be able to produce sensible recognition results. These experiments also challenge the system's capabilities in terms of model matching robustness in the face of mild occlusion and scene clutter. For these experiments, 9 iterations of the 8 longest scene lines were sampled instead of 3, with the longest 12 features for each object again being sampled for matching. Match lists were formulated as they were for the previous singularly presented object recognition experiments. The best verification score, as previously used to determine the best intra-object match for each object as matched to each image, has also been recorded to investigate the utility of the proposed methodology for inter-object verification in more cluttered and occluded scenes.

Although the single object recognition results in the previous section were shown to be most robust when requiring each matched feature to have a high minimum Bhattacharyya match score (0.8), since we are now dealing with occluding objects in cluttered scenes, we must allow for some degradation of valid PGH match scores. A minimum match score of 0.5 has therefore been sampled for these experiments. Each object is listed with an indication of the rank of both its match list score and verification score relative to the other 14 objects in the dataset. Since there were 3 objects in each image, ideally, each object should be ranked third or higher in each regard

8.4.2 Results & Discussion

Although the intention was to produce bar charts (similar to those in Figure 8.1) to indicate the relative amount of support for competing object recognition hypotheses across each of the 5 mixed-object sample images, the results were found to be too confused to warrant such detailed analysis using such limited sets of features. In context, the match list recognition results indicated in the bar charts presented in the previous section (Figures 8.1-15) show how easily confused a number of the objects were using the lower match score bounds and limited sets of reference features before considering the problems of occlusion and distracting scene clutter. There was not enough time on the project to conduct more extensive experiments or to more comprehensively investigate extension of the proposed techniques for dealing with the more challenging aspects of the recognition problem for complex cluttered scenes. The results of the object recognition experiments in cluttered and occluded scenes are however summarised in Table (Figure) 8.19 with reference to the images presented in Figure 8.18.

8 out of 15 objects were correctly recognised as being present in the sample images on the basis of a forced 3 way decision based on match list weight. In the other cases, either enough features were not sampled for recognition or other coincidental match lists arose for confused aspects of other objects. For example, as evidenced in the third row of Figure 8.17, over a hundred image lines would be required to provide basic coverage of the curved-feature funnel object as required for recognition. Further considering that only the 12 longest linear edge features were being used to represent each object view, this meant that the limited feature sets used for experimentation were insufficient to provide an unbiased analysis and there was little point presenting detailed generic recognition results without being able to manage more comprehensive image and model edge feature sets.

Retrospectively, there may be some advantage to be gained by sub-sampling each image into, for instance, rectangular octants, with the longest reference features passing through each region being sampled. This may give some advantage in distributing features more evenly across objects of varying complexity in cluttered scenes without having to exhaustively examine the bulk of the image edge data. The outer bounding rectangle could, for instance, be positioned, scaled and oriented to align with the first 2 sampled eigenvectors of the image edge data distribution to more evenly partition the data. Such an image sampling methodology would also be amenable to parallel processing, where separate processors could synchronously analyse certain image regions, allowing for high speed recognition in complex scenes.This is a potential avenue of future research.

On a number of occasions, spurious object pose matches also occurred across multiple image locations. Retrospectively, an object location constraint on model matching would therefore be of use in enhancing recognition acuity. A reasonably high number of well-supported coincidental feature matches also arose for various objects' views against random image edge feature distributions. A number of the objects used for experimentation have some very simple aspects that were easily confused. One problem here is that if a feature does match to some random image edge structure, as is evidently highly likely, the chances are that a number of other related features may also be well matched, meaning that reasonably well-supported match lists may arise for invalid hypotheses.

Despite any inadequacies of the proposed technique for object recognition using limited sets of reference features, the model match verification scores obtained for each

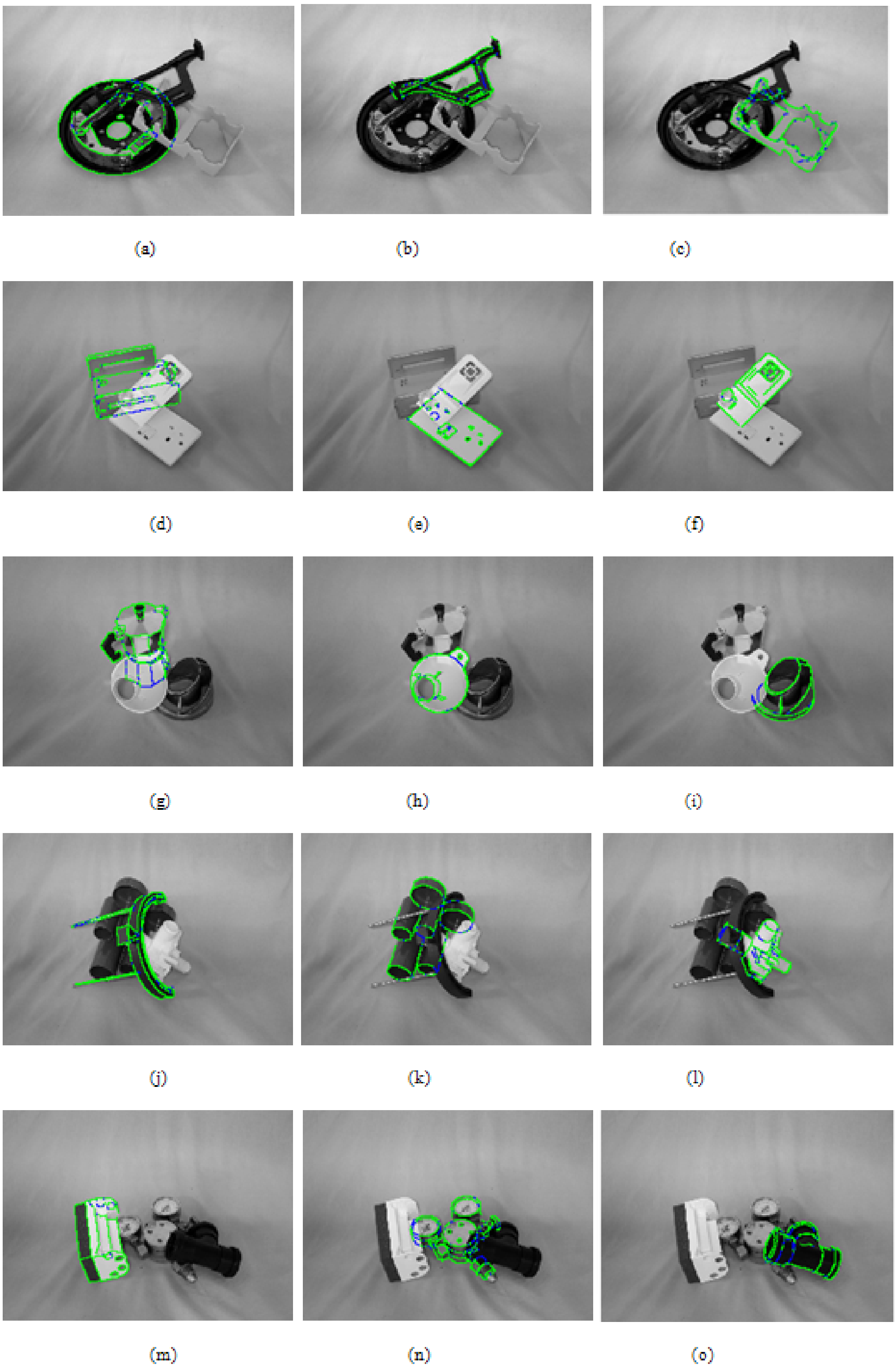


Figure 8.18: *The images above represent 3D wireframe object models matched to their corresponding image poses in mildly cluttered and occluded scenes using the proposed model matching and object recognition system. Projected model feature points passing the edge location hypothesis test with a 1% confidence limit are shaded in green and those failing the test in blue.*

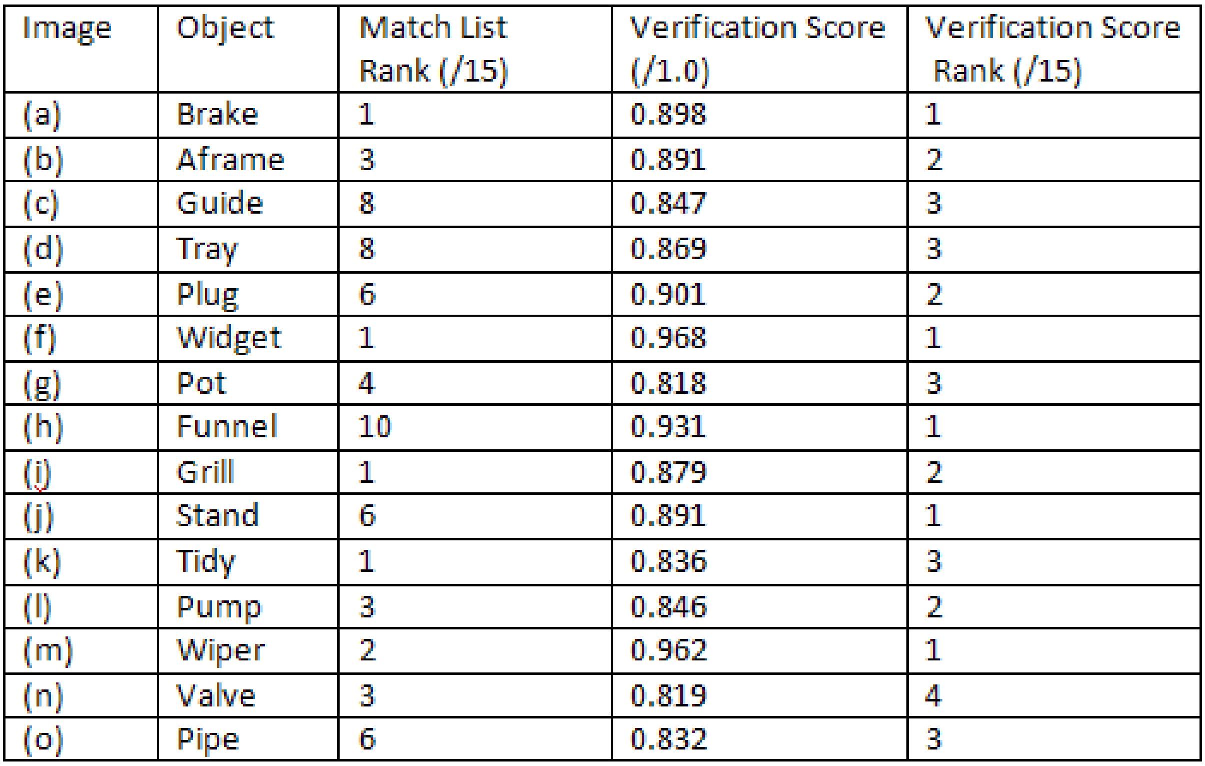


Image	Object	Match List Rank (/15)	Verification Score (/1.0)	Verification Score Rank (/15)
(a)	Brake	1	0.898	1
(b)	Aframe	3	0.891	2
(c)	Guide	8	0.847	3
(d)	Tray	8	0.869	3
(e)	Plug	6	0.901	2
(f)	Widget	1	0.968	1
(g)	Pot	4	0.818	3
(h)	Funnel	10	0.931	1
(i)	Grill	1	0.879	2
(j)	Stand	6	0.891	1
(k)	Tidy	1	0.836	3
(l)	Pump	3	0.846	2
(m)	Wiper	2	0.962	1
(n)	Valve	3	0.819	4
(o)	Pipe	6	0.832	3

Figure 8.19: *The results of the object recognition experiments for mildly cluttered and occluded scenes are tabulated above. Although there is evidently some confusion with regard to match list scores on account of the limited numbers of features used, the final verification scores serve to more accurately identify valid object recognition hypotheses.*

object otherwise proved to be far more robust indicators for object recognition, allowing recognition to be performed much more efficiently using limited sets of focus features. In this regard, each object is effectively model matched to the image, with the possibility for examining longer (and possibly less pronounced) match hypothesis lists, and the best verification score is determined. The results in Table (Figure) 8.19 show that, given an accurate projected model alignment, despite interference from scene clutter and occlusion, the proposed model match verification metric is able to (semi-) reliably differentiate the correct solutions, with 14 out of 15 objects being correctly ranked. In the single case where an invalid model match scored the second highest verification score, this was for a very simple side-profile aspect of the widget object that coincidentally aligned with the corner of the wiper object. Upon close examination, this confusion could be resolved by constraining the tolerable lateral shifting for the widget object's features, which would naturally occur for such an object in the proposed learning framework.

The results of the partially cluttered and occluded scene object recognition experiments are qualitatively illustrated in Figure 8.18. In each example, the best matching result for each imaged object is displayed as an overlaid 3D wireframe projection, with feature sample points passing the 1% edge location hypothesis test shaded in green and those failing in blue. As discussed above, the limited numbers of features used as the basis of these experiments meant that these solutions were not necessarily the best supported in terms of match list scores.

Despite the quality of the model matching results in the images presented in Figure 8.18, image (o) represents an instance where the system had trouble finding the correct solution. The black pipe object in image (o) was very difficult to detect, as observed for the model matching experiments in the previous chapter. This is mainly because all the internal edge features modelled for the object were typically not detectable as image edges and all that remained was a very sparse and uninformative outer profile. The circular nature of the majority of features also meant that many more line reference features were required to represent the object for recognition. Furthermore, interactions between specular highlights across the pipe object's glossy surface and surrounding noisy scene clutter also distracted and distorted object detection. The correctly matched object presented in Figure 8.18 required more features to be matched. Similarly, a symmetric, albeit well-aligned, pose was initially inferred for the funnel object in image (h) on account of the very limited number of associated features captured with the current experimental bounds (see Figure 8.17), although without any cost to the relative ranking of the verification score. The correctly aligned funnel object, as shown, required another couple of image features to be accounted for; again reiterating the need for more comprehensive global feature analysis.

In order to deal with occluded objects more effectively, some form of reasoning about inter-object occlusion is required. For instance, following match list construction for a sample image, any objects that are recognised with very high match scores (e.g. $>$ 95% of feature sample points being verified as valid) should initially be considered as being foremost, allowing any overlapped model match hypotheses to be verified on the basis of any unoccluded projected sample points. An equivalent process would be required for any objects only partially captured by the camera frame. This would be essential functionality for addressing the more demanding aspects of the problem of scene interpretation.

The efficacy of the proposed scheme for matching learned 3D models to cluttered and occluded scenes is indicated for a further set of mixed object images in Appendix B.

8.5 Conclusions

This chapter has analysed the applicability of the PGH representation for the task of view-based 3D object recognition. Object recognition follows as a direct extension to the proposed view-based 3D model matching system described in the previous chapter. Although time constraints limited the extent of associated research, the work described in this chapter represents the culmination of the research described throughout this thesis.

Initial object recognition experiments concerned recognising single instances of each of the 15 objects in the test images used for model matching, with a random pair of such images being sampled for each object. In order to make the scheme temporally tractable, as with the previous chapter's model matching experiments, the longest 12 linear edge features from each of 42 sample model views were matched to the longest 24 from the image-based Canny edge map using a single 15% scale interval. Objects' view-sphere-based triangulated manifolds were again maintained with a minimum 0.9 Bhattacharyya match score reconstruction error. Following match list construction for each of the 15 objects for each image, the system was able to infer the correct solution (forced best choice) 29 out of 30 times with the missing object being ranked third amongst similar aspects of other objects. In each case, the system was able to accurately infer the relative pose and location of the object in physical space relative to the camera's frame of reference. This result affirmed the integrity of the system, providing robustness to missing, distorted and spurious image edge features. The remaining challenges involved maintaining recognition acuity through more extreme forms of scene clutter and occlusion.

In attempting to recognise partially occluded objects in more complex scenes including 3 objects at once, it was soon realised that recognition could not be reliably maintained if using significantly reduced sets of learned reference features as with the previous model matching and single-object recognition experiments. Although the experiments were run using 72 sample image lines instead of 24, on occasions, certain objects' edge features were not well-sampled amongst those image features, skewing recognition to favour other feature-rich objects. In line with original research on 2D object recognition, the true power of the PGH representation for recognising fragments of highly occluded objects stems from the redundancy of the representation with regard to representing the majority of linear edge features as individual PGHs. An analysis of the requirements of such a representation for mildly complex objects and images suggested that the associated processing costs made the present scheme intractable for standard sequential computing machines and large learned 3D object databases. It was estimated that it would take nearly 2 days to run such an experiment for a single image using the current 15 object test database and the proposed object recognition routines. As discussed in the next chapter, there are however a number of possible optimisation routes which may potentially speed things up by an order or two of magnitude.

Despite the limitations of using restricted sets of model and scene features as the basis of object recognition, associated experiments enabled the correct objects to be distinctly recognised on a number of occasions (i.e. for objects with a number of extended visible linear edge features), with the majority of objects being subsequently well matched to the corresponding image edge evidence despite mild clutter and occlusion. The model match verification scores sampled for each object were subsequently shown to more reliably support the object recognition process, correctly verifying 14 out of the 15 objects in the test dataset as the best supported object recognition hypotheses.

Another observable problem encountered by the proposed representation in more complex scenes relates to the normalisation of PGHs in support of the proposed local-range scale invariance. Scale invariance is achieved via interpolation of triangulated sections of the learned view shape manifolds between scaled samples of the image-sampled edge geometry, meaning that each image-based PGH sample requires normalisation. If any spurious data arising from scene clutter and occlusion is encoded by the PGH, the normalisation process will down-weight any bin entries for valid structure, thus obscuring detection of valid feature matches. In order to discount any such interference, any PGH bins corresponding to zero values shared by the 3 triangular reference nodes must be excluded from normalisation and matching. Without such a mechanism, recognition results will become too confused in heavily cluttered and occluded scenes to support object recognition. Under this proposed model of feature detection, if more object and image features can be used for recognition in a reasonable amount of time, object recognition should proceed on account of the proportion of total feature length accounted for across hypothesised objects' projected views. The remaining potential for system optimisation and a critique of the theories outlined in this thesis are presented in the following chapter.

Chapter 9

Conclusions

9.1 Thesis Summary

The main aim of this Ph.D. project was to investigate the potential of the Pairwise Geometric Histogram (PGH) representation as the basis of a machine learning computer vision system for the recognition of projected, edge-defined 3D shape. Additional research aims were to evaluate the relative merits of the pre-existing TINA stereo vision-based model matching system, for which a number of system extensions were also required. Associated research required the design and implementation of a user interface for view-based wireframe model construction and also a generic framework for the analysis of projected edge-features in support of image-based localisation, camera calibration optimisation and matching feature verification.

Following an introduction to the project in Chapter 1, Chapter 2 reviewed the field of computer vision to analyse whether any existing methodologies offered valid solutions to the problems of 3D model matching and object recognition. An introductory review of associated research regarding the complementary processes underpinning human visual recognition was also presented. Despite over 40 years' worldwide research into machine vision and object recognition, it was concluded that there is evidently no established solution to the problem of projected edge-based model matching and object recognition with existing techniques all having significant shortfalls, especially for fine-scale, specific-object recognition. The PGH representation was proposed as an optimal, tailor-engineered solution for these tasks, supported by a number of arguments affirming the significance of image intensity edge-based features for visual object recognition.

Under the hypothesis that PGHs are indeed a valid solution to the edge-based 3D object recognition and localisation problem, Chapter 3 went on to review the nature and history of the Pairwise Geometric Histogram (PGH) representation, detailing the operation of the existing faculties for 2D object recognition and localisation. Previous associated research regarding 2D object recognition proved that the PGH representation was sufficient to support recognition of views of highly occluded objects, in contrast to other systems described in the computer vision literature (see Chapter 2). The remaining challenge was the construction of a system that would be able to learn a 3D object's range of projected appearance around the view-sphere. Problems included the interpolation of PGHs between sampled reference view-points at a range of scales and perspectives and the subsequent pose-determination and localisation of recognised 3D objects' virtual representations relative to the camera's frame of reference.

Accompanying the pre-existing TINA 2D object recognition system, the TINA computer vision system is complemented by a stereo-vision based 3D Model Matching (3DMM) system. One of the main problems with the 3DMM is that the statistics used to sample feature correspondence between model and reconstructed scene geometry are not descriptive enough to support model match verification or comparative object recognition. 3D features are simply validated as matching or not, being weighted by their lengths. It is otherwise difficult, if not, at times, impossible to get perfectly accurate stereo reconstructions of noisy edge data, meaning that more precise accounts of match quality may be too restrictive for practical use. In order to evaluate the validity of any such model match hypotheses, Chapter 4 presented a novel quantitative statistical metric to account for the likelihood that a projected 3D wireframe oriented-edge feature point was supported by the image evidence. Cumulative account of such information across projected object views supports optimised camera-calibration and image alignment of object models in support of accurate 3D object localisation and pose determination and subsequent quantitative model match verification. The integrity of the proposed metrics was proved via analysis of the observed distributions of likelihood terms sampled for real images of the test objects relative to those theoretically predicted. Further to closing the loop on the functionality of the stereo vision-based model matching system, the proposed methodologies provided compatible functionality for the proposed view-based approach to 3D model matching and object recognition.

Chapter 5 introduced the computational requirements of a 3D view-based model matching and object recognition system with particular emphasis being placed on the interpolation of projected shape between view-sphere sampled PGHs. Observing that continuous PGH-sampled shape manifolds may be highly non-linear, high dimensional and irregular, the problems with modelling such continuous manifolds for arbitrary shapes were identified, leading to the conclusion that modelling such manifolds as the basis of interpolation was intractable. The proposed alternative solution to this problem was to adopt a piecewise locally-linear representation. In accordance with the 2 independent axes of rotation realisable at the surface of the view-sphere (i.e. latitudinal and longitudinal rotation), a localised planar manifold representation mediated by triangulation was established. The underlying idea here was that a minimum interpolative reconstruction error be maintained across each learned connected triangular region. In order to account for variance in the projected scale of learned objects, a mechanism was proposed to interpolate triangulated reference regions between scaled intervals of the image geometry. In contrast to previous work on 2D PGH-based object recognition, only a single (or reduced sets of) scale of stored geometry was now used as the basis of model matching and object recognition, with the scale of the image edge geometry being scaled independently.

Chapter 5 went on to formulate a quadratic model as a solution to scale-interval-based PGH triangulation. The effects of perspective distortion were subsequently analysed, indicating that objects' PGH-based appearances were liable to significant distortion through proximal ranges of view-point. Certain classes of objects were shown to be more affected by perspective distortion than others with the conclusion that a separate interpolative model of perspective distortion would be required for invariant 3D object recognition. Subsequent research was however conducted over restricted ranges of scale, for objects positioned approximately 1 metre away from the camera, for which a fixed model of perspective, approximating that observed 1 metre away from the camera, was sufficient. Finally, the utility of the proposed scale-interval invariant model matching mechanism was assessed in comparison to a non-interpolative, nearest-neighbour type matching strategy in terms of the view-sphere area covered by a single PGH relative to computational cost. Associated experiments indicated that despite the computational overheads associated with interpolation, the proposed methods were more efficient for encoding view-sphere-bound shape than simple nearest neighbour matching processes.

Chapter 6 introduced a scheme for partitioning 3D objects' view-spheres into sets of representative views to be used as the basis of automated learning in accordance with the mathematics outlined in Chapter 5. Observing that some objects' views' features may give rise to complex, highly triangulated manifolds, a connected triangle match score gradient descent type optimisation scheme was presented, supporting much more efficient parsing of complex manifolds. Finally, the single PGH feature-based mechanisms underpinning inference of the 3D pose and physical location of detected objects relative to the camera frame were presented.

With a framework for view-based 3D model matching in place, Chapter 7 reviewed the performance of the proposed model matching system across full view-sphere sampled tests of 15 wireframe-modelled 3D objects. In order to make the model matching process temporally tractable, the system was based upon analysis of reduced sets of model and image features. Specifically, the longest 12 linear edge features from each object view were matched to the longest 24 linear edge features sampled from the image geometry. Using 14 sampled image views for each object, the same experiments were then run for stereo pairs of images using the stereo vision-based 3D model matching system (3DMM) with a view to assessing the relative merits of the competing methodologies. The results of the model matching experiments indicated great contrast in the capabilities of the view and stereo-based methodologies. Excluding exceptional cases where the manually composed wireframe models proved inadequate for detection, the view-based system proved to be extremely reliable, serving to autonomously and accurately detect and localise nearly every single imaged object under essentially arbitrary illumination. In contrast, the full view-sphere nature of the sampled object images meant that a number of limitations of the stereo-based model matching routines came to light. More than 6 times as many model matching errors were made by the stereo system over its view-based counterpart, with model matching failing nearly 60% of the time. The main problems with stereo processing related to an inability to model and match view-based features, problems with detection of other types of feature and combinatorial search issues. The results of these experiments discounted stereo as the basis of a prospective object recognition system and retrospectively as that of a model matching system, with view-based recognition mechanisms proving to be far more reliable. With the view-based model matching system in place, subsequent verification experiments revealed that optimal valid/ invalid class separability was maintained by the edge location only metric, with an additional model of corresponding edge orientation information proving unnecessary.

Chapter 8 served to assess the applicability of the proposed view-based model matching techniques for the task of 3D object recognition. The process of object recognition followed as a direct extension to the proposed model matching routines, with match lists agreeing on the pose and scale of detected objects being formulated for each learned object for each sample image. Initial experiments concerned recognising individually presented instances of each of the 15 objects in the sample data set. Using a pair of random images of each of the 15 objects as the basis of experimentation, the proposed methodology proved suitable for the object recognition task, correctly recognising 29 out of 30 objects, with the only confused object being ranked third best under exceptional circumstances. The remaining challenges concerned maintaining recognition acuity for occluded objects in cluttered scenes. Previous research regarding 2D object recognition has proved the viability of the PGH representation for the recognition of highly occluded objects. The power of the PGH representation for this purpose stems from the PGH-based modelling of each and every linear edge feature defining an object's projected shape. Such a redundant representation is deemed essential if recognition is to be based upon random, highly-fragmentary edge evidence. In transferring such capabilities to the task of 3D object recognition, a significant problem arises in terms of the amount of data that requires analysis.

Although the proposed techniques prove adequate for single-object recognition based upon limited sets of focus features, subsequent analysis indicated that potentially hundreds of features may be required for both model and scene geometries in order to provide coverage of the majority of corresponding features. Without these levels of support, certain, especially occluded, objects may be excluded from scene analysis. Even for relatively simple objects and images, such comprehensive analysis was shown to take too much time to make the scheme tractable on current standard sequential computing hardware for large object databases. Although promising recognition results were obtained for mildly cluttered and partially occluded test images using limited sets of features in conjunction with the proposed verification routines, the true power of the representation could not be exploited without being able to manage processing of more comprehensive sets of model and image edge features. Finally, significant problems were identified with the process of PGH normalisation underpinning the proposed scheme for scale interpolation in cluttered and occluded scenes. In these cases, the process of normalisation may significantly devalue any legitimate PGH entries, thus lowering any valid match scores into the depths of the invalid match distribution. These problems may be avoided by instead normalising any scaled image-sampled PGHs across any non-zero bins in the triangulated PGH reference set. Such a process is deemed essential for any prospective PGH-based object recognition applications. Otherwise, if different amounts of clutter are present in each connected-scale PGH sample, interpolation of the normalised triangulated manifold between these scaled-image-sampled PGHs will be impaired.

9.2 Recommendations for Future Research

Unfortunately, the final object recognition results reported in this thesis were constrained by the limited amount of time available to complete the project. Another few months' worth of research would have enabled effort to be expended in optimising the proposed methodologies in support of application to more challenging recognition scenarios involving higher levels of scene clutter and occlusion and more extreme forms of illumination.

Perhaps the main inadequacy of the presented view-based 3D object recognition system is the cost of processing, which, as is, prohibits application of the proposed scheme for real-time object recognition applications involving large object databases with standard sequential computing hardware. As should now be clear, the main motivation supporting this project has been to design the recognition system to make optimal use of the available image data in an attempt to artificially emulate the acuity of the human visual recognition system for such tasks. Ideally, a parallel processing architecture would be utilised in accordance with the nature of visual processing in the human brain. Without immediate access to such specialised hardware, there are a number of potential means by which to speed up the presented system.

In the first instance, there is some scope for optimisation of the storage and access of the stored view-sphere manifolds. PGH-bin data is currently stored to 3 significant figures in standard decimal notation, whereas direct coding of machine language should support more efficient data access. For convenience, triangulated manifolds have so far been encoded as independent triplets of PGHs with obvious advantages to be gained in terms of storage and read/write access with PGHs being referenced from single stored unique instances. Since many objects' edge features are boundary features, in many cases, half the PGH will be completely blank. In these cases, it would be possible to indicate such information with a simple flag, meaning that half of many PGHs would not require storage, access or cross-referenced computation. More sophisticated techniques are otherwise expected to be able to parse arbitrary

PGH data in a similar manner to more pronounced effect. Individual rows (or columns), for instance, could be processed in a similar manner. With further regard to the triangulated manifold search optimisation function described in Chapter 6, a list of the connecting triangles for each making up each view's manifold could also be learned so that global manifold search is not required at each iteration for detecting connected triangular regions. Although the above-cited recommendations for improving the efficiency of the proposed system are likely to make a significant improvement to system performance, as discussed, a supplementary normalisation process is also deemed necessary to account for any adverse effects of scene clutter and occlusion. Although a a significant improvement in system efficiency is still likely to arise through implementation of these recommendations, it must also be noted that the presented experiments have only been performed through a single 15% scale interval for each projected wireframe object. A generic shape recognition system would require analysis over multiple scale intervals. Since this research has indicated that view-based mechanisms are essential for recognising many types of object, the issue of optimisation remains a critical one to ongoing research.

In the current implementation, all objects' shape manifolds have been learned to maintain a minimum 0.9 Bhattacharyya reconstruction match score across their constituent triangles in an attempt to minimise computational requirements. Although this has proved an effective basis for 3D model matching and introductory object recognition, higher levels of precision will benefit both tasks, especially for more challenging scenarios involving high levels of occlusion and those requiring finer-scale object discrimination. The missing or distorted 10% of PGH-encoded features may be essential to recognition for certain object views in these cases. In similar regards, the fixed localised nature of PGH sample regions may also be optimised to better effect. For instance, for simple objects such as the plug socket (as viewed face-on), the sparseness of the features meant that including only local features was suboptimal for model matching and recognition. Objects such as this would benefit from having their complete projected appearances encoded by individual PGHs. Again for extreme cases of occlusion, it may be beneficial to encode arbitrary shapes' entire projected appearances for each constituent feature's PGH. Although constituent match scores may be significantly affected under occlusion with this model, if using the proposed normalisation scheme, any fragments of the same object should have similar match scores (accounting for the same proportion of observed model features for each), which would be useful as a further match hypothesis grouping constraint. Considering the computational demands imposed by the current basic system, any potential benefits to be gained from these other considerations would however be dependent on optimising the existing routines, especially with regard to discounting any blank PGH regions from processing. If such an optimisation scheme could be implemented, it would be possible to match arbitrarily scoped learned PGHs to broadly sampled image PGHs without any particular cost deficit. There may also be some scope for increasing PGH resolution slightly if the scheme was ultimately required to support fine-scale discrimination across, for instance, many thousands of different objects. The redundant, multiple part nature of the PGH representation otherwise ensures that combined results are expected to be highly discriminatory for arbitrary shapes. As mentioned in the previous chapter, the incorporation of an object centroid constraint would also be beneficial for match list construction, which would conveniently support the recognition of multiple instances of specified objects.

Other potential avenues of subsequent research involve automating the process of model acquisition and PGH encoding for real objects given appropriate robotic handling equipment. There is then some scope for optimisation of the presented modelling scheme in such regards. Routines for automatically learning CAD models in formats such as VRML (Virtual Reality Modelling Language) should also be beneficial to any prospective industrial applications. There remains the issue that wireframe models may not even be required long term if PGH manifolds can be learned with enough precision to support fine-scale object localisation, verification and differentiation. Alternatively, virtual 3D surface-based object models could be inferred and associated with the learned manifolds in prospective support of any physical reasoning faculties. This would facilitate learning and recognition of free-form objects which are unamenable to analysis within the current framework. By learning and updating any PGH encoded manifolds directly from real objects, any limitations of the current manual wireframe modelling and visibility assignment processes would be bypassed. Ultimately, higher level learning processes could be implemented to prioritise any particularly distinct PGH elements for certain object views and also to structure the learned representation space so that shared PGH elements could be referenced from single instances, thus improving memory requirements and system efficiency.

An analysis of the proposed technique's applicability for recognising objects in challenging situations involving high levels of scene clutter and occlusion raises the issue of whether any visual cues other than edge geometry may be of use to such tasks. Although suggested that human vision is heavily reliant on edge-based scene analysis, it is clear that our visual systems are able to make use of other information such as colour, shading and texture. In the extreme, we may be able to hypothesise the presence of a known object in a scene, even if all its projected edge features are occluded, with reference to any fragmentary texture, surface shading or colour cues. Similar cues may also be beneficial for disambiguating any competing model match hypotheses, where sparse edge evidence alone may be of limited use for verification purposes. The requirements of any prospective artificial visual faculties ultimately depend on the assigned application domains. In terms of replicating the recognition competences exhibited by the human visual system, these other modalities should therefore be incorporated into the prospective vision system. The medium of projected edge-defined shape however remains critical to the visual recognition task, especially considering that in many cases that may be the only information available upon which to interpret a scene's contents. The PGH representation otherwise appears to fulfil all the criteria for which it was designed, taking appropriate account of all projective invariances and serving as a sound solution to the problem of oriented, spatial-frequency edge distribution encoding for 3D object recognition.

The remaining challenges facing development of artificial vision systems include learning to recognise natural and deformable objects (in all their guises), generic text recognition in support of reading, environment mapping (or multi-situated-object recognition) for navigation and then class-based and contextual recognition and inference. Ultimately, a vision system should be able to infer the likely 3D shape and appearance of a novel object from a single view via feature associations with any learned objects. Reasoning about such objects' material properties and structural characteristics should follow in support of environmental interaction, bridging the gap between vision and intelligence. The artificial emulation of human visual perception stands as a great challenge to mankind.

Appendix A

The following sections derive likelihood results for various measures used in this thesis. The results are obtained in accordance with the theoretical results outlined in TINA [1] Memo 2004-005, which regenerates what are essentially Jeffreys Priors as the correct method for parameter estimation when we have no prior information.

A.1 Bhattacharyya Similarity

In previous work, the similarity measure used for histogram comparison was justified as an approximation to a chi-squared statistic. Later, it was shown that the approach linearises the metric space for similarity between vectors of histogram values. In fact, it is possible to show that the measure is an exact form for the comparison of probability densities. The following proof is taken from TINA Memo 1999-001.

We start by considering the distance measure

$$L \;=\; 1 - \frac{(b-a)}{(4mN)} \sum_{i=1}^{N} \frac{(\sqrt{f}_i - \sqrt{h_i})^2}{var(\sqrt{f}_i) + var(\sqrt{h_i})}$$

where f_i and h_i represent frequency measures from m samples in the range a to b of a quantised variable x. The term in the sum is a maximum likelihood estimator (least squares) which assumes that $\sqrt{f}_i$ and $\sqrt{h_i}$ have Gaussian distributions and from the argument above the expected bias is N as each frequency measurement also represents an independent model parameter.

In the limit that the number of samples m becomes infinite, the estimated frequency ratios tend to conditional probabilities $P(i|f)$ and $P(i|h)$. Moreover, using the law of large numbers and error propagation it is clear that the variance terms become equal and constant [45]

$$var(\sqrt{f}_i) \;=\; var(\sqrt{h_i}) = 1/4$$

while in this limit the distributions for $\sqrt{f}_i$ and $\sqrt{h_i}$ become exactly Gaussian. In this limit we can now rewrite L as

$$L_p \;=\; 1 - \frac{(b-a)}{2N}\sum_{i=1}^{N}(\sqrt{P(i|f)} - \sqrt{P(i|h)})^2$$

Going one step further and allowing $(b - a)/N$ to tend to zero, so that the sum becomes an integration, we get.

$$L_B \;=\; 1 - \frac{1}{2}\int_a^b(\sqrt{p(x|f)} - \sqrt{p(x|h)})^2 dx$$

Where $p(x|f)$and $p(x|h)$ are probability density functions.

Rewriting this by expanding the squared term and forming three separate integrations, we get

$$L_B \;=\; \int_a^b \sqrt{p(x|f)}\sqrt{p(x|h)}dx$$

which is the Bhattacharyya measure, although this derivation from a likelihood measure is completely different to the original motivation [2].

A.2 Binomial Ratio Estimates

For the general case of estimating the most likely generating fraction μ from a finite number n from a total of N samples we write the likelihood as (see TINA Memo 2009-008)

$$Ł \;\propto\; (\mu \;-\; \mu^2)^{1/2}\mu^n(1 \;-\; \mu)^{N-n}$$

Differentiating this, we get

$$\frac{\partial L}{\partial \mu} \;\propto\; (\mu \;-\; \mu^2)^{1/2}\frac{\partial}{\partial \mu}[\mu^n(1 \;-\; \mu)^{N-n}]$$

$$+ \frac{\partial}{\partial \mu}[(\mu \;-\; \mu^2)^{1/2}]\mu^n(1 \;-\; \mu)^{N-n}$$

$$= (\mu - \mu^2)^{1/2}[nN^{n-1}(1 - \mu)^{N-n} - (N-n)\mu^n(1 - \mu)^{N-n-1}]$$
$$+ \frac{1 - 2\mu}{2(\mu - \mu^2)^{1/2}}(\mu^n(1 - \mu)^{N-n})$$

$$= [(\mu - \mu^2)^{1/2}\mu^{n-1}(1 - \mu)^{N-n-1}][n(1 - \mu) - (N - n)\mu + \frac{1 - 2\mu}{2(\mu - \mu^2)}\mu(1-\mu)]$$

The most likely generator of the data is given when this expression is set to zero. Assuming that we are not interested in solutions corresponding to $\mu = 0$ and $\mu = 1$, the most probable estimate μ' for the generator of n from N samples is given by

$$n - N\mu' + \frac{1}{2}(1 - 2\mu') = 0$$

i.e.

$$\mu' = \frac{n + 1/2}{N+1}$$

The above analysis assumes that we can obtain unambiguous integer values from a counting process. For counting noisy data below a threshold t, the logical approach, which is consistent with this idea, is to integrate the noise distribution $p(x|d)$. In this way, the soft rank of d_j can be defined as

$$\sum_i^N \int_{-\infty}^{t=d_j} p(x|d_i)dx$$

However, a bootstrap likelihood for arbitrary values of t computed on this basis would allow probabilities with values outside the $(1/2)/N$ and $(N - 1/2)/N$ limits derived above. We solve this problem by observing that the hypothesis we are testing can be defined such that the value t is itself an observation assumed to be from the required class, i.e.

$$s_t = \sum_i^N \int_{-\infty}^{t} p(x|d_i)dx + \int_{-\infty}^{t} p(x|t)dx$$

where the second term is by definition 1/2 and the ratio for a noisy binomial counting process is now $s_t/(N+1)$.

Appendix B

B.1 Further Examples of Cluttered and Occluded 3D Model Matching

The screen shots presented in Figure B.1 provide further examples of the efficacy of the proposed model matching techniques (see Chapter 7) for detecting and localising 3D objects in scenes containing mild clutter and occlusion.

In order to avoid the time complexities associated with matching extended sets of 3D model features to complex scenes, the software has been developed to allow individual (linear) image edge features to be selected and matched to the object models. For these examples, in cases where an object could not easily be model matched using limited sets of sequentially searched image edge features, model matching was performed for individually selected image edge features.

The corresponding verification scores for each model matched 3D object are tabulated in Figure B.2 according to the proposed edge location only hypothesis test with a 1% confidence limit sampled for each pixel spaced sample point along each projected feature (see Section 7.3).

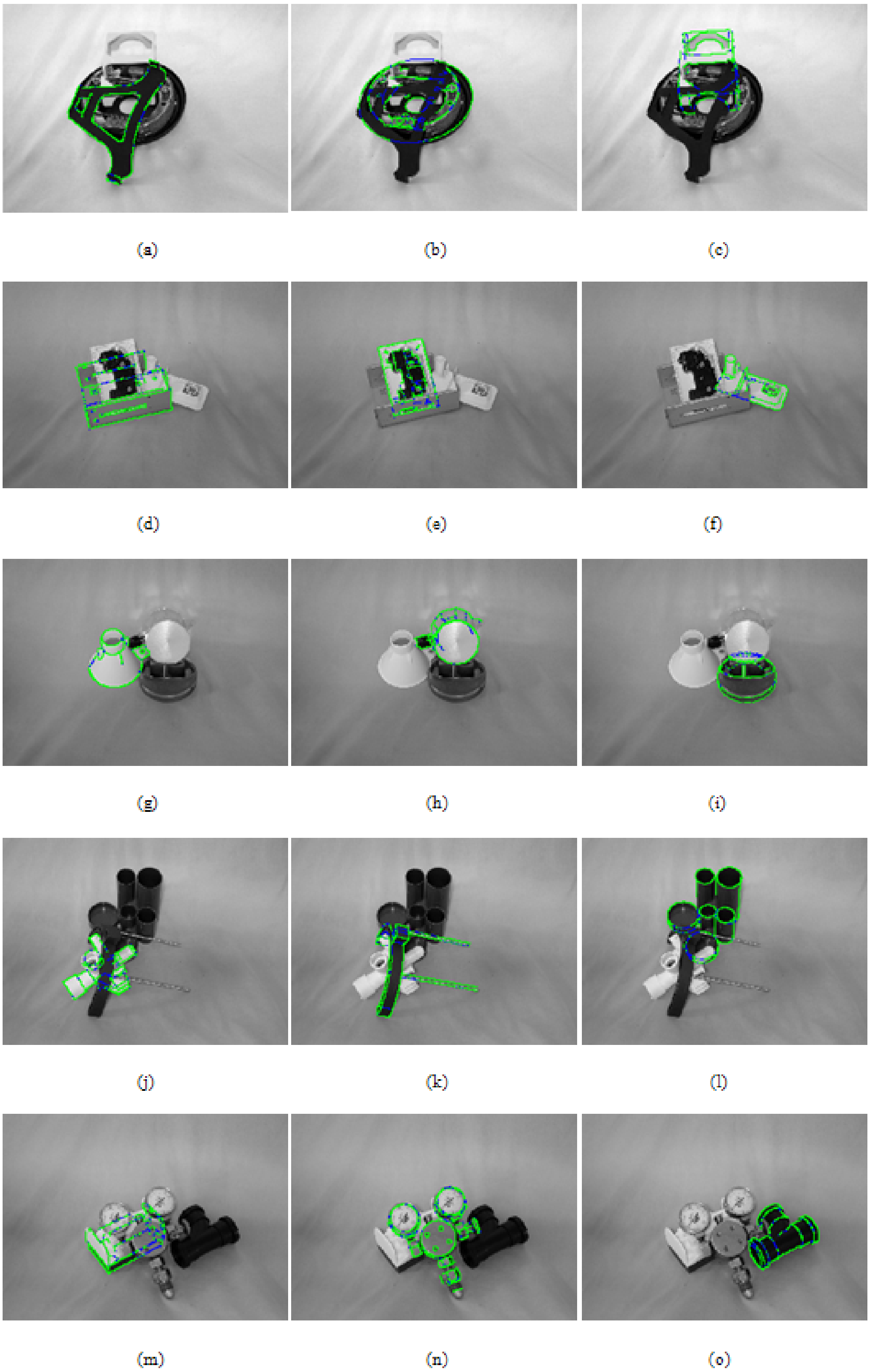


Figure B.1: *Screen shots of projected view-sampled 3D wireframe models matched to their corresponding image poses using the proposed view-based 3D model matching system.*

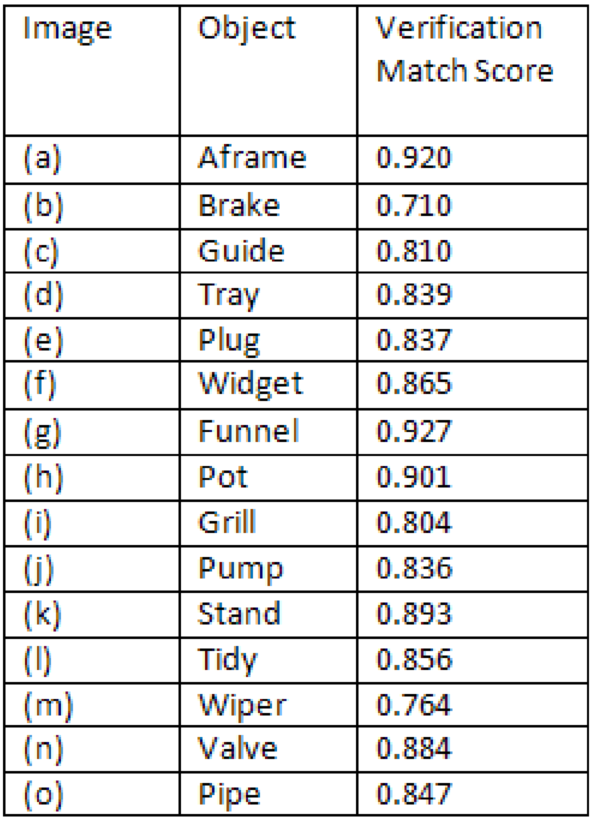


Image	Object	Verification Match Score
(a)	Aframe	0.920
(b)	Brake	0.710
(c)	Guide	0.810
(d)	Tray	0.839
(e)	Plug	0.837
(f)	Widget	0.865
(g)	Funnel	0.927
(h)	Pot	0.901
(i)	Grill	0.804
(j)	Pump	0.836
(k)	Stand	0.893
(l)	Tidy	0.856
(m)	Wiper	0.764
(n)	Valve	0.884
(o)	Pipe	0.847

Figure B.2: *Verification scores for the model matching examples provided in Figure B.1.*

Bibliography

[1] TINA Open Source Computer Vision Development Environment: http://www.tina-vision.net.

[2] A. Bhattacharyya. On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

[3] W. W. Bledsoe. Man-Machine Facial Recognition: Report on a Large-Scale Experiment. *Technical Report PRI-22, Panoramic Research Inc. California*, 1966.

[4] W. W. Bledsoe. The Model Method in Facial Recognition. *Technical Report PRI-15, Panoramic Research Inc. California*, 1966.

[5] R. Shepard and J. Metzler. Mental Rotation of Three Dimensional Objects. *Science*, 171:701–703, 1971.

[6] T. Binford. Visual Perception by Computer. In *Proc. IEEE Conference on Systems and Control*, 1971.

[7] M. Minsky. A Framework for Representing Knowledge. *The Psychology of Computer Vision. P. Winston ed., New York: McGraw-Hill*, pages 644–649, 1975.

[8] J. J. Koenderink and A. J. Van Doorn. The Singularities of the Visual Mapping. *Biological Cybernetics*, 24:145–176, 1976.

[9] J. J. Koenderink and A. J. Van Doorn. The Internal Representation of Solid Shape with Respect to Vision. *Biological Cybernetics*, 32:211–216, 1979.

[10] D. H. Ballard. Generalizing the Hough Transform to Detect Arbitrary Shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[11] S. Palmer, E. Rosen and P. Chase. Canonical Perspective and the Perception of Objects, J. Long and A. Baddeley (eds.). *Attention and performance IX, Hillsdale, NJ: Erlbaum*, pages 135–151, 1981.

[12] D. Marr. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. *Publisher: Henry Holt and Company*, 1982.

[13] P. Jolicoeur. The Time to Name Disoriented Natural Objects. *Memory and Cognition*, 13:289–303, 1985.

[14] J. Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:679–714, 1986.

[15] A. Rosenfeld. Recognizing Unexpected Objects: a Proposed Approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 1(1):71–84, 1987.

[16] D. Lowe. Three-Dimensional Object Recognition from Single Two-Dimensional Images. *Artificial Intelligence*, 31(3):355–395, 1987.

[17] I. Biederman. Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94(2):115–147, 1987.

[18] J. J. Koenderink and A. J. Van Doorn. Representation of Local Geometry in the Visual System. *Biological Cybernetics*, 55:367–375, 1987.

[19] J. Porrill, S. B. Pollard, T. P. Pridmore, J. B. Bowen, J. E. W. Mayhew and J. P. Frisby. TINA: The Sheffield AIVRU Vision System. *Proc. International Joint Conference on Artificial Intelligence*, pages 1138–1144, 1987.

[20] C. Harris and M. Stephens. A Combined Corner and Edge Detector. *Proc. 4th Alvey Vision Conference*, pages 147–151, 1988.

[21] G. Borgefors. Hierarchical Chamfer Matching: A Parametric Edge Matching Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6):849–865, 1988.

[22] Y. Lamdan and H. J. Wolfson. Geometric Hashing: A General and Efficient Model-Based Recognition Scheme. *Proc. Second International Conference on Computer Vision*, pages 238–249, 1988.

[23] M. Tarr and S. Pinker. Mental Rotation and Orientation Dependence in Shape Recognition. *Cognitive Psychology*, 28(21):233–282, 1989.

[24] D. P. Huttenlocher and S. Ullman. Recognizing Solid Objects by Alignment with an Image. *International Journal of Computer Vision*, 5(2):195–212, 1990.

[25] N. A. Thacker and J. E. W. Mayhew. Designing a Network for Context Sensitive Pattern Classification. *Neural Networks*, 3:291–299, 1990.

[26] T. Poggio and S. Edelman. A Network that Learns to Recognize 3D Objects. *Nature*, 343:263–266, 1990.

[27] W. E. L. Grimson and D. P. Huttenlocher. On the Sensitivity of Geometric Hashing. *Proc. Third International Conference on Computer Vision*, pages 334–338, 1990.

[28] M. Turk and A. Pentland. Face Recognition using Eigenfaces. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[29] S. Edelman and D. Weinshall. A Self-Organizing Multiple-View Representation of 3D Objects. *Biological Cybernetics*, 64:209–219, 1991.

[30] S. Ullman and R. Basri. Recognition by Linear Combinations of Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, 1991.

[31] C. A. Rothwell, A. Zisserman, J. L. Mundy and D. A. Forsyth. Efficient Model Library Access by Projectively Invariant Indexing Functions. *Proc. IEEE Computer Vision and Pattern Recognition Conference*, pages 109–114, 1992.

[32] H. Barlow and D. Tolhurst. Why Do You Have Edge Detectors? *Optical Society of America: Technical Digest*, 23:172, 1992.

[33] H. Bulthoff and S. Edelman. Psychophysical Support for a Two-Dimensional View Interpolation Theory of Object Recognition. *Proc. National Academy of Science USA*, 89:60–64, 1992.

[34] I. Biederman. Dynamic Binding in a Neural Network for Shape Recognition. *Psychological Review*, 99:480–517, 1992.

[35] S. Edelman and H. Bulthoff. Orientation Dependence in the Recognition of Familiar and Novel Views of Three-Dimensional Objects. *Proc. National Academy of Science USA*, 32:2385–2400, 1992.

[36] W. Press, S. Teukolsky, W. Vetterling and B. Flannery. Numerical Recipes in C. *Cambridge University Press, 2nd edition*, 1992.

[37] A. C. Evans, N. A. Thacker and J. E. W. Mayhew. A Practical View-Based 3D Object Recognition System. In *Proc. Third International Conference on Artificial Neural Networks*, pages 6–15, 1993.

[38] A. C. Evans, N. A. Thacker and J. E. W. Mayhew. The Use of Geometric Histograms for Model-Based Object Recognition. In *Proc. British Machine Vision Conference*, pages 429–438, 1993.

[39] D. P. Huttenlocher, G. A. Klanderman and W. A. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

[40] R. Bergevin and M. D. Levine. Generic Object Recognition: Building and Matching Coarse Descriptions from Line Drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(1):19–36, 1993.

[41] A. C. Evans. Geometric Feature Distributions for Shape Representation and Recognition. *Ph.D. Thesis, Department of Electronic and Electrical Engineering, The University of Sheffield*, 1994.

[42] P. A. Riocreux, N. A. Thacker and R. B. Yates. An Analysis of Pairwise Geometric Histograms for View-Based Object Recognition. In *Proc. British Machine Vision Conference*, pages 75–84, 1994.

[43] R. A. Lane, N. A. Thacker and N. L. Seed. Stretch Correlation as a Real-Time Alternative to Feature Based Stereo Matching Algorithms. *Image and Vision Computing*, 12(4):203212, 1994.

[44] A. P. Ashbrook, N. A. Thacker and P. I. Rockett. Pairwise Geometric Histograms: A Scalable Solution for the Recognition of 2D Rigid Shape. *Proc. Scandinavian Conference on Image Analysis*, pages 271–278, 1995.

[45] A. P. Ashbrook, N. A. Thacker and P. I. Rockett. The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data. *Kybernetika*, 34(4):363–368, 1995.

[46] A. P. Ashbrook, N. A. Thacker, P. I. Rockett and C. I. Brown. Robust Recogniton of Scaled Shapes Using Pairwise Geometric Histograms. *Proc. British Machine Vision Conference*, pages 503–512, 1995.

[47] C. F. Olson, D. P. Huttenlocher. Recognition by Matching Dense, Oriented Edge Pixels. *Proc. International Conference on Computer Vision*, pages 91–96, 1995.

[48] H. Murase and S. Nayar. Visual Learning and Recognition of 3-D Objects from Appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995.

[49] M. Tarr. Rotating Objects to Recognize Them: a Case Study of the Role of Viewpoint Dependency in the Recognition of Three-Dimensional Objects. *Psychonomic Bulletin and Review*, 1(2):55–82, 1995.

[50] N. A., Thacker, P. A. Riocreux and R. B. Yates. Assessing the Completeness Properties of Pairwise Geometric Histograms. *Image and Vision Computing*, 13(5):423–429, June 1995.

[51] N. A. Thacker, P .A. Riocreux and R. B. Yates. Multiple Shape Recognition Using Pairwise Geometric Histogram Based Algorithms. In *Proc. IEEE Image Processing*, 1995.

[52] T. F. Cootes C. J. Taylor, D. H. Cooper and J. Graham. Active Shape Models - Their Training and Application. *Computer Vision and Image Understanding*, 61:38–59, 1995.

[53] V. Ramesh. Performance Characterisation of Image Understanding Algorithms. *Ph.D. Thesis, University of Washington*, 1995.

[54] A. Pope and D. Lowe. Learning Appearance Models for Object Recognition. *International Workshop on Object Representation for Computer Vision, Cambridge, England. Proceedings published as Object Representation in Computer Vision II, J. Ponce, A. Zisserman and M. Hebert (eds.), Springer-Verlag*, pages 201–221, 1996.

[55] C. F. Olson and D. P. Huttenlocher. Automatic Target Recognition by Matching Oriented Edge Pixels. *IEEE Transactions on Image Processing*, 6(1):103–113, 1997.

[56] C. Huang, O. I. Camps and T. Kanungo. Object Recognition Using Appearance-Based Parts and Relations. *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 877–883, 1997.

[57] C. Schmid and R. Mohr. Local Grey-Value Invariants for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, 1997.

[58] D. J. Simons and D. T. Levin. Change Blindness. *Trends in Cognitive Science*, 1:261–267, 1997.

[59] P. Belhumeur, J. Hespanha and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[60] S. J. Dickinson, R. Bergevin, I. Biederman, J. Eklundh, R. Munck-fairwood, A. K. Jain and A. Pentland. Panel Report: The Potential of Geons for Generic 3-D Object Recognition. *Image and Vision Computing*, 15:277–292, 1997.

[61] A. P. Ashbrook. Pairwise Geometric Histograms for Object Recognition: Developments and Analysis. *Ph.D. Thesis, Department of Electronic and Electrical Engineering, The University of Sheffield*, 1998.

[62] D. J. Simons and D. T. Levin. Failure to Detect Changes to People During a Real-World Interaction. *Psychonomic Bulletin and Review*, 5:644–649, 1998.

[63] D. M. Gavrila. Multi-Feature Hierarchical Template Matching Using Distance Transforms. *Proc. 14th International Conference on Pattern Recognition*, 1:439–444, 1998.

[64] F. J. Aherne. Towards an Automatic Object Recognition Scheme. *Ph.D. Thesis, Department of Electronic and Electrical Engineering, The University of Sheffield*, 1998.

[65] J. Chen and G. Stockman. 3D Free-form Object Recognition using Indexing by Contour Features. *Computer Vision and Image Understanding*, 71(3):334–355, 1998.

[66] K. Ord and S. Arnold. Advanced Theory of Statistics: Classical Inference and the Linear Model. *Arnold*, 1998.

[67] R. Nelson and A. Selinger. A Cubist Approach to Object Recognition. *Proc. International Conference on Computer Vision*, pages 614–621, 1998.

[68] R. Nelson and A. Selinger. Large-Scale Tests of a Keyed, Appearance-Based 3-D Object Recognition System. *Vision Research: Special Issue on Computational Vision*, 38:15–16, 1998.

[69] T. Cootes, G. Edwards and C. Taylor. Active Appearance Models. *Proc. European Conference on Computer Vision*, 2:484–498, 1998.

[70] B. Huet and E. R. Hancock. Line Pattern Retrieval Using Relational Histograms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1363–1370, December 1999.

[71] B. Mel. SEEMORE: A View-Based Approach to 3D Object Recognition using Multiple Visual Cues. *Advances in Neural Information Processing Systems*, 8:865–867, 1999.

[72] D. G. Lowe. Object Recognition from Local Scale-Invariant Features. In *Proc. IEEE International Conference on Computer Vision*, pages 1150–1157, 1999.

[73] I. Biederman. Visual Object Recognition. *S. Kosslyn and D. Osherson (eds.). An Invitation to Cognitive Science, 2nd edition, Volume 2, Visual Cognition. MIT Press.*, pages 121–165, 1999.

[74] J. Beis and D. Lowe. Indexing Without Invariants in 3D Object Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):1000–1015, 1999.

[75] R. Nelson and A. Selinger. A Perceptual Grouping Hierarchy for Appearance-Based 3D Object Recognition. *Computer Vision and Image Understanding*, 76(1):15–16, 1999.

[76] S. Palmer. Vision Science. Photons to Phenomenology. *Cambridge, MA: MIT Press*, 1999.

[77] T. Cootes, G. Edwards and C. Taylor. Advances in Active Appearance Models. *Proc. IEEE International Conference on Computer Vision*, 1:137–142, 1999.

[78] A. Pope and D. Lowe. Probabilistic Models of Appearance for 3-D Object Recognition. *International Journal of Computer Vision*, 40(31):149–167, 2000.

[79] A. Lacey, N. Thacker, P. Courtney and S. Pollard. TINA 2001: The Closed Loop 3D Model Matcher. *Proc. British Machine Vision Conference*, pages 203–212, 2001.

[80] T. F. Cootes and C. J. Taylor. On Representing Edge Structure for Model Matching. *Proc. IEEE Computer Vision and Pattern Recognition Conference*, 1:1114–1119, 2001.

[81] B. Rossion and I. Gauthier. How Does The Brain Process Upright and Inverted Faces? *Behavioral and Cognitive Neuroscience Reviews*, 1(1):62–74, December 2002.

[82] M. Yang, D. J. Kriegman and N. Ahuja. Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[83] O. Carmichael and M. Hebert. Object Recognition by a Cascade of Edge Probes. *Proc. British Machine Vision Conference*, pages 103–112, 2002.

[84] K. Mikolajczyk, A. Zisserman and C. Schmid. Shape Recognition with Edge-Based Features. *Proc. British Machine Vision Conference*, pages 779–788, 2003.

[85] P. A. Bromiley, M. L. J. Scott, M. Pokric, A. J. Lacey and N. A. Thacker. Bayesian and Non-Bayesian Probabilistic Models for Magnetic Resonance Image Analysis. *Image and Vision Computing, Special Edition; The use of Probabilistic Models in Computer Vision*, 21:851–864, 2003.

[86] W. Hayward. After the Viewpoint debate: Where Next in Object Recognition? *Trends in Cognitive Sciences*, 10(7):425–427, 2003.

[87] B. A. Draper, K. Baek and J. Boody. Implementing the Expert Object Recognition Pathway. *Machine Vision and Applications*, 16(1):27–32, December 2004.

[88] I. M. Scott. Searching Image Databases Using Appearance Models. *PhD Thesis, The University of Manchester*, 2004.

[89] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[90] X. Teng, Y. Liu and C. Liu. AAM Based Matching of Hand Appearance for User Verification. *Lecture Notes in Computer Science: Advances in Biometric Person Authentication, Springer Berlin*, pages 690–695, 2004.

[91] E. C. Leek, I. Reppa and M. Arguin. The Structure of Three-Dimensional Object Representations in Human Vision: Evidence from Whole-Part Matching. *Journal of Experimental Psychology: Human Perception and Performance*, (31):668–684, 2005.

[92] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[93] D. Cristinacce and T. F.Cootes. Feature Detection and Tracking with Constrained Local Models. *Proc. British Machine Vision Conference*, 3:929–938, 2006.

[94] F. Rothganger, S. Lazebnik, C. Schmid and J. Ponce. 3D Object Modeling and Recognition Using Local Affine-Invariant Image Descriptors and Multi-View Spatial Constraints. *International Journal of Computer Vision*, 66(3):231–259, 2006.

[95] H. Ling and K. Okada. Diffusion Distance for Histogram Comparison. *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:246–253, 2006.

[96] S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2006.

[97] H. Ling and K. Okada. An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:840–853, 2007.

[98] N. A. Thacker and E. C. Leek. Retinal Sampling, Feature Detection and Saccades; A Statistical Perspective. *Proc. BMVC*, pages 990–999, 2007.

[99] P. J. Philips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott and M. Sharpe. FRVT 2006 (Face Recognition Vendor Test) and ICE 2006 (Iris Challenge Evaluation) Large-Scale Results. *National Institute of Standards and Technology (USA) 7408*, 2007.

[100] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool. SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[101] D. Hubel. Eye, Brain and Vision (eBook): http://hubel.med.harvard.edu. 2009.

[102] Wikipedia The Free Encyclopedia: http://www.wikipedia.org. 2009.