# The Application of Range Imaging for Improved Local Feature Representations

Euan Strachan

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy
School of Computing Science
College of Science and Engineering
University of Glasgow

October, 2012

# Abstract

This thesis presents an investigation into the integration of information extracted from co-aligned *range* and *intensity* images to achieve pose invariant object recognition. Local feature matching is a fundamental technique in image analysis that underpins many computer vision-based applications; the approach comprises identifying a collection of *interest points* in an image, characterising the local image region surrounding the interest point by means of a *descriptor*, and matching these descriptors between example images. Such local feature descriptors are formed from a measure of the local image statistics in the region surrounding the interest point. The interest point locations and the means of measuring local image statistics should be chosen such that resultant descriptor remains stable across a range of common image transformations. Recently the availability of low cost, high quality range imaging devices has motivated an interest in local feature extraction from range images. It has been widely assumed in the vision community that the range imaging domain has properties which remain quasi-invariant through a wide range of changes in illumination and pose. Accordingly, it has been suggested that local feature extraction in the range domain should allow the calculation of local feature descriptors that are potentially more robust than those calculated from the intensity imaging domain alone. However, range images represent differing characteristics from those represented within intensity images which are frequently used, independently from range images, to create robust local features. Therefore, this work attempts to establish the best means of combining information from these two imaging modalities to further increase the reliability of matching local features.

Local feature extraction comprises a series of processes applied to an image location such that a collection of repeatable descriptors can be established. By using co-aligned range and intensity images this work investigates the choice of modality and method for each step in the extraction process as an approach to optimising the resulting descriptor. Additionally, multimodal features are formed by combining information from both domains in a single stage in the extraction process. To further improve the quality of feature descriptors, a calculation of the surface normals and a use of the 3D structure from the range image are applied to correct the 3D appearance of a local sample patch, thereby increasing the similarity between observations.

The matching performance of local features is evaluated using an experimental setup comprising a turntable and stereo pair of cameras. This experimental setup is used to create a database of intensity and range images for 5 objects imaged at 72 calibrated viewpoints, creating a database of 360 object observations. The use of a calibrated turntable in combination with the 3D object surface coordiantes, supplied by the range image allow location correspondences between object observations to be established; and therefore descriptor matches to be labelled as either true positive or false positive. Applying this methodology to the formulated local features show that two approaches demonstrate state-of-the-art performance, with a ~40% increase in area under ROC curve at a False Positive Rate of 10% when compared with standard SIFT. These approaches are range affine corrected intensity SIFT and element corrected surface gradients SIFT.

Furthermore,this work uses the 3D structure encoded in the range image to organise collections of interest points from a series of observations into a collection of *canonical views* in a new model local feature. The canonical views for a interest point are stored in a *view compartmentalised* structure which allows the appearance of a local interest point to be characterised across the view sphere. Each canonical view is assigned a confidence measure based on the 3D pose of the interest point at observation, this confidence measure is then used to match similar canonical views of model and query interest points thereby achieving a pose invariant interest point description. This approach does not produce a statistically significant performance increase. However, does contribute a validated methodology for combining multiple descriptors with differing confidence weightings into a single keypoint.

# Acknowledgements

I would like to thank my PhD supervisor Dr J.Paul Siebert for all the hard work he has put in over the past 4 years to shape my way of thinking about problems, reasoning about results and communicating the principal ideas behind my work. I certainly have not found the journey to completing this thesis easy, and at times the belief he has given me in my work has been all that has seen me through. I would like to thank my second supervisor Professor Joemon Jose, for helping me keep sight of the goal of finishing the work and structuring a plan to ensure that it happened. Thanks also to the EPSRC for funding the work carried out in this thesis and to the examiners for their time and effort in reviewing. All of my office mates and members of the Computer Vision and Graphics Group; our conversations on topics outside of PhD life which somehow worked their way back to the current problems that we were facing in our work definitely helped me step back from the smaller problems and see each as part of a bigger picture. Likewise, to all my friends in the Mountaineering Club; the fresh air and hillwalking trips cleared my head and allowed me to keep going. To my brother, Andy, for his support in living in Glasgow. To Laura Wright for all her love, for standing by me and always believing that I would be able to complete this work. Lastly I would like to thank my father, Mel, for having always encouraged me to reason about the world and to pursue what I could achieve, I could not have hoped for a better role model.

# Contents

# List of Figures

# List of Tables

# List of Algorithms

# Chapter 1

# Introduction

The main aim of this research project was to investigate the role of range and intensity imaging modalities in point based correspondences between imaged representations of objects. As an approach to investigating the role of range and intensity imaging in a local feature representation this thesis adopts the context of the  local feature matching. Local features have been widely adopted by the computer vision community as an approach for formulating local feature representations for a range of processing tasks. Such features comprise a distribution of intensity image properties such as gradient magnitudes and orientations of a limited area surrounding an *interest point*, comprising a corner or similar compact 2D discontinuity. However, standard local feature approaches have known limitations for matching features under changes in the 3D pose of the object. In this thesis the use of multimodal information from the range and intensity domains are used as an approach for mitigating the effects of changes in view point. Following an initial investigation of modifications which can be made directly to the existing feature extraction process to accommodate range domain information, the local feature extraction processing pipe-line is extended to incorporate additional processing stages which rely on the range domain information for structuring a view independent feature descriptor. As part of the investigation an experimental procedure is outlined, whereby range and intensity images of real-world objects are captured using a stereo camera pair and calibrated turntable configuration.

## 1.1 Aims and Objectives

A range image is a matrix where each pixel encodes the distance from the camera to each point on the object imaged. This distance or range value is recorded in place of an intensity value in a conventional digital image (Besl and Jain, 1986). The range image representation encodes a measure of the 3D structure of an object. Therefore, the range domain information has been noted to have the potential to be more invariant to a variety of common transformations observed in images such as, illumination changes and 3D pose changes (Gordon, 1992). However, the range imaging modality differs in characteristics from the intensity domain and many approaches which utilise the range imaging modality do not realise the expected performance benefit (Pears et al., 2010). Additionally, by using a stereo pair camera configuration for range image capture it is possible to capture co-aligned images in both the range and intensity modalities. Combining information from these domains has demonstrated improved recognition rates in face recognition (Faltemier et al., 2007), although there has been limited interest in incorporating multimodal information within local features. This thesis addresses the application of multimodal information within local feature representations.

### 1.1.1 Scientific Questions

At the beginning of this research project, the following questions were posed:

- How can range and intensity domain representations of an object be integrated together in order to form a consistent description?
  - What possibilities are there for range information to be utilised in feature extraction?
  - Is it possible to create locations in images where both range and intensity information are diagnostic?
  - How can the optimum representation combining features from range and intensity information be formulated?
- How can a standardised test to evaluate the performance of keypoints on 3D free form objects under *out-of-plane* transformation changes be devised? See Figure 1.1 for an example of the challenges of recognition under out-of-plane transformations in viewer perspective changes.

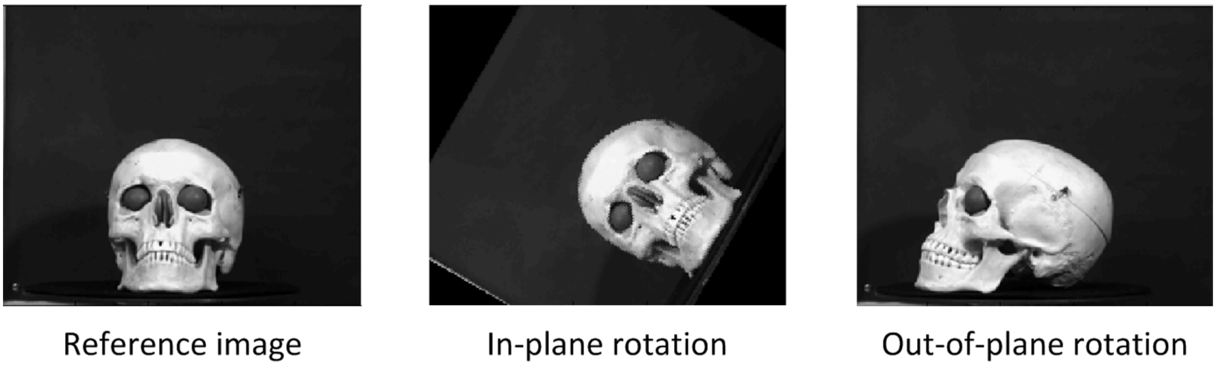


Figure 1.1: Example of in-plane and out-of-plane object rotations. In-plane rotation maintains visual similarity with a linear transformation between pixels and pixels in the reference image. Out-of-plane rotation no-longer demonstrates a pixel-wise correspondence with the reference image and, due to occlusions, includes differing information from reference image.

  - How can correspondences between pixel locations in range images be established between observations modified by out-of-plane viewer pose changes?

- Are there additional stages for keypoint extraction which can be introduced to exploit the characteristics of the range domain?

### 1.1.2 Motivation

Range images or 2.5D depth images are frequently cited as a representation with characteristics which are better suited for matching under changes in 3D pose and illumination (Lo and Siebert, 2009; Lo, 2009; Pears et al., 2010). Based on the assumption that range data is more stable than intensity data and the recent availability of 3D scanning devices, such as the Microsoft Kinect (Janoch et al., 2011), laser scanning (Taati and Greenspan, 2011) and stereo capture (Siebert and Marshall, 2000; Calonder et al., 2008; Hartley and Zisserman, 2005), there have been many proposed 2.5D local surface descriptors. Many of these descriptors extend 2D local feature matching into the range image domain using a similar structure for feature extraction, and exploit range image data at the surface representation level (Ohbuchi et al., 2008). Other techniques have approached the problem by extending 3D local feature matching into the 2.5D range domain for matching point clouds, with the assumption that the range image is a close approximation to the surface (Kofman and Knopf, 1999). In addition to the above there are systems which are capable of using a combination of 2D intensity image data and 2.5D range image data to form a multimodal feature descriptor (Mian et al., 2008; Bowyer et al., 2006). The main goal of this work is to produce a high quality keypoint descriptor that can be used for finding point-to-point correspondences between 3D free form objects, for tasks such as rigid body matching, range map registration, object recognition (Ohbuchi and Furuya, 2009) and object landmarking. Range images are an important class of 3D representations as these comprise the fundamental output generated by 3D triangulation systems, and have been suggested by many to offer greater object representational stability (Pears et al., 2010; Gordon, 1992). However range data is challenging to interpret by means of local feature descriptors as the modality only provides partial 3D information, while sharing many of the characteristic of 2D intensity images, such as: occlusions, perspective distortion and a lack of 3D connectivity. Range images also inherit the noise characteristics of the 2D intensity images from which they have been constructed. In this thesis the principal objective is to answer the question: what is the most effective use of range and intensity image data to achieve greater invariance in feature matching under changes in 3D pose given the afore mentioned limitations?

## 1.2 Background

Local feature matching is a widely adopted approach for determining correspondences between images for a variety of applications. Local features comprise a set of localised interest points in an image of an object and a collection of associated measurements from a limited region surrounding the interest points which forms a *feature descriptor*. Image matching and analysis by local feature approaches are processed using matches between sets of local interest points collected from examples and query images. As analysis is achieved by the matching of a number of local interest points between images, analysis may be achieved in complex or cluttered scenes as the process relies on determining statistically a number of good interest points matches. However, to achieve successful analysis through local feature matching, the feature descriptor representation of a local interest point must remain invariant to a range of possible image transformations. Additionally the feature descriptors must also be distinctive enough to reduce ambiguity between instances of local interest points which do not represent the same object location.

Therefore, surface representations and descriptor extraction processes are key to forming robust local features. Progressing the representation of local features in this respect will improve the performance of any application utilising local features. This goal is addressed in this thesis through the inclusion of range and intensity domain information.

### 1.2.1 Keypoint performance evaluation

The performance of local feature matching systems is difficult to establish without ground truth for full 3D out-of-plane motion. In a number of reported evaluation systems this problem is avoided by evaluating the performance of local feature matching systems stochastically on images as a whole (Lai et al., 2011; Van De Sande et al., 2010; Janoch et al., 2011). However, this approach does not account for the modularity of many local feature matching systems, where the local intensity or surface descriptor may perform well, although the overall performance is degraded by a poor selection of local feature locations; or when a post processing stage which can be applied to both local feature matching systems is responsible the performance increase. The focus of this thesis is on improving the repeatability of the surface descriptors used for local feature matching. Here, the stochastic approach is not appropriate as it shows the result of matching the corpus of features present in an image, as opposed to the point-to-point correspondence performance. Other systems have been proposed which do evaluate the point-to-point correspondence performance of local feature matching systems. In these systems the performance to out-of-plane rotation is inferred from the performance under affine transformations (Mikolajczyk and Schmid, 2004, 2005; Schmid et al., 2000). However, this assumption causes the evaluation system to be inherently biased towards local features which correct for affine transformations, and discriminate against the cases where the local range data is highly distinctive, the very cases that are of interest in this thesis. The use of turntables and 3D geometry has been exploited to evaluate the performance of local descriptors under 3D out-of-plane orientation changes (Moreels and Perona, 2007, 2004). Using local feature matching between stereo image pairs and the constraints of epipolar geometry to find the ground truth for matched keypoints in a 3rd query image, the effectiveness of the local feature matching system can be evaluated. Winder and Brown have used a similar approach in unconstrained outdoor scenes to investigate sample patch arrangements (Brown et al., 2011; Winder and Brown, 2007). While this approach gives a clearer indication of the performance of local feature descriptors, it however discards around 30 to 50% of the keypoints due to failed matches. Therefore, in Winder and Browns work the performance of the failed matches cannot be evaluated. By adopting range imaging of the scene it is however possible to find the transformation of all points on an object applied by the turntable deterministically without the use of local features, this approach additionally offers the opportunity to evaluate the performance of 2.5D local features, and the multimodal combination of 2.5D range features with 2D intensity features. Therefore, one of the main contributions of this work is an evaluation system framework capable of applying a known transformation to an object and capturing range and intensity images together with the 3D motion of points on the object surface when rotated.

## 1.3 Overview of General Approach

Figure 1.2, gives an overview of the SIFT pipeline and the stages where modifications are made in this thesis to accommodate range domain information. This section gives an overview of the general approach adopted in this thesis. Stereo pair images are captured and processed using a stereophotogrammetry package, C3D, to create co-aligned pairs of range and intensity images (Ju et al., 2003). Keypoints are detected in a scale space representation derived from both range and intensity images using the scale space corner detector methodology from SIFT (Lowe, 2004, 1999). The combination of keypoint locations from both imaging domains is investigated as a means to finding consistent locations on an object which may be detected within different views. From the resulting collection of keypoint locations, a statistical measure of the local surface surrounding the keypoint in the range and intensity images is derived by subsampling the measurement region of the keypoint with a spatial arrangement of *receptive fields* (Schiele and Crowley, 2000). The spatial subsampling of the keypoint local area with receptive fields allows the composition of the local area surrounding a keypoint to be encoded. An investigation into the optimum receptive field arrangement is conducted.

In addition to subsampling the local region surrounding a keypoint, pose normalisation of the whole sample patch may be applied, such that the keypoint region sampling aperture consistently covers the same region between observations. The pose normalisation applied to the sample patch may be based on measures of the intensity or range surface surrounding the keypoint. Using the range surface as an estimation of the 3D object surface, the transformation applied to the sampling patch was corrected to sample the range and intensity images to compute the pose corrected appearance of the keypoint from the surface normal viewing angle.

The choice of surface measures used to create a feature descriptor will determine the performance of feature descriptor matching under pose changes. To investigate this, a number of possible range and intensity image surface measures and their combinations are proposed for computing a feature descriptor.

Post processing of extracted features is investigated as a means to create a set of keypoints which describe the variation of feature descriptors across a view sphere. In order to investigate post processing of feature descriptors two approaches are presented: the first approach attempts a statistical characterisation of keypoints in feature descriptor space, to give a space of variation in which keypoints may be match. A second approach is applied whereby the observation space of the keypoint is partitioned into a series of compartments, each containing a feature descriptor expression of the keypoint from a given view and a confidence measure.

To validate the approaches proposed in this thesis, a test set of co-aligned range and intensity images were collected using a calibrated stereo camera pair and turn-table configuration. Using this configuration the location of re-occurring keypoints between observations of an object under a pose change has been established.

## 1.4 Contributions

This thesis makes the following key contributions:

- Formation of a methodology for the capture and position control of 3D free form objects under out-of-plane orientation changes. This is used to create a database of range and intensity images with known transformations.

- A methodology utilising the database of range and intensity images with known transformations, for evaluating keypoint matching under pose changes.

- The evaluation of range data implemented in several proposed keypoint matching approaches.

- The proposal of an extension to the SIFT feature extraction structure to accommodate range domain pose information.

The work presented in this thesis has appeared in:

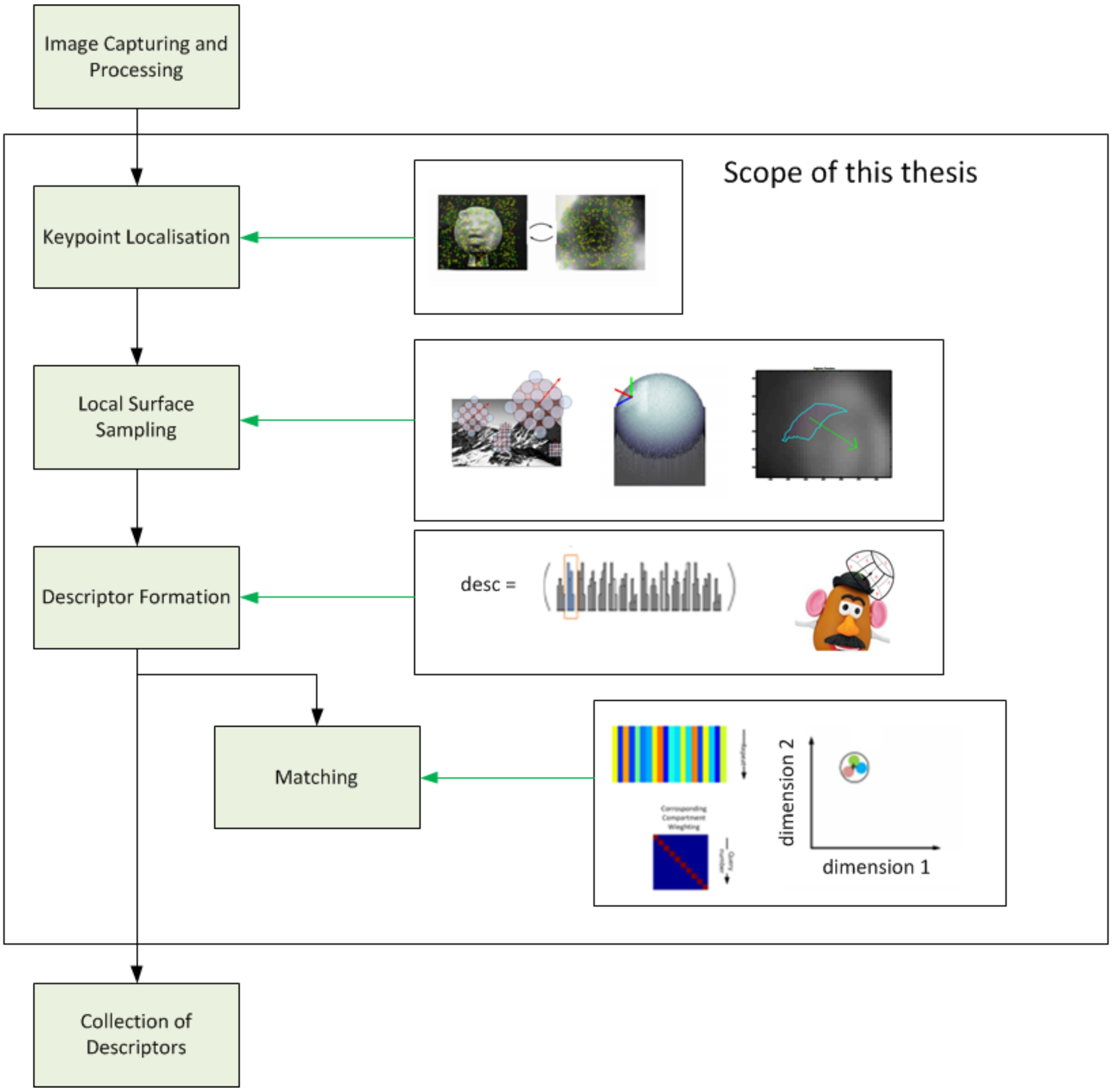


Figure 1.2: Scope of Thesis

- **Euan Strachan and J.Paul Siebert:** 2.5D local feature matching system, *In Proc. of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum*, April 2011, Vasteras, Sweden
- **Euan Strachan and J.Paul Siebert:** Local Multi-Modal SIFT Features in Co-registered Range and Intensity Images, *International Conference in Signal and Image Technologies and Internet Based Systems*, November 2012, Sorrento, Italy
- **Euan Strachan and J.Paul Siebert:** A View Based Approach for matching the 3D Appearance of Local Features, *International Conference on Image Analysis and Recognition*, June 2013, Póvoa de Varzim, Portugal

## 1.5 Hypothesis

This thesis argues that it is possible to achieve a performance improvement over existing local feature matching approaches by exploiting information from co-aligned range and intensity domain images. Range images offer a partial representation of the 3D object surface which can allow keypoint feature descriptors to encode further information regarding the 3D appearance of the keypoint, which is unavailable when using only the intensity imaging modality. This additional information regarding the object structure can be encoded together with the intensity domain information to form a robust local feature descriptor for an image location.

## 1.6 Overview of Thesis

This thesis is organised as follows: Chapter 2 gives a background and literature review for this research project, covering imaging modalities and recognition approaches; Chapter 3 evaluates the influential design choices in SIFT features for matching in 2D images; Chapter 4 extends the experimental methodology presented in Chapter 3 to capture range and intensity images of 3D free form objects under out-of-plane orientation changes; Chapter 5 examines the implementation of range domain information in the standard SIFT pipe-line; Chapter 6 uses multiple observations of an object to create a set of SIFT descriptors for an object which account for descriptor variation introduced by pose changes, the resulting characterisation is used to match to any single observation; Chapter 7 creates local features which partition the view sphere based on a local pose estimation from the range image; Chapter 8 overviews the contributions of this research project, draws conclusions from the findings and suggests future work.

# Chapter 2

# Background and Literature Review

The aim of this research project is to exploit the use of range images to improve local feature invariance, to exploit the cross modal information from range and intensity images and to model the appearance change in imaged representations of objects across time due to pose changes. As such this literature review chapter first establishes the range imaging modality, its formulation, characteristic and context in modern applications. It then reviews the goals of image representations, the progress which has been made towards characterising underlying object information and other approaches where a limited range of deformations can be accounted for and characterised. This chapter concludes by identifying a possible avenue of research where range data can contribute to the performance of object recognition.

## 2.1 Introduction

This chapter introduces the key paradigms in the current literature for 3D free form object recognition . From this corpus, local features are shown to be an important low level image representation for many generalised tasks in computer vision. These low level image features typically form the input to higher level reasoning algorithms (Freeman et al., 2008) which are in turn used as sensor data for the completion of many autonomous tasks (Meger et al., 2008). With many high quality and inexpensive cameras available much of the literature addresses low level feature extraction from intensity based images. Whereas, other imaging systems such as lidar have created images in the range image domain, these have different characteristics from intensity based images and new approaches have been required to create appropriate representations. More recently, imaging systems capable of extracting aligned intensity and depth images, such as stereo imaging, provide simultaneous information regarding the intensity and underlying 3D structure present in the scenes imaged. However, in all camera based approaches, the initial stages involve the extraction of high quality information from image data to form a mathematical description of the image contents (Szeliski, 2010; Gonzalez and Woods, 1992). The task of forming a mathematical description of the contents of an image is challenging, as vision approaches are required to be sensitive to changes in the image relating to the image contents, while remaining robust to changes in the image resulting from observation conditions. This chapter first identifies the characteristics of images in each domain, then focuses on the constraints and design choices made by others to create invariant object representations. The chapter concludes by identifying areas in the existing literature where progress may be made.

## 2.2 Imaging Modalities

To characterise a more fundamental representation of a scene which can in turn be useful for processing tasks, we must first define the initial means of representation. In this thesis a scene is defined as a physical environment which we desire to sense or measure with a computer vision system. In order to sense and reason about the environment, or scene, a visual representation must be constructed out of observable measurement signals, such a representation is referred to as an image. As recognition is an important problem to solve in the case of almost any signal, and as most approaches involve forming fundamental representations, there is an abundance of literature offering solutions for differing signal domains (Zeng et al., 2009; Dror et al., 1995; Fanelli et al., 2010). Many of these approaches have influenced the field of computer vision and pattern recognition which is principally concerned with the representations discussed in this chapter. This chapter begins by covering the main scene representations, and their characteristics, discussed in the literature.

By far the most common image representation is the 2D intensity image. This is a 2D matrix which records the intensity value of light focused on an array of optical sensors. Each element in the 2D matrix is known as a pixel. This rectilinear arrangement of measurements gives a regular structure to the representation of the object on the imaging plane. Cameras which record 2D intensity images are cheap, common and produce high quality data. Typically 2D intensity images have additional colour information associated with the intensity value for each pixel. Resources such as Google Image Search, Pisca and Flikr have allowed large databases of examples of scenes in intensity images to be created (Hays and Efros, 2008; Ponce et al., 2006). However, due to the projective nature of the light focused on the optical sensor, intensity images are a perspective representation of an object, unique to the object viewing angle. The perspective intensity images are a complex combination of object information and additional information regarding the object pose, lighting, and setting as well as suffering from degradations such as occlusions, noise and clutter. More fundamental scene representations resulting from imaging in the intensity domain are challenging as this additional information cannot be easily disentangled (Palmer, 1999).

Other representations exist which give a more intuitive representation of an object or scene, such as 3D meshes or 3D computer aided design drawings (Turk and Levoy, 1994; Krishnamurthy and Levoy, 1996); where the whole object is represented as it exists in 3D space independent of viewing perspective. These approaches represent an object via a series of points joined together to form polygons defining the local connectivity constraint. Mesh and CAD approaches are capable of giving a whole view of an object. However, creating 3D mesh or CAD drawing representations requires either an interaction with the object to fully explore the view space, or that the object is manually created, as such a single observation instance is insufficient to create a full 3D representation (Bustos et al., 2005). In addition to meshes, 3D point clouds have recently become a popular representation (Rusu and Cousins, 2011), these are representations which store the series of point locations without the connectivity information. Both representations can be created from the integration of partial 3D measurements from range devices, or from multiview geometry. The individual measurements from a range imaging device can be used to create a point cloud of X, Y, Z locations, which can be incorporated across measurements in an approach such as Simultaneous Localisation and Mapping, SLAM (Calonder et al., 2008). In recent years SLAM approaches have been able to form 3D point clouds and mesh models from time series collections of intensity images (Parsley and Julier, 2008). Meshes representations typically have complicated surface topology where an even sampling of the surface manifold in not possible in three dimensions. In addition to this the points used to create the polygon surface are unevenly distributed and there is no standard means by which to create a unique mesh structure for any given object without imposing a landmarked case for an object class.

A compromise between distal and proximal representations can be reached using the range image domain. Range images are a 2D matrix, similar to the 2D intensity image, where each pixel in the image encodes the distance from the range imaging device to a projected point on the object imaged. Range images have similar characteristics to 2D intensity images and are based on a projective camera model. Range images are therefore a limited 3D model, or a 2.5D model of the object imaged. They lack information regarding 3D structure unavailable from the given viewpoint. However, range images do give explicit 3D compositional information and illumination invariance absent in 2D intensity images (Fanelli et al., 2010). As range and intensity images are both in matrix form these can be co-aligned to represent a new imaging modality which has so far been under-utilised.

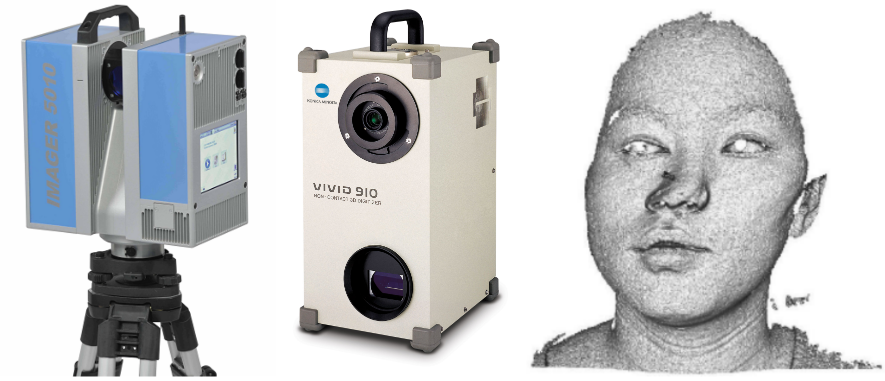
To frame the characteristics of the range and intensity domain data used in this thesis, this section is arranged as follows. Firstly a selection of scanners which acquire range data are introduced to highlight the general availability of the range medium and the requirement to process the resultant data; section 2.2.1 covers the specific details of the capture of range domain data by means of stereo imaging; section 2.2.2 identifies the characteristics of range and intensity data and sources of noise for the stereo image configuration; section 2.2.3 then gives the configuration of the stereo capture rig used in the laboratory for the capture of all range data used in this thesis.

### 2.2.1 Commercially available range imaging devices

The aim of this section is to show the wide spread availability of range images capture using commercially available devices. The characteristics of each approach are discussed as well as introducing some datasets created with each. Common issue across all range imaging devices are: image noise; and surface specularity (Cyganek and Siebert, 2011). Specific issues experienced when generating range images through triangulation approaches additionally include: occlusions; and failure to localise a pixel match (Cyganek and Siebert, 2011).

#### 2.2.1.1 LIDAR laser scanners

Laser scanners such as the Minolta Vivid 910 or Z+F IMAGER 5010, offer co-aligned colour and range data with highly accurate distance values. These scanners are an active triangulation approach



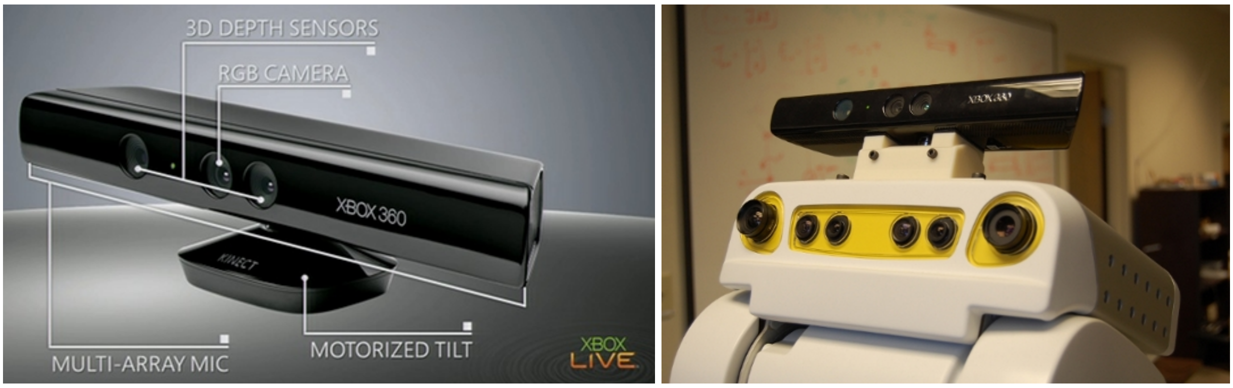
(a) Z+F IMAGER 5010 (b) Minolta Vivid 910 Laser Scanner (c) Distortions in the FRGC v2

Figure 2.1: Laser Scanner

where a laser scan line is projected from a known location and the intersection of this line and the object can be used to calculate the distance. The laser is powerful enough for outdoor use. The scanners have been used to form datasets of range and intensity images such as the Notre Dame Facial Recognition Grand Challenge, FRGC v2, dataset (Faltemier et al., 2007) and as ground truth in the EPFL multiview stereo benchmark (Calonder et al., 2008). However, it has been noted that errors exist in laser scanning data due to object motion during the time taken for the laser to scan the object (Boehnen and Flynn, 2009). Figure 2.1a shows the Z+F IMAGER 5010 used to create the EPFL multiview stereo benchmark dataset (Calonder et al., 2008). Figure 2.1b shows the Minolta Vivid 910 laser scanner used for the FRGC v2 dataset. Figure 2.1c, gives an example of distortion in a range image resulting from object motion during the scan process (Boehnen and Flynn, 2009).

#### 2.2.1.2 LIDAR time of fight scanners

Time of flight cameras work using the same principle as RADAR, i.e. by illuminating a scene with light of a given frequency and measuring the time taken for the light to return when it is reflected by a surface. For a single observation these methods typically have a range resolution error of 1cm, due to noise in the measurements and the short times differences when the distances involved are small (Cui et al., 2010).



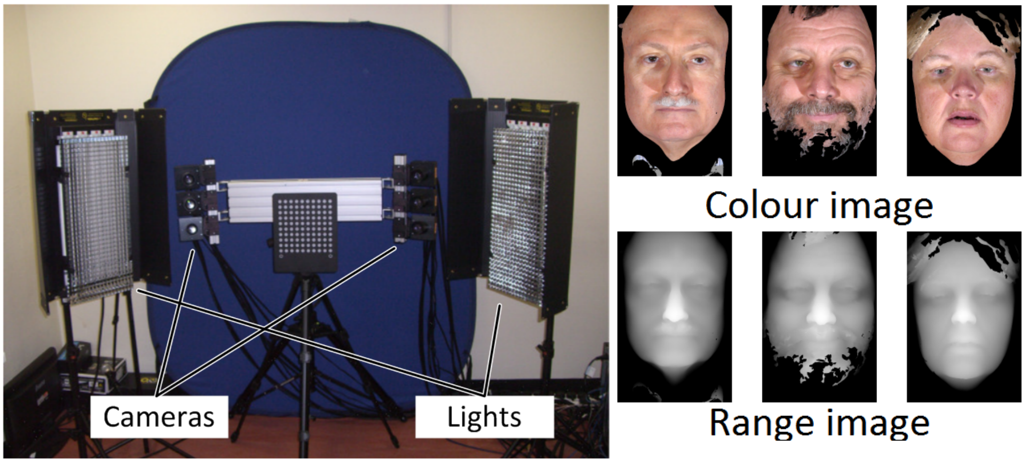
(a) Kinect Diagram (b) Kinect mounted on a Willow Garage PR2 robot platform

Figure 2.2: Microsoft Kinect

#### 2.2.1.3 Structured Light

The most common example of a structured light device for creating depth representations of objects or scenes is the Microsoft Kinect. The Kinect illuminates its field of view with a structured pattern in infra red. From the distortions of the imaged infra red light pattern, a depth image can be created. The Kinect has an additional colour camera to capture colour images of the scene, allowing the device to capture a range, colour and intensity domain representation. The Kinect has been widely adopted by the vision community and has open source support through OpenCV (Willow-Garage and Intel, 2011), the Point Cloud Library (Rusu and Cousins, 2011) and ROS (Conley, 2012). The Kinect has the ability to create range and intensity image representations in real-time, offering the additional time dimension to potential representations. A large database of range and intensity images for a range of common objects and object poses collected with the Kinect is available from Washington State University (Lai et al., 2011).

The Kinect does however have known limitations. The resolution of the camera and range sensor is only 640x480 pixels giving a low resolution image. Due to the need to detect the projected infra red pattern, the Kinect is limited to indoor use. There is not a direct correspondence between pixels in the range and intensity images, as these are captured by cameras or sensors at differing locations, as such the captured range and intensity images are only roughly aligned. Figure 2.2a shows the setup of the Kinect, Figure 2.2b shows the Kinect in use on a Willow Garage PR2 robot.



(a) Dimensional Imaging Stereo Capture System

(b) Examples of Stereo holes in Texas Dataset (Gupta et al., 2010b)

Figure 2.3: Stereo Imaging

#### 2.2.1.4 Stereo Capture

Stereo capture rigs are a passive range measurement device which can be created from standard off the shelf digital cameras. This device is capable of forming high quality range images with co-aligned texture images. Stereo capture is however a computationally expensive process, as the displacement of every pixel with respect to each camera must be calculated. Recent use of GPUs and machine learning approaches have allowed this processing to occur in real-time (Mei et al., 2011; Di3, 2008). Stereo capture rigs rely on triangulation of points imaged in the intensity images from each camera, and typically perform poorly when this correspondence cannot be established such as with hair. As stereo imaging uses 2 images captured simultaneously, the approach is robust to object motion distortions affecting laser scans. However motion blur present in images does affect the stereo match quality. Figure 2.3b, shows range images from the Texas dataset, the range values for the human hair in these range images cannot be established (Gupta et al., 2010b). Figure 2.3a, shows a Dimensional Imaging Ltd. stereo imaging device capable of capturing range, intensity and colour time series data (Matuszewski et al., 2011).

Stereo photogrammetry is the means of range imaging used in this thesis. As such the remainder of this section is focused on the specific characteristics of range imaging by means of stereo photogrammetry and the limitations of range images produced.

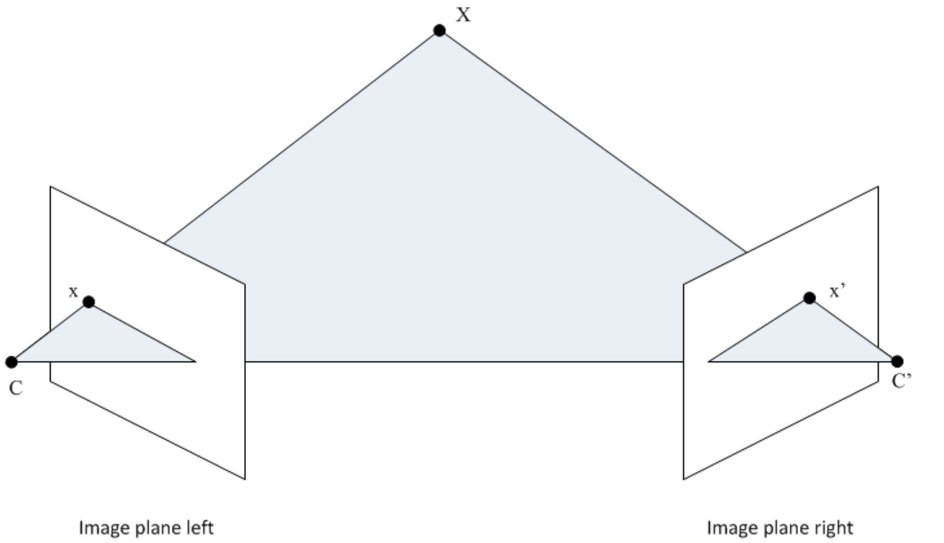


Figure 2.4: Point Correspondence Geometry (Hartley and Zisserman, 2005)

### 2.2.2 Range image formulation

Range images are a matrix of distance values representing the distance from the perspective centre of a camera, through which all rays pass, to the point on the object imaged by the pixel of the given camera. The range images in this thesis are formed using a stereo capture configuration. The images are created in a 4 step process outlined in this section. A more complete study involving an in-depth analysis of the stereo vision approach used in this thesis can be found in (Siebert and Marshall, 2000; Cyganek and Siebert, 2011; Ju et al., 2003; Urquart, 1997). An introduction to stereo vision can be found in (Trucco and Verri, 1998), and a more general overview of the field of multiview geometry can be found in (Hartley and Zisserman, 2005).

Figure 2.4 shows the standard approach for establishing the 3D location of points in a scene using stereo camera configuration (Hartley and Zisserman, 2005). In this figure, X denotes the 3D location of a point in a scene, x denotes the imaged location of the given point, and C denotes the perspective center of the camera. The convention for geometry symbols used in this thesis is: [X, Y, Z] denotes a 3D location in a scene, and [x, y] denotes a pixel location in an image.

1. **Camera Calibration**; each camera in the stereo configuration is calibrated with respects to the other. The calibration establishes the location of the perspective centre of each camera with respect to both the image formed by the camera, and the perspective centre of the

other camera. This step allows the intersection of pixel correspondences between the stereo intensity images to be projected into 3D space.

2. **Stereo Matching**, establishes correspondences of pixels in both stereo images. In this thesis pixel correspondences are found using the C3D matcher. This is a coarse to fine correlation approach. Stereo correspondence is an active area of research with advances continually being made. The Middlebury Stereo Evaluation benchmark (Scharstein and Szeliski, 2002) is the standard benchmark for contemporary approaches. The current state of the art in stereo matching relies on Markov Random Field, or loopy belief approaches (Szeliski et al., 2008; Mei et al., 2011). These can achieve an accuracy of around 96% of correct pixel correspondences.

3. **Intersection**; the location of each pixel correspondence between the cameras is then projected through the perspective centre to find the ray intersection, or closest point between two rays, in $R^3$ space (Abdel-Aziz and Karara, 1971; Hartley and Zisserman, 2005). The length of the vector between a camera perspective centre and the intersection gives the range value for the given camera.

4. **Post processing**; smoothing of the resulting range image or bundle adjustment to correct for erroneous ray intersections.

As each pixel represents a distance measure from the perspective centre of the camera to the object, the pixel divisions in the x, y image plane are proportional to angular divisions in azimuth and elevation. In this thesis the weak perspective camera model is used, where the angular divisions of the x, y image axes are assumed to be proportional to a linear change in distance measured in metres, Equation 2.1. This assumption holds provided the maximum angular deviation across the image is small, Equation 2.2. To achieve this, the mean distance to the object must be much greater than both the distance across the field of view at the object distance, and the total deviation of the object distances across the viewing space. As the range and intensity images use the same perspective model, the space in which the weak perspective model assumption holds is the same for both imaging modalities.

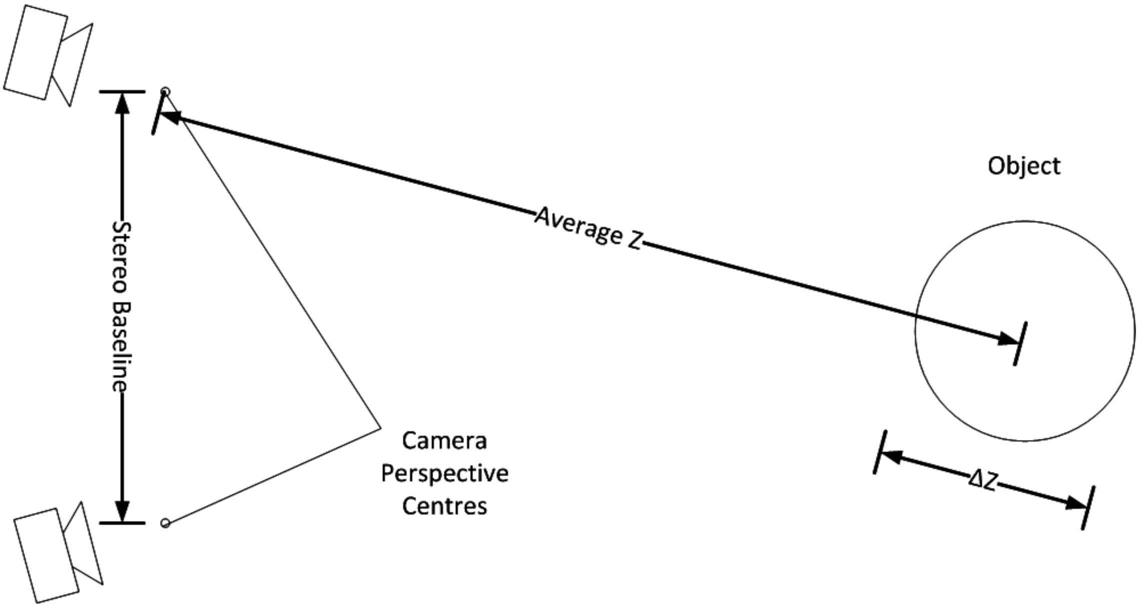


Figure 2.5: Stereo Capture Setup

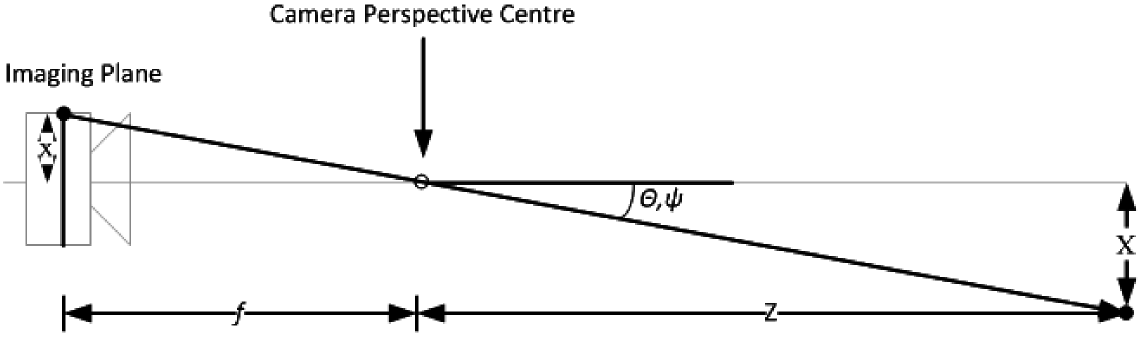


Figure 2.6: Imaging Plane

$$\begin{aligned} x &= f\frac{X}{\bar{Z}} \\ y &= f\frac{Y}{\bar{Z}} \end{aligned} \quad , \bar{Z} \gg \Delta Z,\ X, Y \tag{2.1}$$

$$\tan(\theta, \psi) \approx \theta, \psi \tag{2.2}$$

### 2.2.3 Range imaging setup

Figure 2.5, shows the layout of the stereo capture used in this thesis. This figure shows the average distance with respect to the perspective centre of the left camera. The cameras used are two, 5M pixel Prosilica GC2450C cameras; the left camera captures intensity and colour while the right camera captures intensity only. The monochrome camera in the stereo capture configuration has

| Parameter | Symbol | Value |
|---|---|---|
| Focal length | $f$ | 50mm |
| Imaging plane x, y values | *x, y* | 1:2040,1:2448 pixels |
| Real world X, Y | *X, Y* | ±19 centimetres |
| Range values | Z | 1.66 $\rightarrow$ 2.06 metres |
| Range variation | $\Delta Z$ | 20 centimetres |
| Average range | $\bar{Z}$ | 1.86 metres |
| Field of view in x, y direction | $\theta, \psi$ | 0.94 degrees |
| Stereo baseline | N/A | 25 centimetres |

Table 2.1: Stereo capture setup

lower noise characteristics compared to the colour camera. Using both colour and monochrome cameras in the stereo capture arrangement allows for co-algined range and colour images; where the range image demonstrates higher pixel correspondence confidence than could be achieved using two colour cameras. The parameters of the camera capture setup are shown in Table 2.1, here the view angle of the cameras can be seen to be less than $1^\circ$ giving the capture configuration a weak perspective model.

### 2.2.4 Summary of range image characteristics

Range images are a compromise between a 3D object representation and an imaging modality which can be observed from a single viewing perspective. As a 3D object representation, range images include occlusions of greater than 50% of the object surface, due to the occluded back face of the object (Besl and Jain, 1986; Frome et al., 2004). Additional characteristics which range images inherit from using a projective camera model are that the resultant images include perspective distortions (Geiger et al., 1995). Using stereo capturing approaches there exist circumstances where matches between left and right images cannot be established, in these cases a range value will appear erroneous. These cases are the result of two processes: firstly where the appearance change of a pixel location is greater than a matching process will allow, this can occur when imaging protrusions such as hair; secondly when a view of an object is available in the image from one camera, but not the other. To mitigate this the baseline between the cameras can be kept small.

## 2.3 Feature Based Image Analysis

Feature based image analysis is an approach with which it is possible to reduce the information present in a scene representation, so that the remaining information characterises only the desired image property, or label referring to the image content. In addition to removing superfluous information, the resulting description should be robust to a range of appearance variations which an object can exhibit in the chosen imaging modality. In this section an introduction to object recognition approaches based on extracting and matching features in range and intensity images is presented.

This section is organised as follows: Section 2.3.1 discusses the main paradigms in object recognition, and Section 2.3.2 elaborates on the local feature structure showing research relating to each stage in the processing pipeline.

### 2.3.1 Feature Extraction Paradigms

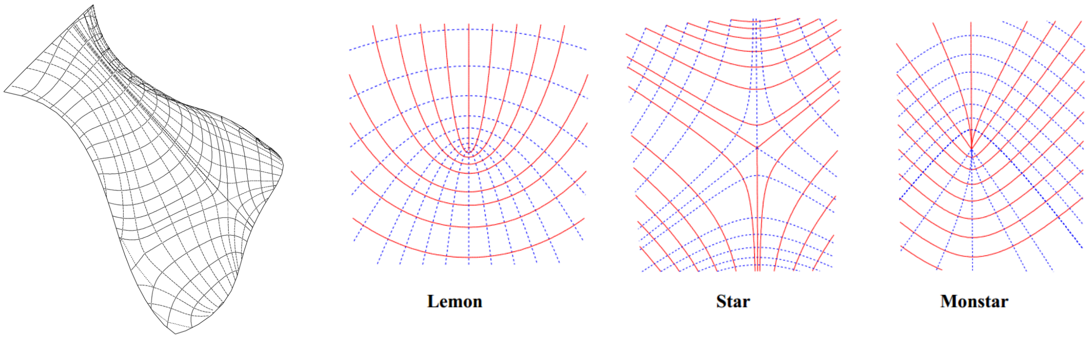
Since the 1970's the computer vision literature has tackled the problem of recognition from a number of different angles and for a number of applications. Initially attempts were made to recognise rigid unarticulated objects through analysing the structure present in the image representation as a whole. Line and edge structures from detectors such as the Canny edge detector (Canny, 1986) and Hough transform (Duda and Hart, 1972; Illingworth and Kittler, 1987) were used as a global geometric representation for the object which could then be used for matching directly to an object database. Matches in an object database could be established through finding a possible viewing angle of the object, which accounts for the edge structure observed in the image (Lowe, 1987; Roberts, 1963). As range data became available in the 1980s, the output of these scanners could be used as geometric models where the structure present could be matched directly (Kanade, 1987; Faugeras, 1993; Reid and Brady, 1995). However, these approaches were found to be susceptible to noise and clutter in the image scene even when identifying rigid unarticulated models (Faugeras and Hebert, 1986). To address these issues the problem was decomposed into segmentation (Pal and Pal, 1993) and recognition (Murase and Nayar, 1993; Pontil and Verri, 1998); to first find the object to be identified, then to characterise the object. Segmentation approaches are able to partition an image into regions with similar surface statistics which can be derived from texture (Belongie and Malik, 1998; Shi and Malik, 2000), colour (Klinker et al., 1990; Liu and Yang, 1994), surface curvature (Hoover et al., 1996; Powell et al., 1998), silhouettes in intensity or range (Pal and Pal, 1993).

In the late 1990's and early 2000's there was a revolution in object recognition allowing objects to be recognised in scenes under challenging conditions, such as occlusion, clutter and a range of poses. By dividing larger scenes up into smaller yet distinctive subimages, or sample patches, localised around interest points (Schmid and Mohr, 1996; Lowe, 1999), it was possible to apply standard recognition approaches which were previously used to characterise the scene as a whole (Murase and Nayar, 1993; Wallraven et al., 2003). This approach is thought to have much in common with biological vision present in mammals (Koenderink and Van Doorn, 1992; Burton et al., 1986; Jones and Palmer, 1987; Biederman, 1987). In applications where enough local features of sufficient quality could be captured under viewing conditions similar to those stored in a database the object recognition problem becomes solvable and robust as the composition of these features can be used to create a hypothesis regarding the object location and pose (Lowe, 2004).

More recently, much of the current research in the field has been directed to analysing large quantities of image data using standard local features such as SIFT or SURF (Blaschko and Lampert, 2008; Bay et al., 2008) in conjunction with text based retrieval approaches, to create a bag-of-visual-words (Freeman et al., 2008; Ponce et al., 2006; Sivic and Zisserman, 2003; Weinberger and Saul, 2004). However, these processes all introduce a bottleneck of information at the feature description stage. This problem has been noted for the conditions of view point change, where an investigation into improving the structure of the sampled intensity image data has shown improvements in overall performance (Winder and Brown, 2007). The creation of local features which retain their similarity under these conditions while still remaining distinctive is a challenging problem, with many approaches for increasing the invariance of sample patches to image transformations (Mikolajczyk and Schmid, 2005; Lo and Siebert, 2009; Brown et al., 2011; Winder and Brown, 2007; Lowe, 1999). The recent availability of co-aligned multi-modal range and intensity image data for robotics applications has created the possibility of integrating this cross modal data as a means to increase the robustness and distinctiveness of the extracted local features.

This section presents a selection of literature relating to extracting surface measures for representation. These surface measures are derived from:

1. **Global** approaches where the surface measures have been used to characterise areas with similar structure;



(a) NURBS global description

(b) NURBS features

Figure 2.7: NURBS Surface Parametrisation, (Ko et al., 2003)

2. **Local features**, which typically investigate local feature structure.

3. **Adaptive** approaches where the object class is known, and the variation or presence of the object is required.

#### 2.3.1.1 Global Feature Extraction Approaches

This section reports surface characteristics which may be used for matching between object instances. Where the full 3D information for an object is present in an uncluttered scene representation, the direct representation of the object can be used for recognition and matching using the Iterative Closest Point (ICP) algorithm (Besl and McKay, 1992; Zhang, 1994). This approach has been successfully applied to clutter free range images where there is a high overlap in the data available between examples, and has shown high recognition rates in facial recognition (Pears et al., 2010; Islam et al., 2009; Faltemier et al., 2008). However, the approach requires a comparison of two objects over many iterations and the representations of the object must contain sufficient overlap of the object surface. Additionally, poor initialisation of ICP can cause the representation search to find local minima as opposed to a true alignment. These issues make ICP prohibitively expensive as a surface representation for large databases. Non Uniform B-Splines (NURBS) have been used to recognise CAD models through a parametrisation of their object surfaces to create shape intrinsic fingerprints (Krishnamurthy and Levoy, 1996; Ko et al., 2003; Ling et al., 2010), see Figure 2.7. However, NURBS require large feature spaces, and full 3D information. These are typically used for matching point clouds, and suffer similar problems as ICP with data which include occlusions.

Frequency approaches such as Zernike moments (Novotni and Klein, 2003; Canterakis, 1999) or Spherical Harmonics (Frome et al., 2004; Zhang and Hebert, 1999) have been applied to 3D free form structures to create more generalisable rotationally invariant descriptions. These approaches can create a unique spectrum for an object which can be formed from a range of different local surface measures, such as colour, texture, surface normals (Zhang and Hebert, 1999; Dorai and Jain, 1997). Surface normals however are view dependant, therefore a more intrinsic measure of local surface structure has been required. This limitation can be accounted for using Splash Images (Stein and Medioni, 1992) or with approaches similar to NURBS where the surface is parametrised based on its local structure to allow a view independent representation to be created. Surface parametrisation can be achieved by fitting low order polynomials to local surface patches, an example of this is Principal Curvatures where the surface gradient directions and magnitudes can be used as a local representation (Besl and Jain, 1986; Gordon, 1992). Konderink and van Doorn advance this surface measure to create Shape Index and Curvedness (Koenderink and van Doorn, 1992). Shape Index extends Principal Curvatures by creating a single parameter to define the magnitudes of both principal curvatures, and second, Curvedness, which defines the ratio between the two principal curvatures. The Shape Index surface measure allows all surfaces to be categorised as a smooth transition between differing surface types using a single measure, Figure 2.8, with a second measure defining the scale of the surface type. Shape Index, in addition to creating an intuitive taxonomy of surfaces types creates a resulting representation which is invariant to changes in in-plane orientation and changes in object scale. Shape Index has proved a popular choice of surface measure for object recognition in the range imaging domain (Hetzel et al., 2001; Lo and Siebert, 2009; Dorai and Jain, 1997). In comparison to Principal Curvatures which combine measurements of surface type and surface scale in a single measure for each principal direction. The use of principal directions allow deformations to the surface to exist which preserve the Principal Curvature. These deformations change the principal directions, and therefore the *shape* of a feature. Shape Index however characterises any deformation of the surface as a change in either the Shape Index or Curvature (Koenderink and van Doorn, 1992).

Cross modal and multi-modal surface representation have been investigated. The facial recognition community have a large body of research and approaches for global recognition using a

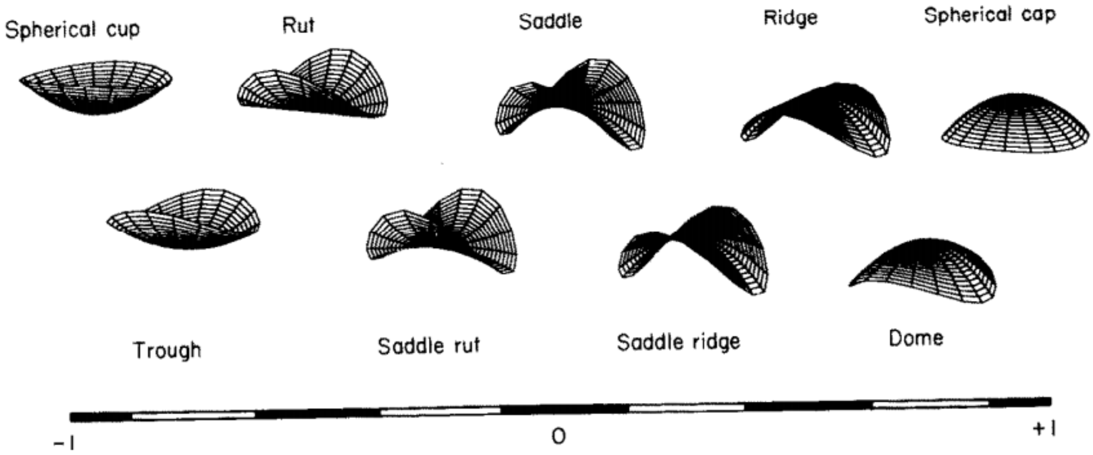


Figure 2.8: Shape Index (Koenderink and van Doorn, 1992)

combination of range and intensity images (Mian et al., 2007). Bowyer (Bowyer et al., 2006) and Abate (Abate et al., 2007) independently present two surveys of multimodal information applied to facial recognition. Both surveys report upwards of 95% recognition rates on the specific object example of faces, and both report a marked improvement when using the combination of information from range and intensity domains. Campbell and Flynn give a good summary of a number of recognition approaches for the recognition of 3D free form objects in range images (Campbell and Flynn, 2001). In their survey the best recognition rates are achieved using the local feature approaches of Spin Images (Johnson, 1997) and Point Signatures (Chua et al., 2000), see Section 2.3.1.2.

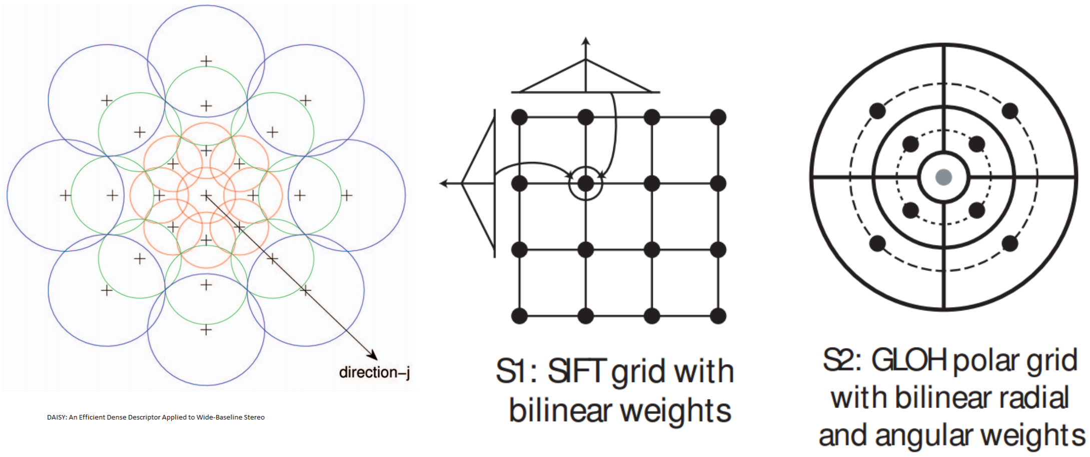
#### 2.3.1.2 Local Feature Extraction Approaches

Interest in local feature matching as an approach for object recognition in challenging environments began with the work of Schmid and Mohr in the late 90's (Schmid and Mohr, 1996). These approaches find keypoints in a scene representation which act as repeatable key locations or keypoints, which may then be identified in subsequent examples of the object in differing scenes. Since the introduction of local feature matching, Tuytelaars and Mikolajczyk (Tuytelaars and Mikolajczyk, 2008) have identified the key desirable qualities of any local feature extraction approach as follows:

- **Repeatability,** features when detected in one example of an object scene, must be detected in subsequent examples of the object. This measure is composed of two factors: invariance to deformations and robustness to changes in the keypoint expression; these factors are covered in greater detail in Section 2.3.2.2.

- **Distinctiveness,** the areas on the object selected as keypoints must demonstrate informative variation.
- **Locality,** keypoints are required to use a sufficiently small patch of the image such that the effects of occlusion and distortions due to out-of-plane viewer location changes are reduced.
- **Quantity,** for a scene representation, a large number of keypoints are required to increase the supporting evidence for a hypothesis regarding the scene content.
- **Accuracy,** the attributes of the localised feature should be accurately determined, these include the scale and orientation.
- **Efficiency**, for practical applications the keypoint extraction process should have the potential to be optimised.

In order to satisfy these conditions a range of local feature approaches have been proposed in the literature. SIFT features are the most common local feature extraction approach. Local features are detected in position and scale, and characterised by a histogram of image intensity gradients (Lowe, 2004, 1999). SURF features are an efficient alternative to SIFT (Bay et al., 2008), using integral images for scale localisation (Viola and Jones, 2001), and Haar wavelets for descriptor characterisation (Gonzalez and Woods, 1992). GLOH, features use a modified patch sampling approach and dimensionality reduction to compress the descriptor vector for increased efficiency in the matching stage (Mikolajczyk and Schmid, 2005), Figure 2.9b. PCA-SIFT applies dimensionality reduction directly to the SIFT sample patch. Ke and Sukthankar presents an evaluation of this approach (Ke and Sukthankar, 2004a). Tola et al introduce the DAISY local feature descriptor, which uses a SIFT-like structure with a densely represented image sampling patch, Figure 2.9a. This creates highly distinctive features required for stereo matching (Tola et al., 2010). SPIN Images are pose invariant local features which were originally proposed for matching between range images, or full 3D CAD models (Johnson, 1997; Johnson and Hebert, 1999). Spin images are formed from histogramming vertices in the feature local, based on the distance from the surface normal plane, and the radial distance from the feature surface normal. Lazebnik et al have extended these into the intensity


direction–j
DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo
S1: SIFT grid with bilinear weights
S2: GLOH polar grid with bilinear radial and angular weights

(a) DAISY, (Tola et al., 2010)

(b) SIFT and GLOH, (Winder and Brown, 2007)

Figure 2.9: Sample patch arrangement

domain and investigated the combination of intensity domain SPIN images with other descriptors (Lazebnik et al., 2005). (Chua et al., 2000) formulate Point Signatures, a descriptor similar to SPIN images, which demonstrate ~90% recognition rates for facial recognition. Point Signatures encode a range surface as a histogram of the vertices in the feature local, based on the angular distance from the surface normal the radial distance from the keypoint location.

The majority of the local feature matching approaches presented in the literature follow the processing pipeline outlined by SIFT. In order to give a fuller discussion of advances in local feature extraction methodology Section 2.3.2 structures various advances in the context of the SIFT processing pipeline. Figure 2.10, shows an example of local feature image sample patches used to form feature descriptors. Each subfigure in Figure 2.10 shows local sample patches extracted at a single scale, each sample patch is aligned to the dominant orientation of the patch.

#### 2.3.1.3 Adaptive Feature Characterisation

The previous sections on global and local features have discussed recognition when the given object is rigid, or does not exhibit significant changes at the scale of the local features. However, changes in 3D observation positions or deformations of the object, do result in a transformation being applied to the object surface affecting both global and local feature interpretations (Tuytelaars and

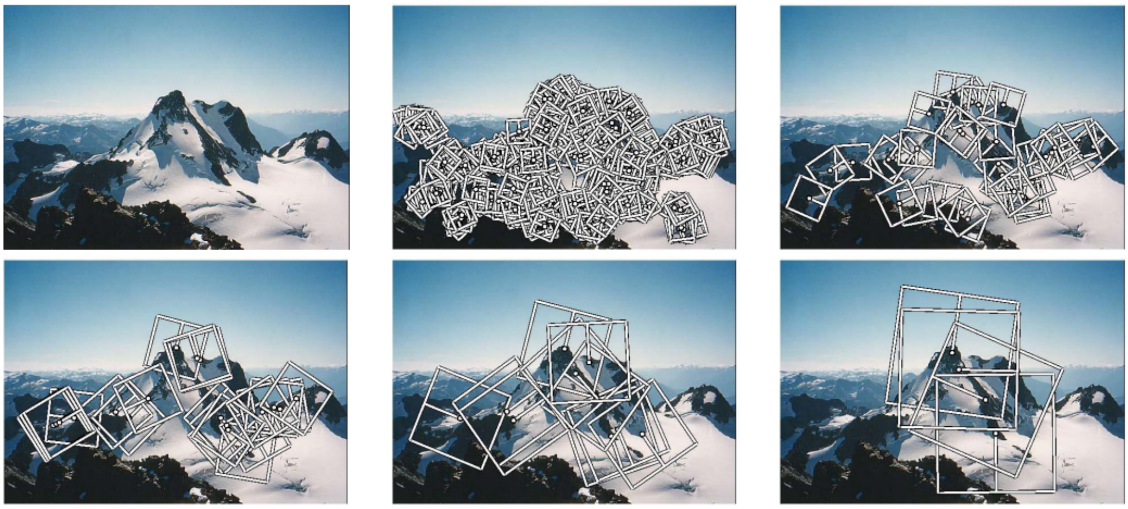


Figure 2.10: Local feature sample patches at differing scales (Brown et al., 2005)

Mikolajczyk, 2008; Gopalan et al., 2010). The problem has typically been constrained by considering the effects of deformation when the given object class is known. Pentland and Turk have presented *Eigenface,* where an image patch covering a face is modelled, with Principal Component Analysis, as a descriptor describing the common mode patch structure and an additional descriptor to define the informative variations characterising the specific example (Turk and Pentland, 1991). This is advanced using Fisher's Linear Discriminant to form Fisherfaces, which have shown improved performance (Belhumeur et al., 1997). Lee and Seung learn the characteristics of an image patch using non-negative matrix factorisation. The response of an image patch to a series of filters is learned to find a description for an object example (Lee et al., 1999). Viola and Jones present an instance locator, which is capable of learning object classes and finding image patches containing an instance of an object in larger scenes (Viola and Jones, 2001, 2004).

In addition to patch based approaches there are others which consider the underlying structure of the object. These include Elastic Bunch Graph Matching, EBGM (Wiskott et al., 1997; Albiol et al., 2008). EBGM marks faces with landmarks based on Gabor jet local features and a graph structure joining the features; the landmarks are iteratively moved to satisfy an energy minimisation function of the graph structure and local appearance. Active Shape Models generalise this approach to allow arbitrary objects to be recognised with the requirement of prior landmarking of training data (Cootes et al., 1995). Active Appearance Models extend this work further to fit a standard fully textured model to the variations shown in the example (Cootes et al., 2001). The description of the object can then be composed of the standard model and the deformations required to be applied to fit the model data to the object, Figure 2.11.

Felzenszwalb et al create a more detailed model of the possible variations which allows more complex free form objects, such as human bodies, to be represented and recognised in varying poses (Felzenszwalb and Huttenlocher, 2004, 2005). Many of these graph approaches have limited connections between collections of local features to optimise the performance for a single object class. Constellation models create a fully connected graph structure for more generalised application (Chung et al., 2009; Fergus et al., 2003). Recently partial 3D information has allowed further constraints for the motion of points which has aided simplifying the model (Matuszewski et al., 2011).

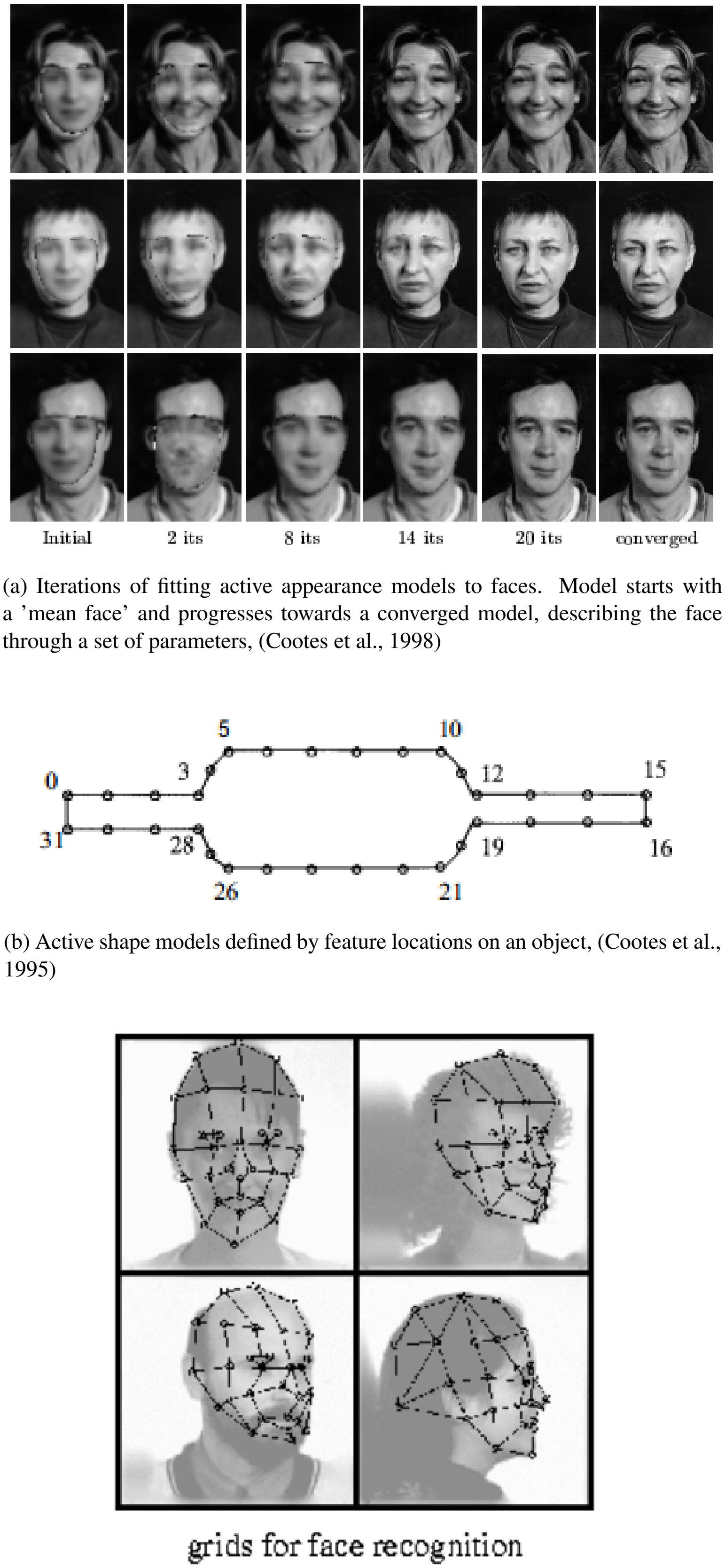
Stark et al learn the 3D shape of features from 3D CAD models and apply the models learned to recognition under out of plane orientation changes and deformations (Stark et al., 2010). Assumptions regarding 3D shape have been used to normalise intensity data for recognition of instances of faces for head pose tracking (Zabulis et al., 2009). Rough parametrisation of objects can be used to define contours and create Shape Context descriptors (Belongie et al., 2002). These have successfully used to recognise deformable silhouettes (Ling et al., 2010) and have been recently extended for describing full 3D objects (Kokkinos et al., 2012).

### 2.3.2 Local Feature Structure

SIFT local features outline a processing pipeline which has been followed by other approaches which adopt the local feature paradigm (Lowe, 2004, 1999). This section seeks to demonstrate the advances made in local feature matching in this context. Local feature approaches which do not follow the SIFT structure have their advances discussed under the subsection which most relates to the novelty of the approach. Figure 2.12 shows the SIFT feature extraction pipeline.

#### 2.3.2.1 Scale Space

Scene representations contain a range of spatial frequencies where informative features can appear with differing size. Furthermore perspective changes and image resizing require that image features may be created at any arbitrary scale, and matching must be done irrespective of the scale at which the feature appears in an example (Lindeberg, 1993). In order to characterise features of differing



(a) Iterations of fitting active appearance models to faces. Model starts with a 'mean face' and progresses towards a converged model, describing the face through a set of parameters, (Cootes et al., 1998)

(b) Active shape models defined by feature locations on an object, (Cootes et al., 1995)

(c) Elastic Bunch Graph Matching, fitting bunch graphs to faces based on Gabor jets, (Wiskott et al., 1997)

Figure 2.11: Recognition based on adaptive features

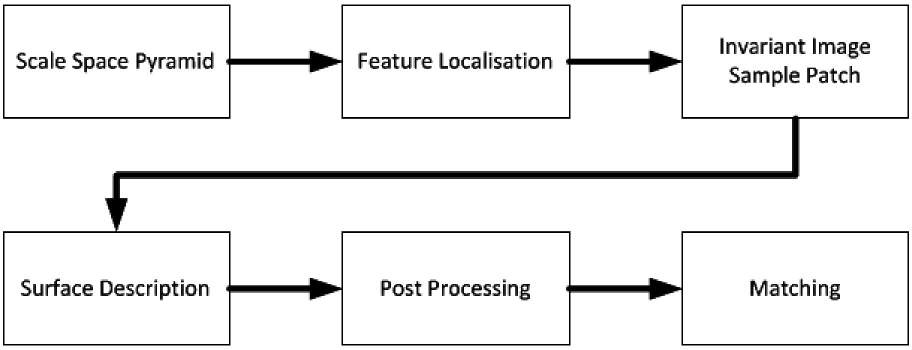


Figure 2.12: SIFT Feature Extraction Pipeline

scale SIFT divides the source image up into a Gaussian Image Pyramid where each image in the pyramid contains a limited range of spatial frequencies (Lowe, 2004). Alternatively, SURF uses integral images where gradient responses can be efficiently found at any scale without the requirement of building and storing an image pyramid (Bay et al., 2008; Viola and Jones, 2001). Wavelets are another approach for calculating gradient responses at differing scale; wavelets calculate the magnitude of the gradient response to wavelet filters characterising differing spatial frequencies (Mallat, 1989).

#### 2.3.2.2 Feature Localisation

Many types of features exist in images, such as lines, contours, corners, gradients etc (Szeliski, 2010). Localising these features within a repeatable measurement window is referred to as the aperture problem, Figure 2.13. Corner detectors are the most localisable out of the taxonomy of image features and therefore the best suited for localising image sample patches within larger images. The detection of corners can be established by a range of approaches designed to minimise the effects of viewer location, lighting conditions. Moreels and Perona (Moreels and Perona, 2007) investigate differing corner detectors approaches and report similar results to Mikolajczyk et al (Mikolajczyk and Schmid, 2005), with Hessian Affine corners located in a Difference of Gaussians Image Pyramid features remaining the most stable across a range of transformations.

Maximally Stable Extrema Regions, MSER, were proposed by Matas et al (Matas et al., 2004). These features were originally proposed as repeatable features for 3D reconstruction of scenes from multiview geometry, Figure 2.14. These features have been shown to demonstrate high performance on planar scenes, however reduced performance on objects with complex 3D geometry (Moreels and Perona, 2007). Gupta and Mittal show improved repeatability in feature localisation by find-

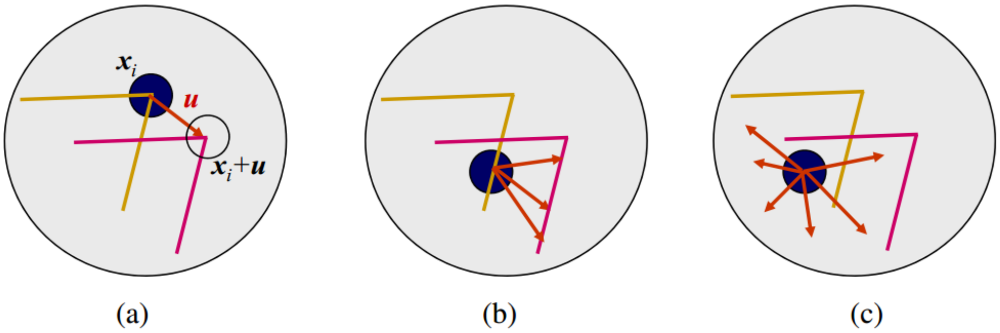


Figure 2.13: Aperture Problem, highlighting the difficulty in finding a localisable correspondences between image locations. Here the effects of attempt to match (a) corners, (b) lines, and (c) areas, are shown (Szeliski, 2010)

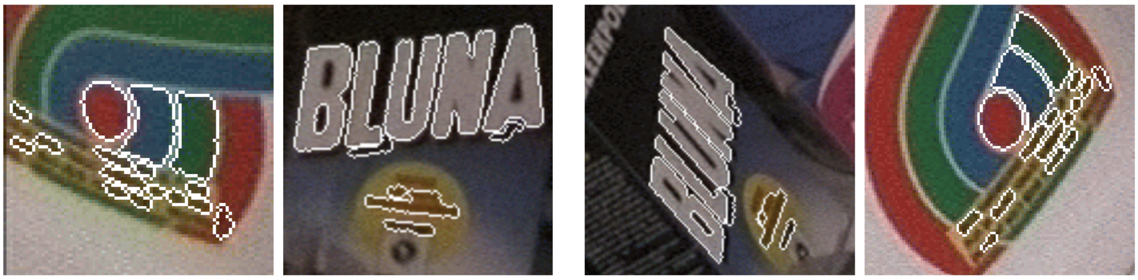


Figure 2.14: Outline of MSER features repeatably localised between pairs of images with large changes in viewpoint , (Matas et al., 2004)

ing line intersections of distinct lines in intensity images; however their approach produces fewer keypoints than corner based approaches (Gupta et al., 2007).

Histograms of Orientated Gradients, HOG, descriptors identify all locations in the image as potential interest points. Through densely sampling the image the significant locations are learned in a post processing stage (Agarwal and Triggs, 2006; Dalal and Triggs, 2005; Nowak et al., 2006).

#### 2.3.2.3 Image Sample Patch

For localised feature instances to be matched between examples, a unique identifier tag, or keypoint descriptor is required. The keypoint *feature descripto*r is formed from a measure of an extracted image sample patch from the local area surrounding the interest point. To increase the distinctness of the resultant descriptor and allow a measure of local image composition to be formed, image

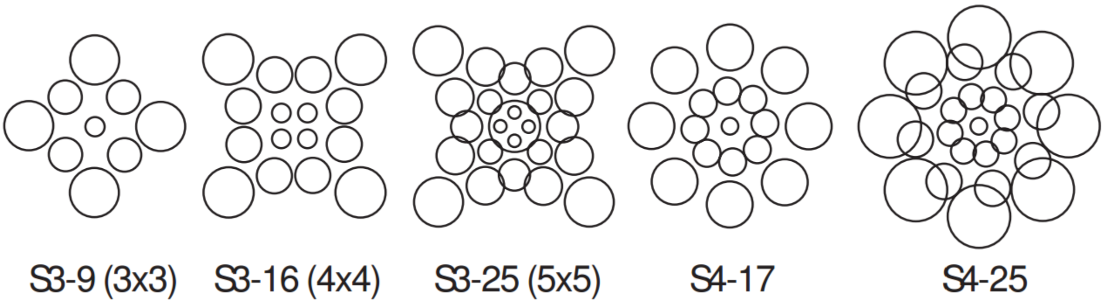


Figure 2.15: Collection of image sample patches, sampling patches are labeled Sx-n, where x refers to the spatial sampling scheme, and n defines the number of receptive fields (Winder and Brown, 2007)

sample patches are subdivided into a collection of receptive fields or subfield arrangements. Due to their use in SIFT, rectilinear arrangements of receptive fields are a popular image sample patch decimation (Lowe, 2004; Bay et al., 2008). GLOH uses a log polar image patch sample patch decimation more similar to a mammal retina and focuses the descriptor information content on the keypoint location (Mikolajczyk and Schmid, 2005). The GLOH arrangement is advanced by DAISY (Tola et al., 2010) which is composed of an overlapping and foveated arrangement. The composition of receptive fields using the DAISY approach is investigated by Winder et al (Winder and Brown, 2007; Winder et al., 2009), Figure 2.15. Fan et al propose LIOP, which forms a unique receptive field arrangement for every descriptor by segmenting the sample patch based on the local intensity (Wang et al., 2011), Figure 2.16.

In addition to choices of receptive field arrangements, image sample patches may have a variable shape based on a calculated measure of the local interest point pose. SIFT calculates the canonical orientation of the interest point, the image sample patch is then aligned with this orientation (Lowe, 2004). Mikolakczyk and Schmid apply an Affine warping function based on the local characteristics of the 2D intensity image to give partial invariance to changes in 3D observation location changes (Mikolajczyk and Schmid, 2004), Figure 2.17. Lo uses range image surface gradients to apply a similar Affine warping function to a range image sample patch (Lo and Siebert, 2009). SPIN images extracted from point clouds apply a full 3D out-of-plane rotation normalisation to sample patches, these however have no sensitivity to in-plane patch orientation (Johnson and Hebert, 1999).

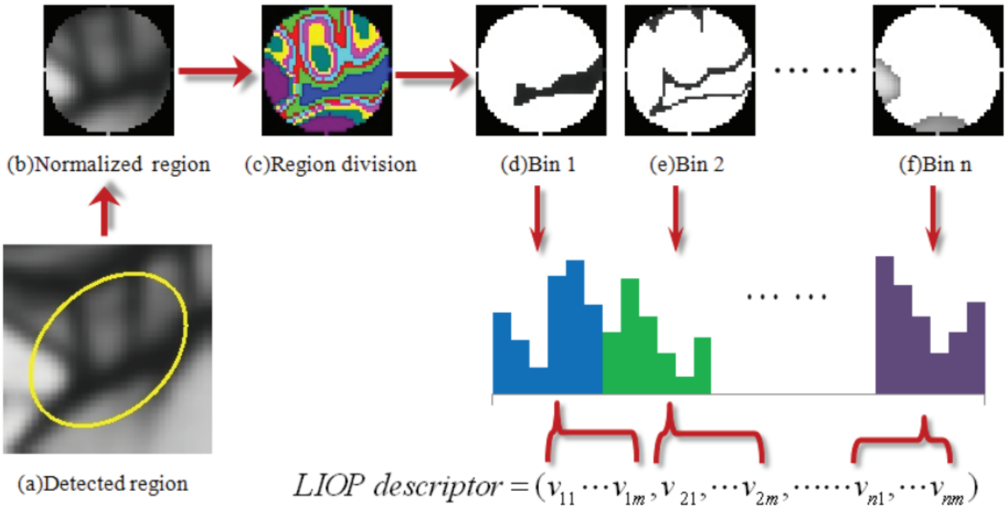

(b)Normalized region
(c)Region division
(d)Bin 1
(e)Bin 2
(f)Bin n
(a)Detected region
LIOP descriptor = ($v_{11} \cdots v_{1m}, v_{21}, \cdots v_{2m}, \cdots\cdots v_{n1}, \cdots v_{nm}$)

Figure 2.16: LIOP Regions segmenting a sample patch, (Wang et al., 2011)

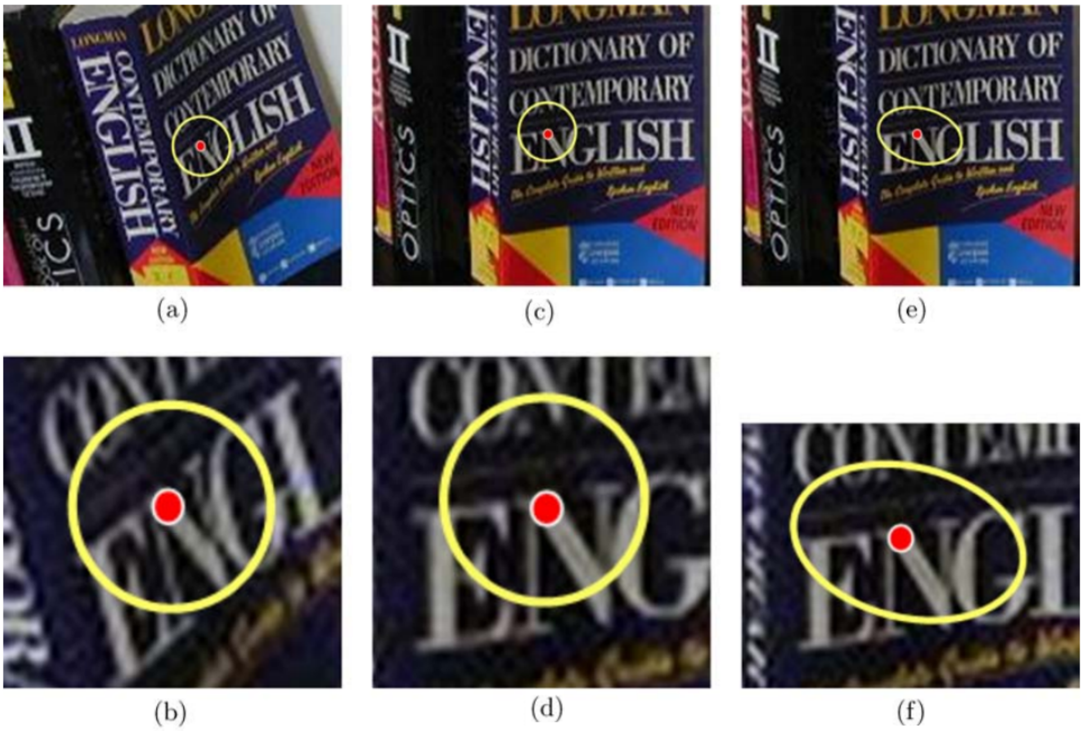


Figure 2.17: Affine corrected sample patches: (a, b) canonical view, figures (c-f) show the object viewed from a new view point, the lower row of figures show an enlargement of the sample patch. Subfigures (c, d) use a sample patch with in-plane orientation correction only, (e, f) use an Affine corrected sample patch, covering the same image region as in the original image (a, b), (Mikolajczyk and Schmid, 2004)

#### 2.3.2.4 Surface Description

To create a unique descriptor for a keypoint location, the extracted image sample patch and receptive field arrangements are used to sample the local image statistics. The local statistics are histogrammed to form a key mathematical representation of the keypoint local, termed a *feature descriptor*. A variety of surface measures have been proposed in the literature, these include intensity image gradient orientation and magnitude measures. SURF uses Haar wavelets, these are a fast approximation where the local response to a binary gradient operation is measured (Bay et al., 2008). To reduce quantisation noise Local Binary Patches, LBP, have been proposed. These are a local binary operation which encodes the comparison between a pixel intensity and that of its neighbours (Ojala et al., 1996; Heikkilä et al., 2009). Gupta et al investigate LBP approaches further using different pixel comparisons for forming descriptors (Gupta et al., 2007; Zabih and Woodfill, 1994). Additionally, descriptors may be formed from the relationship of shapes and lines (Gupta et al., 2010a). Belongie et al use shape context to form rotationally invariant descriptions of objects (Belongie et al., 2002). Zitnick uses a local parametrisation approach to form descriptions of lines in local image patches (Zitnick, 2010). Shape Index and Curvature have been shown to be more robust measure of local surface statistics in the range imaging modality (Lo and Siebert, 2008, 2009; Dorai and Jain, 1997; Lukins and Fisher, 2006; Hetzel et al., 2001; Atmosukarto et al., 2010). Campbell and Flynn provide an overview of description metrics for range images (Campbell and Flynn, 2001). Bowyer provides a survey on the choices of surface measures for facial recognition in the range and intensity domains (Bowyer et al., 2006).

#### 2.3.2.5 Postprocessing

Having established a collection of keypoints, with the desirable properties outlined in Section 2.3.1.2, it is possible to apply these keypoints to a range of applications. The information presented in the keypoint structure may be optimised for a given application through a series of post processing steps. This subsection details a limited selection of these applications and their associated post processing steps. The goal of this subsection is to place local features in a chain of processing from image capture to application, and identify potential steps which may be incorporated into the extraction process or additional information which extracted keypoints may present to allow for further post processing stages.

Local features have been used for object matching (Lowe, 2004) and image stitching (Brown and Lowe, 2007). In these application, extracted keypoints must form a consensus on the composition of keypoint locations between image examples. In order to identify keypoints which conform the generalised Hough transform has been used on local feature location and orientation (Pope and Lowe, 2000; Ballard, 1981). Object recognition from individual keypoints in a large database requires that the processing power required to match two keypoints is minimised. To reduce the processing required to match two points an efficient search through the descriptor space, such as a k-d tree, may be used (Muja and Lowe, 2009), or the descriptor length may be reduced by applying PCA to the keypoint feature descriptor (Ke and Sukthankar, 2004b). Efficient keypoint matching has applications in mobile phone, and optical flow (Liu et al., 2011). An additional level of abstraction may be introduced so that extracted keypoints themselves form data from which keypoints may be extracted. This bag-of-features approach has been used in applications such as pedestrian detection (Dalal and Triggs, 2005), instance recognition (Fei-Fei and Perona, 2005; Agarwal and Triggs, 2006) and scene interpretation (Sivic and Zisserman, 2003). The application of instance detection allows for use of positive and negative keypoint matches to enforce the resulting representation of an instance (Lee et al., 2011). Recently, further levels of abstraction to bag-of-features models have been added to create a hierarchy of recognition for a semantic representation allowing for a comprehensive vision system based on local features (Lai et al., 2011), Figure 2.18. However, the performance of all post processing stages outlined in this section rely on the successful representation of the initial image data in local feature descriptor form.

## 2.4 Summary and Conclusions

This chapter has introduced the characteristics of various imaging modalities and the restrictions these place on the available information regarding an object. Specifically the characteristics of range images have been discussed in detail. These are shown to offer a partial 3D representation of the underlying structure of an object, which is unavailable when using only intensity images. Range images are shown to be a challenging representation which share many characteristics with intensity images, such as the perspective projection and a similar level of surface occlusion. A number of range scanning devices have been introduced in this chapter to demonstrate the availability of

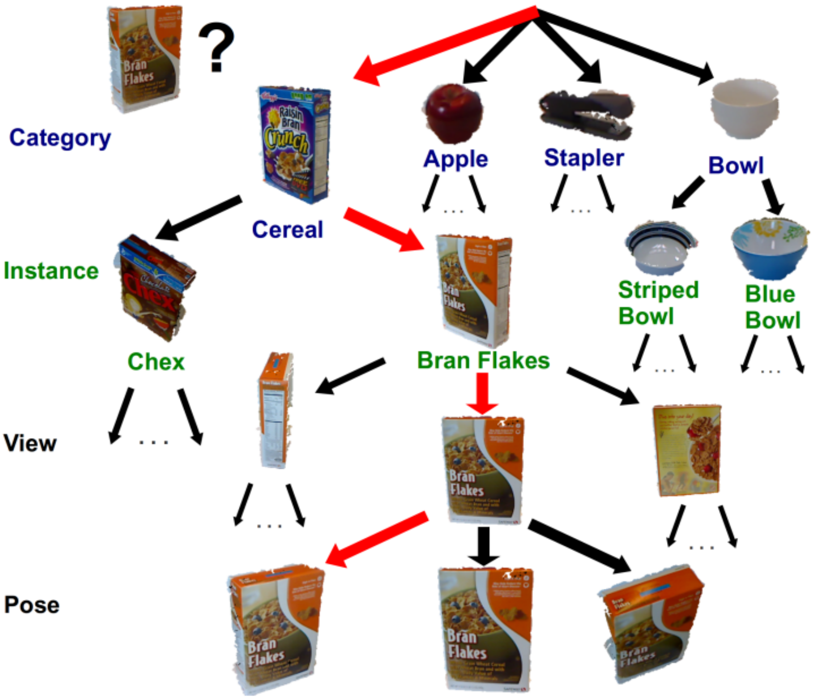


Figure 2.18: Bag-of-features hierarchy showing the use of local features in an advanced vision system preforming category level recognition with additional meta data (Lai et al., 2011)

capture devices supporting this imaging domain. Furthermore, stereo imaging has been identified as an area of specific interest, as scanners are passive, cheap, and form co-aligned range and intensity images. Additionally, given these qualities there is a recent body of research into reducing processing time and increasing the quality of the resultant range image (Scharstein and Szeliski, 2002).

The approaches to analysing the underlying data from imaged representations of objects were separated into 3 categories for discussion: global approaches, characterising the whole image scene through measures of surface types; local features where portions of an image are used to form point correspondences between images; and adaptive approaches which are capable of accounting statistically for variations between imaged objects. These 3 approaches are frequently used in combination where measures of local surface topology categorisations are used to find point correspondences in images to form a collection of feature descriptors. The statistical variation may then be accounted for over this set of descriptors. Recently, many of the approaches adopted by the research community for the recognition of free form objects have been based on adopting a local feature representation and then applying an analysis to the resultant descriptor (Fei-Fei and Perona, 2005; Pontil and Verri, 1998; Sivic and Zisserman, 2003; Sivic et al., 2005). However, it has been noted that the local feature representations themselves form an information bottle-neck, where sufficient information is required to characterised the local image region surrounding the interest point. In addition, the representations formed should not be sensitive to transient information which may change between observations (Tuytelaars and Mikolajczyk, 2008). Therefore, local feature extraction processes which are capable of characterising an invariant representation of a interest point local region will improve the results of any subsequent processing and analysis (Winder and Brown, 2007).

The adoption of 2.5D SIFT local features in range images has demonstrated the ability to form robust local range feature descriptors (Lo and Siebert, 2009). However, 2.5D SIFT does not incorporate additional information from the intensity domain, which has independently been used to form distinct local feature representations. There is evidence that features calculated using multimodal information from co-aligned range and intensity images can demonstrate an improvement in local feature distinctiveness, as multimodal information has been used to increase the performance of facial recognition systems on the FRGCv2 (Bowyer et al., 2006). The concept of using multimodal information from co-aligned range and intensity images has not been investigated in a local feature context.

In order to investigate the application of co-aligned range and intensity images in local feature extraction processes, the underlying processing pipeline architecture that implements the Scale Invariant Feature Transform, SIFT, is adopted (Lowe, 2004). SIFT is a frequently cited extraction approach which has been applied to a number of computer vision tasks from robotic vision, to image retrieval. The algorithm comprises a pipeline of low level image processing steps which divide a scene representation into a series of patches surrounding interest points and characterises each interest point with a unique descriptor. This chapter has decomposed this processing pipeline into a combination of choices of modality and method for each stage in interest point localisation, sample patch transformations, and surface measures. An advance on the state-of-the art for any recognition system based on local features can be made through careful design choices and through the integration of information from an appropriate modality at each stage of the processing pipeline. This thesis therefore seeks to investigate the integration of multimodal range and intensity domain images as an approach to improving the distinctiveness of robust local features.

The next chapter introduces an initial investigation into the choices of design parameters for each stage in the SIFT algorithm; a test configuration using 2D intensity images is developed to investigate the success of variations on SIFT; and a series of rules for design parameters is devised. Chapter 4 extends this experimental configuration allowing the combination of range and intensity domain to be investigated in Chapter 5. Chapters 6 and 7 seek to use adaptive recognition approaches for extending descriptor representations for handling surface patch deformations resulting from changes viewing position.

# Chapter 3

# Initial Investigation

This chapter follows from the background and literature review to conduct an initial investigation into the implementations of presented concepts and extends existing SIFT code towards an implementation of 2.5D SIFT. To evaluate changes made to the existing SIFT code an experimental methodology capable of tracking keypoint locations under a range of image transformations is outlined. The outlined experimental setup is used to test all changes made and results are presented as ROC curves.

## 3.1 Objectives

The main objective of this chapter is to investigate the stages in SIFT which influence feature matching performance in order to formulate underlying rules for algorithm configurations; from this set of rules an optimum configuration is found. The modified SIFT configurations examined in this section focus on the surface interpretation and extraction of local surface statistics for creating keypoint feature descriptor vectors. In order to conduct this investigation this chapter develops a readily reconfigurable and efficient experimental setup with the ability to directly compare any modified SIFT formulation against an implementation of Lowes original SIFT (Lowe, 2004). The experimental setup presented in this chapter focuses on point-to-point matching of individual keypoints; the composition of keypoints and higher level feature matching stages which analyse collections of keypoints are omitted.

Using the experimental setup presented in this chapter the performance of all modified SIFT algorithms can be compared directly against the performance of standard SIFT to isolate the effects of each modification. All implementations of SIFT variants presented are based on freely available SIFT MATLAB code (F. El-Maraghi, 2008). The initial stages of the SIFT algorithm involve the preprocessing steps of feature detection and scale space filtering. In these steps the Harris corner blob detection threshold level, number of octaves and intervals in scale space parameters are fixed focusing this investigation to the effects of surface representation in the feature descriptor. This results in the SIFT variations investigated in this chapter using the same finite support region window to sample the data, allowing the match quality of the resulting feature descriptor to attributed to the intended changes in the local receptive field arrangement, angular resolution and local surface statistic representations. It is anticipated that the influencing factors in the SIFT algorithm investigated in this chapter can be extended to sample range data similarly. The results from this chapter can then be used to create local features in an approach which produces an increase in the distinctiveness and robustness of the feature descriptor in both domains.

The modified SIFT algorithms investigated in this chapter are evaluated against each other by considering their ability to correctly identify a location instance in a transformed image based on local surface statistics while minimising the probability of incorrect labelling of the location instance. In order to examine this an experimental configuration was developed which is capable of producing a range of transformed query images from a collection of target images. In addition to applying the image transformation, each query image also stores ground truth for the transformation of each pixel location between the target and query images. The range of image transformations investigated as characteristic of the environment in which the SIFT algorithms are intended to be used. The variant SIFT algorithms under investigation are run on both the query and the target image sets to produce a dataset of feature descriptors with keypoint locations for each image. The feature descriptor matches are found using the Euclidean distance between features with a log-likelihood distance measure between the first and second nearest neighbour. The descriptor matches between the query and target datasets are then compared against the keypoint location correspondences from each image to find the correct and incorrect matches based on the feature descriptor matches. The ratios of correct-to-incorrect feature descriptor matches are presented as ROC curves, this gives an indication of the performance of each algorithm independent of the selection of log-likelihood threshold.

## 3.2 Validation experimental design

The various SIFT implementations presented in this chapter were evaluated on their ability to identify location matches between transformed images based on the local surface statistics of these locations. This section shows how the test data was created, the ground truth established and finally how the ROC curve for each SIFT variation was created. An overview of the principles behind the system is shown in Figure 3.1. This figure shows SIFT keypoints created for every target image. The target images have transformed copies made to create a test set of query images from which SIFT keypoints can be extracted, see Figure 3.3. The transformation from target to query image is used as ground truth to associate the positional matches to the feature descriptor matches. From the correspondences between the the positional and feature descriptor matches, a ROC curve can be created to show the performance of each of the SIFT variations. This section describes the process described above, and concludes with an overview of the implementation of the bench marking system.

### 3.2.1 SIFT Keypoint Matches

Descriptor matches are found through a comparison of query and model feature descriptors. Feature descriptors are compared using the Euclidean distance to determine for a query feature descriptor the first and second closest matches in the model feature descriptor database. The first nearest neighbour will be used as a candidate match, given that the distance between the model and query feature descriptors is below an adaptive thresholded level, termed the *log-likelihood* threshold. The log-likelihood threshold uses the ratio of the first nearest neighbour to the second nearest neighbour match to reject keypoints which are indistinctive and likely to lie within the portion of the density function describing false matches. Figure 3.2 shows the probability density function of correct and incorrect matches for a database of 40,000 keypoints using Lowe's SIFT implementation Lowe (2004). In this figure 0.8 is found to be an optimum log-likelihood threshold. However, depending on the desired application of the local feature extraction approach, more correct keypoints can be found by lowering this threshold, or fewer incorrect matches can be found by raising the log-likelihood threshold.

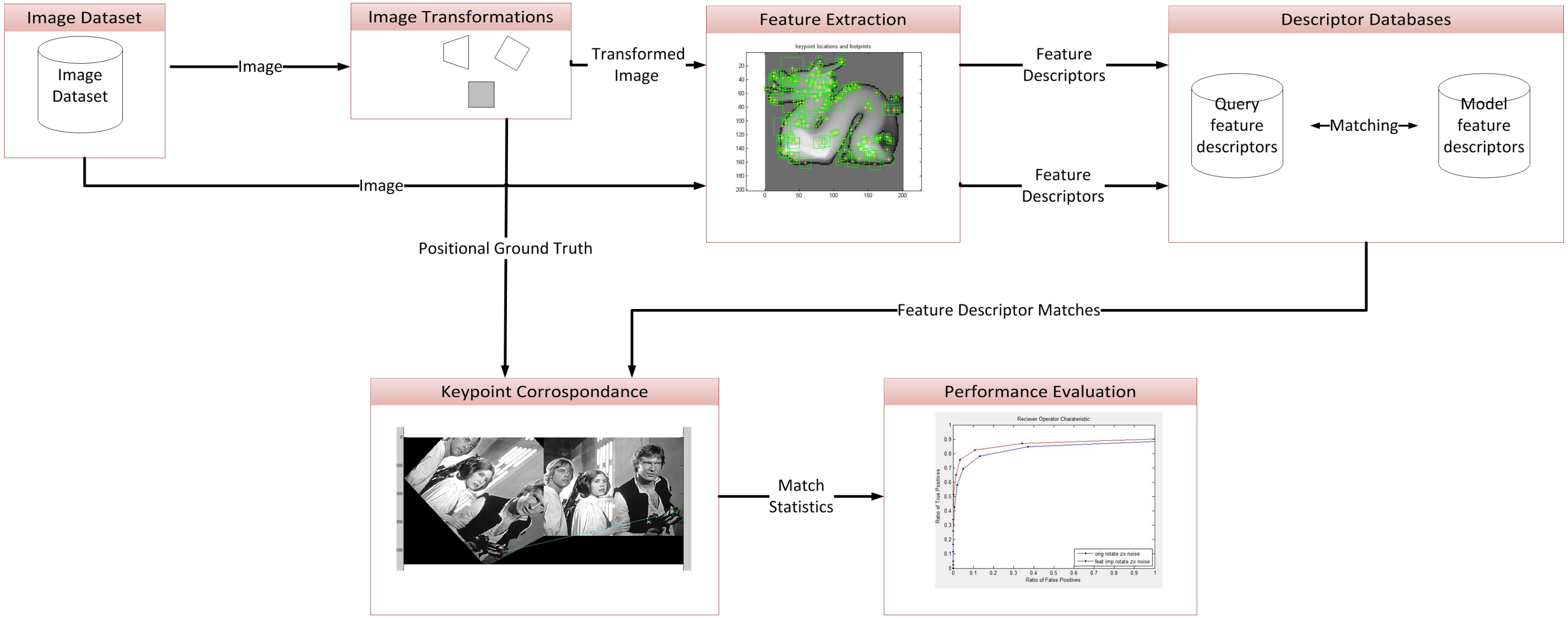


Figure 3.1: Flow chart showing an overview of the experimental design

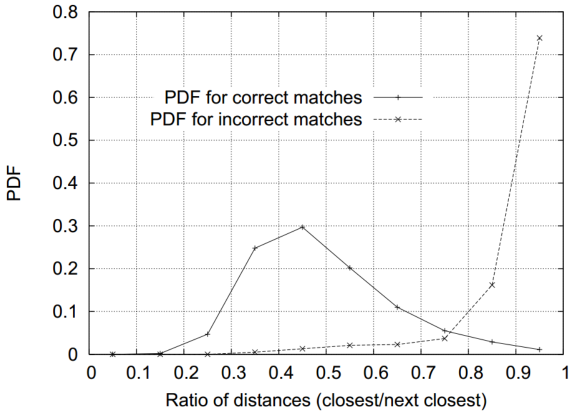


Figure 3.2: Probability density function for correct and incorrect matches based on a ratio between nearest neighbour to second nearest neighbour (Lowe, 2004)

Each keypoint is described by the keypoint attributes, shown in Equation 3.1. These attributes are:

1. keypoint x and y position in the image
2. keypoint scale, $\sigma$
3. keypoint dominant orientation, or canonical orientation, $\theta$
4. and a keypoint feature descriptor vector, in the case of SIFT a 128 element vector.

Distances between model and query keypoints are found using the Euclidean distance, shown in Equation 3.2. The distance between a query keypoint and every model keypoint in the database is found, $desc_1$ denotes a 128 element feature descriptor vector from the query image, and $desc_2$ denotes a 128 element feature descriptor vector from the model database.

$$keypoint = \{x, y, \sigma, \theta, desc\} \tag{3.1}$$

$$dist = \sqrt{desc_1^2 - desc_2^2} \tag{3.2}$$

The log-likelihood nearest neighbours matching criteria is shown in Equation 3.3. The value $distn_1$ denotes the distance for the nearest neighbour match, $distn_2$ denotes the distance of the second nearest neighbour match, and $keyn_1$ denotes the index of the first nearest neighbour. In Section 3.2.2.2, the sensitivity level for the ROC curve is varied by varying the value $Threshold$.

$$match = \begin{cases} keyn_1, & \frac{distn_1}{distn_2} \leq Threshold \\ 0, & \frac{distn_1}{distn_2} > Threshold \end{cases} \tag{3.3}$$

### 3.2.2 Benchmarking Tool

The performance of the local feature matching algorithms examined in this chapter was determined through a comparison of keypoints location correspondences against feature descriptor matches between keypoints extracted from a target database and a query image. In order to achieve this goal a benchmark tool was programmed in Matlab. The benchmark tool allows a query image with a known transformation to be synthesised from an image in the database and to have keypoints extracted from it to be compared against those extracted from its untransformed counterpart.

#### 3.2.2.1 Test Images

The test image database is a set of 20 2D images collected from a Google image search. These images are intended to cover a variety of scenes with a good range of image textures at a range of scales, see Figure 3.4. The local feature matching algorithm being tested is run on each of these images to create a database of keypoints specific to the algorithm being examined. Query images are then created by randomly selecting an image and applying a random transform defined in Section 3.2.3.

#### 3.2.2.2 Performance Evaluation

The performance of local feature matching algorithms is evaluated by comparing the keypoint position correspondences against keypoint descriptor matches found in the target and query images. The ROC curves are created by plotting the ratio of true positive of keypoints in the query image against the ratio of false positive keypoints in the query image, for a range of sensitivity levels. The true positive and false positives keypoints are determined using the confusion matrix defined in Figure

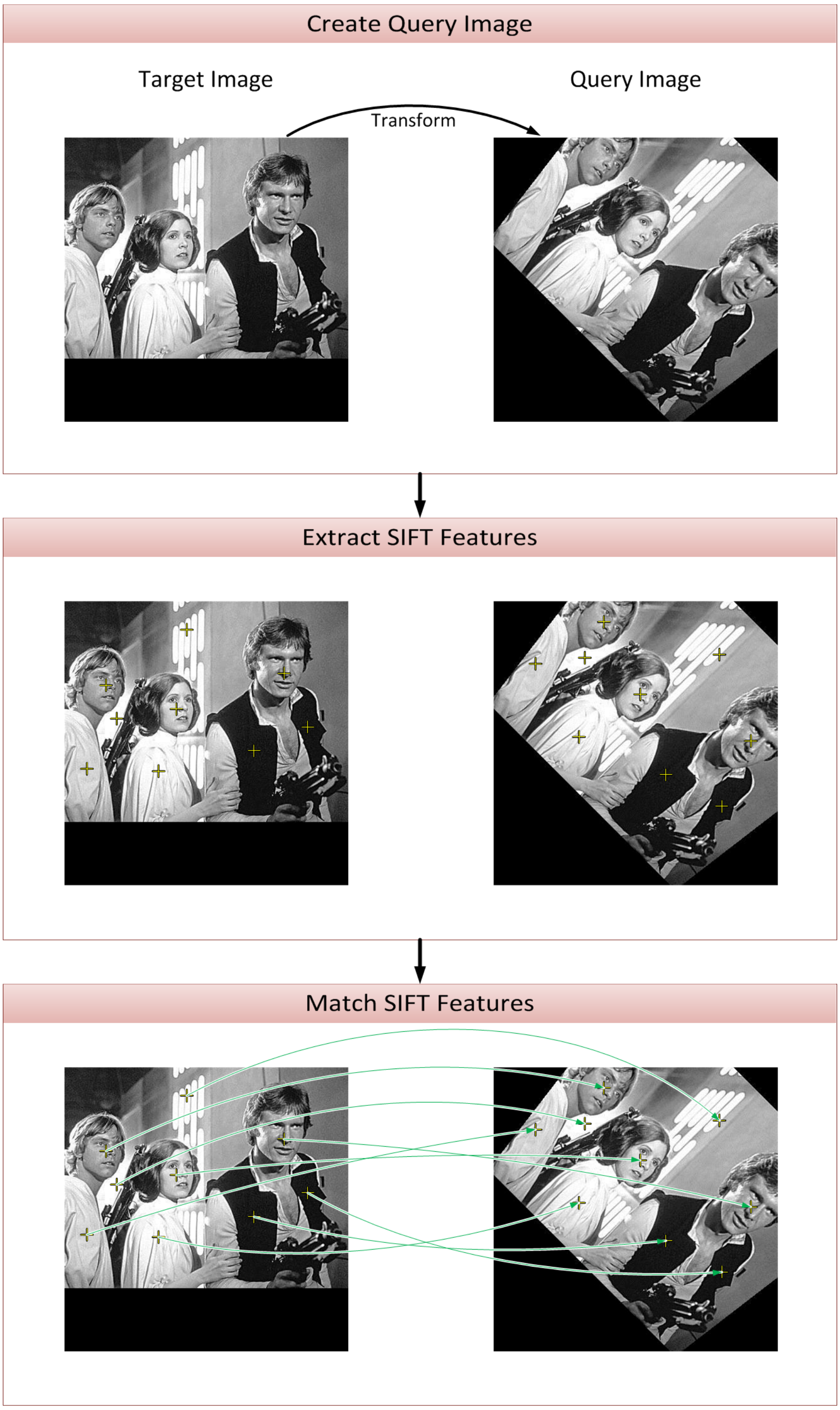


Figure 3.3: Keypoint Matches

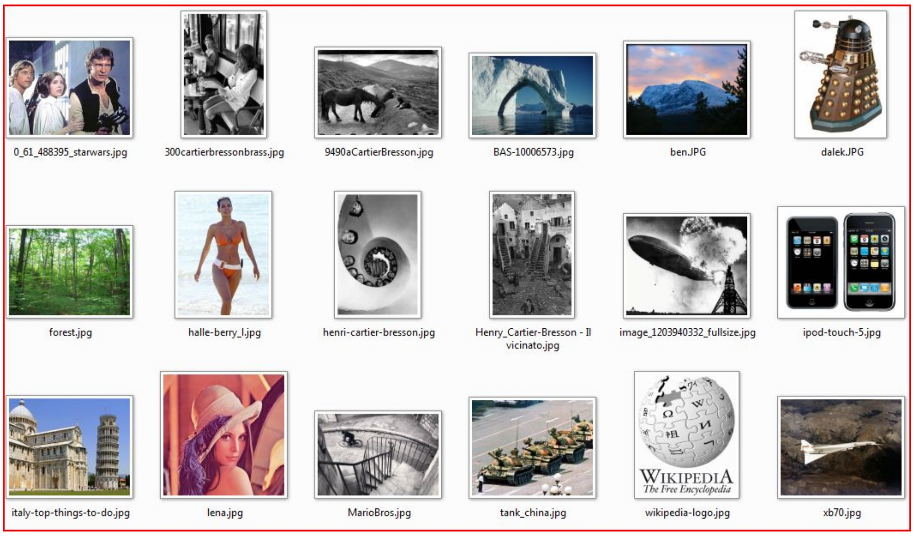
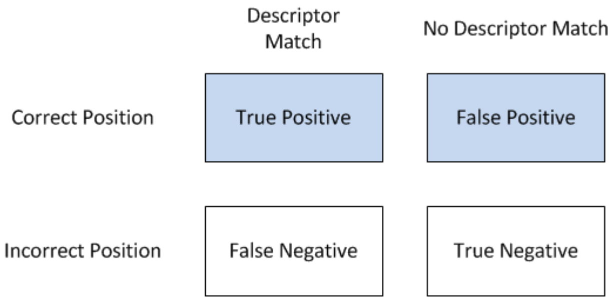


Figure 3.4: Image Database



| | Descriptor Match | No Descriptor Match |
|---|---|---|
| Correct Position | True Positive | False Positive |
| Incorrect Position | False Negative | True Negative |

Figure 3.5: Confusion Matrix for Evaluation ROC Curve

3.5. The ideal ROC curve performance would be demonstrated by a matching algorithm capable of finding all true positive matches and no false positive matches. This performance is shown as the point tp = 1 fp = 0 on the ROC curve, Figure 3.6.

Keypoint position correspondences are defined as keypoints from the query image which, when their positions are transformed using the ground truth, are within a catchment region surrounding a target keypoint position. The catchment region surrounding the keypoint is region of 1.5 sigma of the scale of the query keypoint and has a tolerance of a change of up to 0.5 divisions in scale space between the keypoint location in the target image and the transformed location of the query keypoint, Figure 3.7. The ground truth for the transformed location is outlined in Section 3.2.3. Descriptor matches are found using the approach outlined in Section 3.2.1.

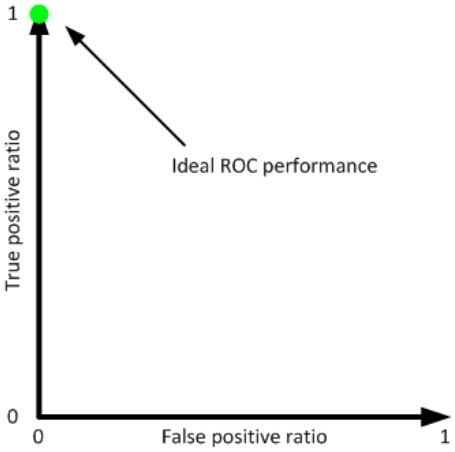


Figure 3.6: Ideal ROC performance

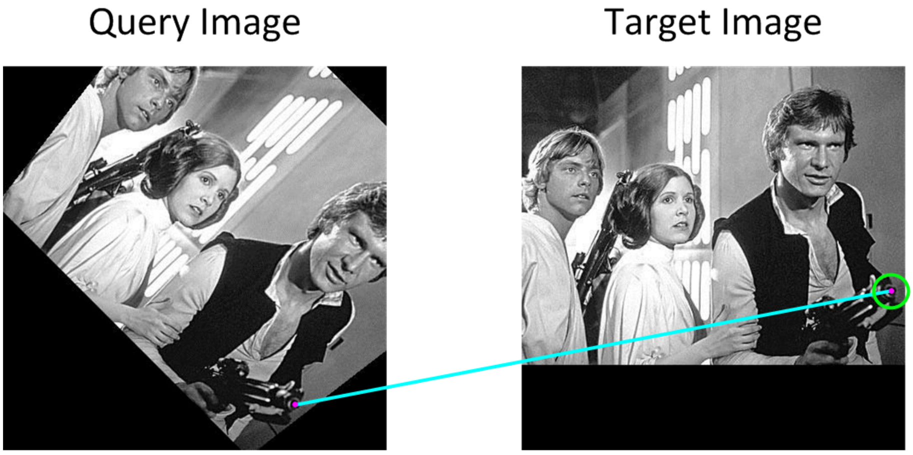


Figure 3.7: Keypoint detected in query image and corresponding location and catchment region shown in the target image

A single keypoint position in a query image can produce multiple keypoints varying in orientation, only one of these orientations is required to form a correct position and descriptor correspondence for a query keypoint to be labelled as a true positive. The other orientations are ignored and not considered in the evaluation of the SIFT variation. Similarly if all of the multiple keypoints for the same image position fail to find a descriptor match, the keypoints only count as a single failed match.

### 3.2.3 Ground Truth Transform

This section focuses on the development of 2.5D feature descriptors and increasing the performance of matching algorithms with respect to distortions characteristic of changes in 3D view point. As such, a homography transform was used as an initial approximation of changes to an image resulting from 3D out of plane pose change. The homography transformation approximates the 3D motion of points with the assumption that all points on the image lie on a plane equidistant from the camera. As part of the initial investigation, the performance of matching systems under this transformation was investigated. An outline of the implementation is given here. Setting some parameters in the homography transform to zero can allow the resulting system to create other key image transformations, such as rotations, translations and Affine transformations.

#### 3.2.3.1 Applying image warp

To apply the homography transform, the target image is resampled with a grid of new pixel locations. The grid of new pixel locations are created by normalising the original image co-ordinates to be in the range of -0.5 to 0.5 in x and y, if the image is not square the smallest dimension is zero padded so as ensure that it becomes square. The normalised pixel co-ordinates can now have the chosen warping applied. To create the homography warp a camera model as shown in Figure 3.8 is used as a mathematical model to create the effect of perspective distortion, Equations 3.4 to 3.10. This mathematical model allows changes in the camera position to be represented as a transformation matrix for the motion of pixel co-ordinate locations from one camera position to another. The resampling of the target image is taken by resampling the pixel co-ordinates $x$ and $y$ to create the new pixel co-ordinates $B_x$ and $B_y$, $s$ defines the change in scale between the images, Equations 3.11 and 3.12.

$$M_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_x) & \sin(\theta_x) \\ 0 & -\sin(\theta_x) & \cos(\theta_x) \end{bmatrix} \tag{3.4}$$

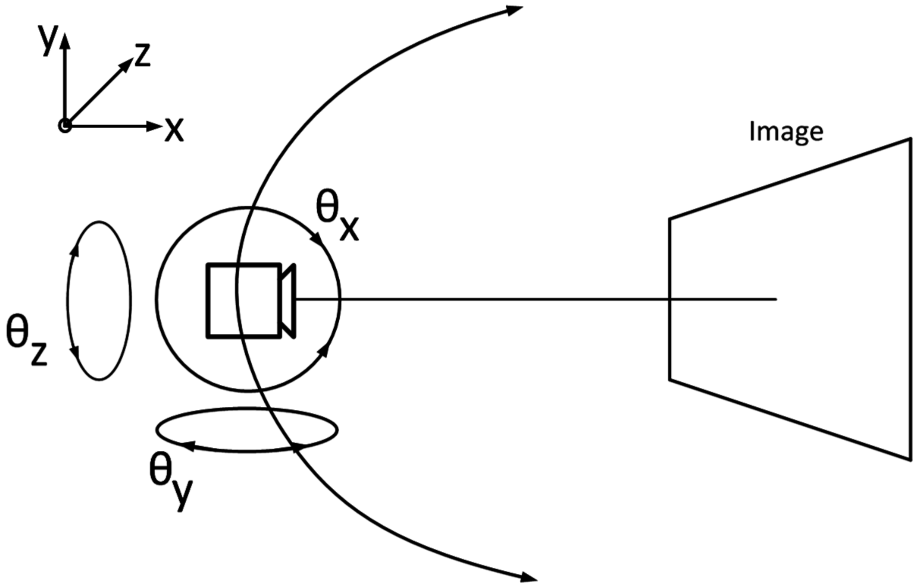


Figure 3.8: Mathematical Model

$$M_y = \begin{bmatrix} \cos(\theta_y) & 0 & \sin(\theta_y) \\ 0 & 1 & 0 \\ -\sin(\theta_y) & 0 & \cos(\theta_y) \end{bmatrix} \tag{3.5}$$

$$M_z = \begin{bmatrix} \cos(\theta_z) & \sin(\theta_z) & 0 \\ -\sin(\theta_z) & \cos(\theta_z) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.6}$$

$$M = [M_x]\,[M_y]\,[M_z] \tag{3.7}$$

$$offset = \begin{bmatrix} \sin(\theta_y) \\ \sin(\theta_x) \\ \cos(\theta_x)\cos(\theta_y) - 1 \end{bmatrix} \tag{3.8}$$

$$T = s \begin{bmatrix} [M] & [offset] \\ \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} & 1 \end{bmatrix} \tag{3.9}$$

$$D = s \begin{bmatrix} [M] & [offset] \\ \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \end{bmatrix} \tag{3.10}$$

$$B_x = \frac{D_1}{D_3} \tag{3.11}$$

$$B_y = \frac{D_2}{D_3} \tag{3.12}$$

#### 3.2.3.2 Tracking keypoints in warped image

Position correspondences between keypoints collected from query and target images are established from the image transformation used to create the synthetic query image. To find the keypoint position correspondences, the keypoint positions are normalised by the image size as done for the creating the synthetic query image. The normalised query keypoint locations then has the inverse image transformation matrix applied, Equation 3.13. The keypoint positions are then have the inverse normalisation applied, $x$ and $y$ represent the keypoint location in the synthetic image, $K_x$ and $K_y$ represent the keypoint location back projected into the model image, and $T$ represents the transformation from the model image to the synthetic image.

$$\begin{bmatrix} x \\ y \\ 1 \\ 1 \end{bmatrix} = T^{-1} \begin{bmatrix} K_x \\ K_y \\ 1 \\ 1 \end{bmatrix} \tag{3.13}$$

An example of the homography warp being applied to perform image rectification is shown in Figure 3.9. Here the cover of a book is taken at a skewed angle. This out of plane rotation and perspective distortion are then corrected. The bounds of the first image are shown in the second image to show the effect of the distortion on the image shape. Landmarks were manually placed in the distorted image on the 'H' in the book title, these are then projected back into the original image as an empirical validation test for returning keypoint locations from query images to target images.

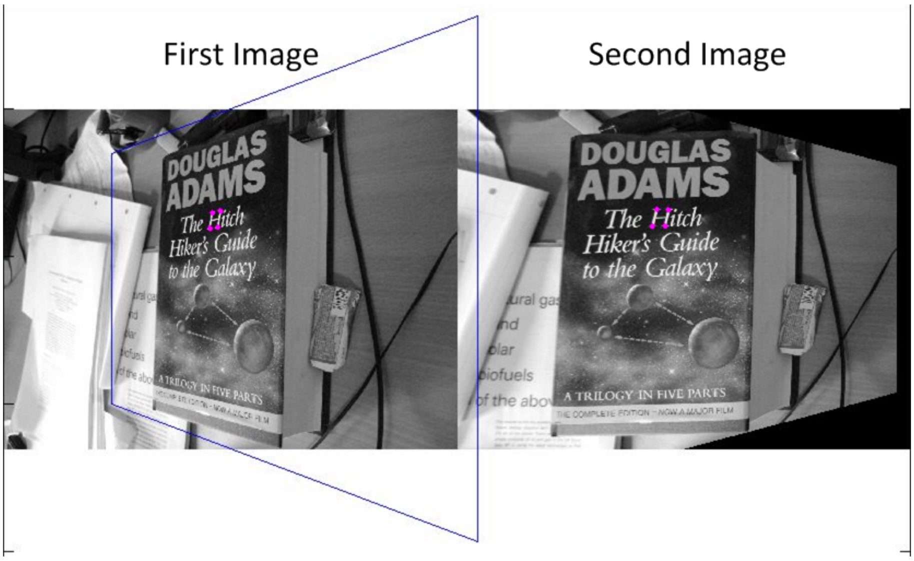


Figure 3.9: Homography Example, showing the rectification of a book cover imaged in the first image, to estimate the appearance from an new viewing perspective shown in the second image. The blue trapezoid shows the outline of the sampled region of the second image in the space of the first image. Four magenta dots manually placed in the second image, and back projected into the first image appear in the same location with respect to the image texture, validating the back projection transformation between the two images.

### 3.2.4 Error in ROC Curves

In this thesis a number of Local Feature Matchers, LFMs, are presented and compared. The performance of each LFM is presented using a graph of Receiver Operator Characteristics, or ROC curve Fawcett (2006). The ROC curve is a parametric plot of the ratio of *true positive keypoint matches* against the ratio of *false positive keypoint matches* for a varying *operating sensitivity*. Figure 3.10 shows an example of a ROC curve, where each point on the ROC curve, or ROC point, is the ratio of *true positive keypoint matches* and ratio of *false positive keypoints* at a fixed *operating sensitivity*. Modifying the operating sensitivity will *move* the ROC points along the ROC curve. The ROC performance of a LFM can be gauged by the proximity of the curve to the point of ideal operation; where the false positive ratio is 0, indicating no incorrect keypoints have been matched; and the true positive ratio is 1, indicating that all keypoints which have a correct match have been correctly matched. Therefore, the ROC curve allows the reader to visually assess the performance of a number of LFMs simultaneously and independent of any differences arising from the choice of *operating sensitivity*.

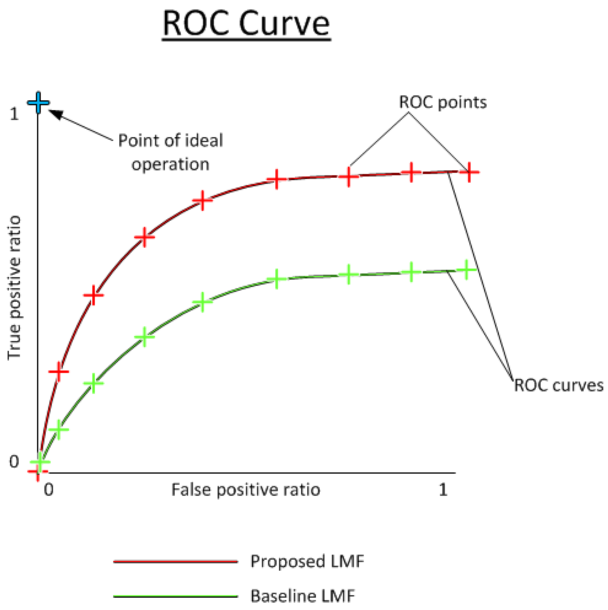


Figure 3.10: ROC Curve

The ROC curve for each new LFM is compared against the ROC curve for SIFT (Lowe, 2004), a well established LFM which serves as a baseline performance measure. In order for a given ROC curve to be recognised as demonstrating a statistically significant improvement in performance error bars are attached to each ROC point. Overlapping error bars indicate that a result is not statistically significant and that more data must be captured to reinforce the hypothesis. Whereas, non-overlapping error bars will indicate a statistically significant result.

The error bars in this work are set at a p-value of 0.05, demonstrating the 5% significance level, see Figure 3.12. This work assumes a Normal distribution of individual measurements of ROC point location. The mean value of ROC point location, $\mu$, is taken as the population mean of individual measurements of the ROC point location, $\bar{x}$. The standard deviation, $\sigma$, is calculated from the variance, $v$, and the number of ROC points used to calculated the mean, $n$, see Equation 3.14. In this work the variance is calculated as a result of the experimental design, see Subsection 3.2.4.2.

$$\sigma = \frac{v}{\sqrt{n}} \tag{3.14}$$

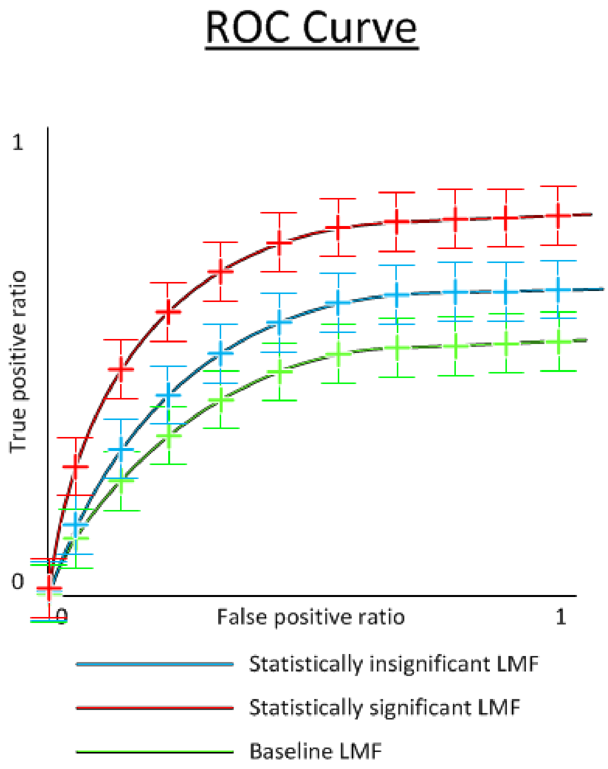


Figure 3.11: Statistically significant ROC curves

#### 3.2.4.1 Error bars

Fawcett calculates the error bars on a ROC curve from the variance of individual measurements of ROC point locations from multiple runs of the experiment (Fawcett, 2006), see Figure 3.13. Multiple runs of an experiment are used to establish the mean ROC curve for each *true positive* and *false positive rate* as the population mean of multiple ROC curve measurements interpolated at these measurements. . In this thesis the mean ROC curve, $\mu$, is the mean of multiple ROC points at a fixed operating sensitivity, $\bar{x}$, calculated from multiple experimental runs. However, the error in ground truth can be established through an analysis of the experimental design. Therefore, the ROC curve error bars are calculated from the error in the experiment ground truth, see Subsection 3.2.4.2.

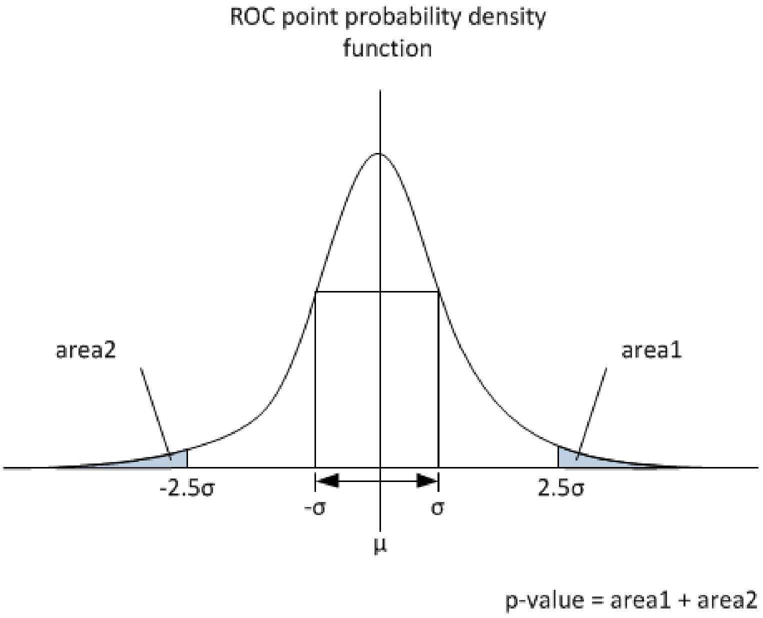


Figure 3.12: ROC point probability density function, showing mean ROC point location at $\mu$, and 5% significance at $2.5\sigma$

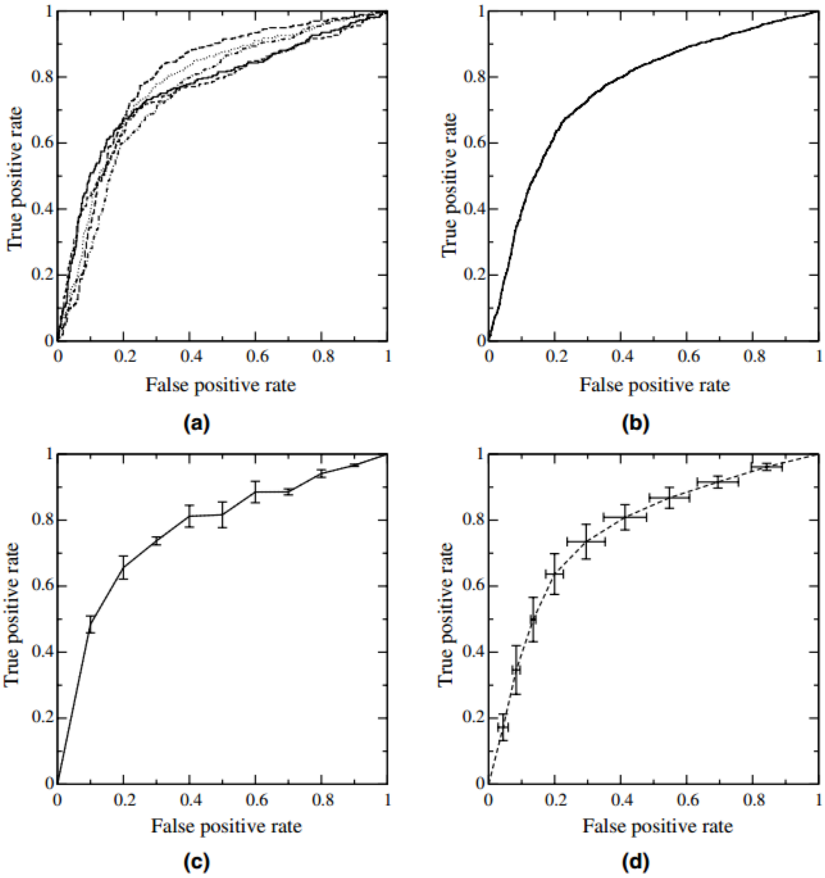


Figure 3.13: Error bars on ROC curve (subfigures c and d) established using multiple experimental runs (subfigure a) to find the mean ROC curve (subfigure b)(Fawcett, 2006)

#### 3.2.4.2 Error introduced through the experimental design

The experimental design used to investigate LFM performance uses ground truth for keypoint locations to form a decision on whether there is a keypoint which should match to a given location and, if a keypoint match does exist, whether the location of the matched keypoint is correct. In order to establish the ground truth for a keypoint location correspondence a catchment region around a keypoint location correspondence is used. The catchment region allows for a translation error in keypoint location. The catchment region is based on the exact location correspondence and the area used to form the keypoint feature descriptor, see Figure 3.7. However, the use of a catchment region also allows for the possibility of incorrectly labeled keypoints. In this work the probability of an incorrectly labeled keypoint in the experiment is used to evaluate the significance of the results.

The statistical significance of the experiment is based on the probability of incorrectly labeling any of the detected keypoints. The probability of labeling any one keypoint incorrectly is the probability of a keypoint falling within the catchment region of a keypoint with which it should not form a location correspondence, see Figure 3.14. The probability of labeling a single keypoint incorrectly can be calculated as the proportion of the image occupied by the catchment region, see Equation 3.15, where $r$ is the radius of the catchment region, $h$, is the image height, $w$, is the image width and $a$ is the probability of the keypoint being incorrectly labeled. The radius of the catchment region, $r$, is proportional to the scale of the keypoints. Here the radius, $r$, is taken as the mean keypoint scale.

$$a = \frac{\pi r^2}{hw} \tag{3.15}$$

The error in an individual experiment can now be calculated from the probability of an incorrectly labeled a keypoint, $a$, and the number of keypoints detected in the query image, $k$. Repeating the experiment $n$ times, reduces the error by a factor of $\sqrt{n}$. The standard deviation, $\sigma$, for the experimental design is shown in Equation 3.16. ROC error bars with a statistical significance level of $5\%$ can now be placed either side of the mean ROC point at $\pm 2.5\sigma$.

$$\sigma = \frac{1 - (1 - a)^k}{\sqrt{n}} \tag{3.16}$$

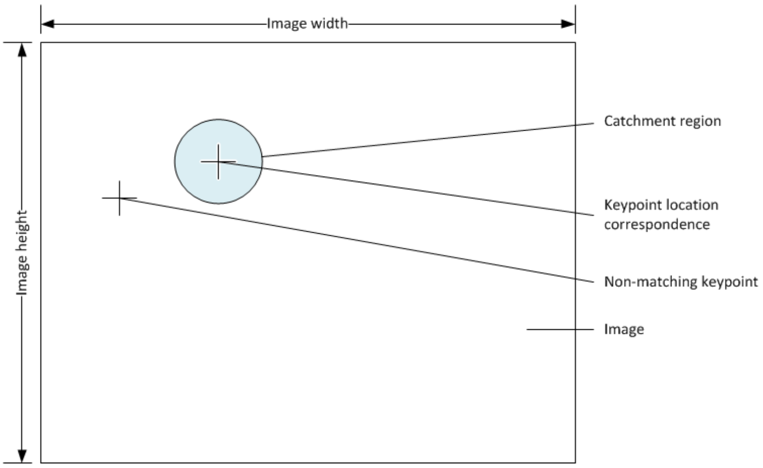


Figure 3.14: Catchment region of keypoint, shown on target image

### 3.2.5 Modes of invariance

The experimental configuration outlined in this chapter is capable of tracking keypoint instances between a model image and a synthetic image. From the comparison of keypoint position and feature descriptor matches it is possible to gauge the performance of the point based matching algorithm. The experimental configuration presented tests the invariance of point based matching algorithms to a range of transformations. However due to a lack of 3D information regarding the structure of the objects represented in the intensity image database, it is not possible to implement a number of realistic image transformations in the configuration outlined here. The configuration presented in this chapter can be used to investigate the invariance of feature matching algorithms to the following range of image transformations:

- Scale changes
- Translation changes
- Orientation changes
- An approximation to perspective change
- Gaussian image noise.

Due to a lack of 3D information in the test data, a range of image transformations could not be implemented in the experimental configuration. The range of image transformations which the experimental configuration presented in this chapter cannot examine are:

- Changes in illumination
- Changes in 3D view point
- The effects of depth of field

The ROC curve performance of the local feature matching algorithms presented in this chapter evaluate the performance of individual point-to-point correspondences between images. The combination of local feature extraction and post processing steps are not considered in this experimental configuration. This study is focused on optimising only the performance of the feature extraction stage, the resulting feature extraction algorithms can then be used in conjunction with post processing steps to further increase performance. In order that results are comparable with other studies presented in the literature all results are compared against the unmodified MATLAB SIFT implementation as a benchmark.

### 3.2.6 Overview

An overview of the system is presented here to show the order of execution of each of the steps detailed in this section.

1. **Collect image dataset**, each of the images placed in the image set folder is read into the MATLAB code.
2. **Create model keypoints for each target image**, the SIFT variation is run on all images in the target database. The image name, and keypoint locations, scales and feature descriptors for each of the target images are stored in a database ready to be used at the matching stage.
3. **Create query keypoints**, this stage is executed in a loop for the number of query examples required. Increasing the number of examples will give a better approximation of the ROC curve for the SIFT variation under examination, although it will also increase run time.
   (a) **Create transformed image**, a transformed image is created by applying the chosen image warping, or noise function with random parameters within the defined boundaries.
   (b) **Extract keypoints**, the feature descriptors, image name, location and scale are extracted and stored.

(c) **Find ground truth keypoint correspondences**, the positions and image names of each of the query keypoints are matched against the positions and image names of the target keypoints. Positional matches are stored as the ground truth for correct matches.

(d) **Find feature descriptor matches**, this stage is executed in a loop to find the performance on matching each image for every log-likelihood detection threshold. The feature matches are stored as the matches for the SIFT variation.

(e) **Associate**, find the true positives as the feature descriptor matches which also have a positional match, find the false positives as the feature descriptor matches which do not have a positional match.

(f) **Save ROC points**, the ROC point for each log-likelihood detection threshold is stored as the performance of the SIFT algorithm on the given image.

(g) **Save experimental setup statistics**, calculate the average catchment area, $a$, and store number of keypoints collected, $k$.

4. **Create the ROC curve**, the ROC curve is created by taking the average over the set of images for each of the ROC points at a given log-likelihood sensitivity level.

5. **Attach ROC error bars**, calculate the $5\%$ significance level and display on ROC curve.

## 3.3 SIFT Variations

All SIFT variations are based on freely available MATLAB SIFT code (F. El-Maraghi, 2008). This section investigates modifications to SIFT which are presented in the literature and evaluates the performance gains on the experimental configuration outlined in the previous section. This section investigates each stage of the feature extraction process and draws conclusions on the parameter choices at each stage.

### 3.3.1 Subpixel Localisation

The MATLAB SIFT code used in the experiments here does not include code to implement subpixel localisation of keypoints. In order to make the MATLAB SIFT code equivalent to other SIFT code available the subpixel keypoint localiser was implemented as detailed by Brown and Lowe (Brown

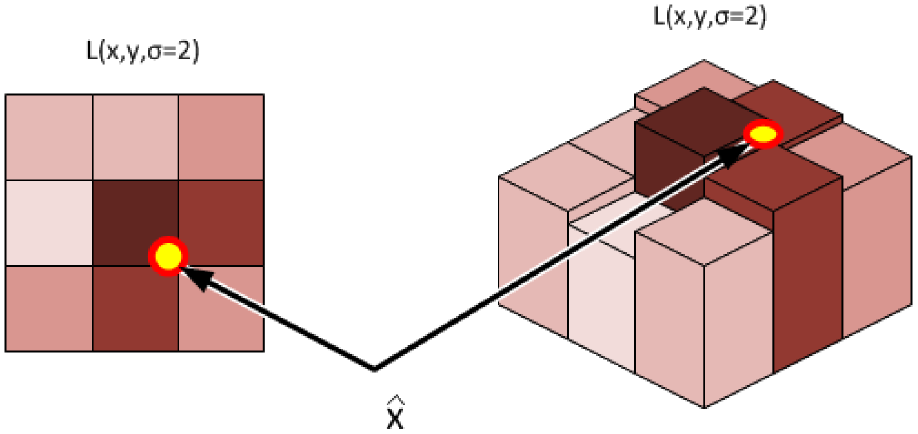


Figure 3.15: Taylor approximation for subpixel keypoint location

and Lowe, 2002; Lowe, 2004). To find the local maxima or local minima the pixels surrounding the course grain pixel localised local minima or maxima are taken as a 3x3x3 image, $L(x, y, \sigma)$, which can be fit to a second order polynomial function, see Equation 3.17. This function which approximates the data contains a turning point, which is the true location of the local minima or maxima. By finding the value of the vector $x$, representing the x, y axes of the 3x3 image with 2 levels in scale space either side, which sets the first derivative to zero, the true location of the keypoint, $\hat{x}$, can be found to subpixel accuracy, see Equation 3.18. The image patch, $L$, surrounding an area of local maximum together with the Taylor approximation of the true keypoint location, $\hat{x}$ are shown in Figure 3.15.

$$L(x) = L + \frac{\partial L^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 L}{\partial x^2} x \tag{3.17}$$

$$\hat{x} = -\left(\frac{\partial^2 L}{\partial x^2}\right)^{-1} \frac{\partial L}{\partial x} \tag{3.18}$$

Furthermore Brown and Lowe use the subpixel location of the keypoint to interpolate the grayscale value of the subpixel keypoint location, see Equation 3.19. The grayscale value $L(\hat{x}_{x,y,\sigma})$ is then used to determine whether a local maxima is sufficiently stronger than its neighbours to be accepted or rejected as a keypoint. This approach was also adopted and implemented in this section.

$$L(\hat{x}_{x,y,\sigma}) = L(x) + \frac{1}{2} \frac{\partial L}{\partial x_{x,y,\sigma}} \hat{x}^T_{x,y,\sigma} \tag{3.19}$$

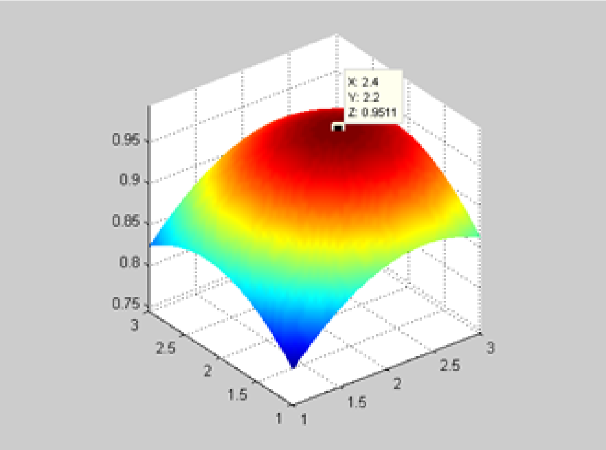


Figure 3.16: Subpixel Test Data

| | Gaussian Maxima Location | Estimated Maxima Location Output | Estimated Location with Offset | Difference |
|---|---|---|---|---|
| x | 2.4 | 0.3960 | 2.3960 | 0.004 |
| y | 2.2 | 0.1954 | 2.1954 | 0.0064 |
| $\sigma$ | 2 | 0 | 2 | 0 |

Table 3.1: Test Data Example

#### 3.3.1.1 Validation

To validate the subpixel implementation a 3x3x3 synthetic image patch was created using a Gaussian function, Equation 3.20. The standard deviation was set to be 3 pixels, the $x$, $y$ and $\sigma$ offsets can be used as variable to determine the performance of the subpixel localiser. If the subpixel localiser is working correctly it should return the same values as outputs in the $\hat{x}$ vector. An example of the subpixel localiser is shown in Figure 3.16. Table 3.1 shows the comparison of actual location against estimated location for this example. The Euclidean distance between the localised keypoint and the true Gaussian Maxima is shown in Figure 3.17 for the range of $x$ and $y$ offsets in the range of -1 to 1 for a selected range of offsets in $\sigma$. The mean of the discrepancy between the offset and Gaussian maxima in the region of -0.5 to 0.5 in $x$, $y$ and $\sigma$ offsets is 0.0069 pixels.

$$G(x, y, \sigma) = \exp\left(\frac{-(X_{sample} - x_{offset})^2 - (Y_{sample} - y_{offset})^2 - (\sigma_{sample} - \sigma_{offset})^2}{2std^2}\right) \tag{3.20}$$

#### 3.3.1.2 Results

The performance of SIFT with the inclusion of the subpixel keypoint localiser was compared using the experimental setup outlined in Section3.2. A comparison was carried out for two cases:

1. Lowe's demo C release of SIFT (Lowe, 2008) with subpixel keypoint localisation compared against Colour SIFT applied to the grayscale images, with and without subpixel localisation (Burghouts and Geusebroek, 2009), Figure 3.18a; and

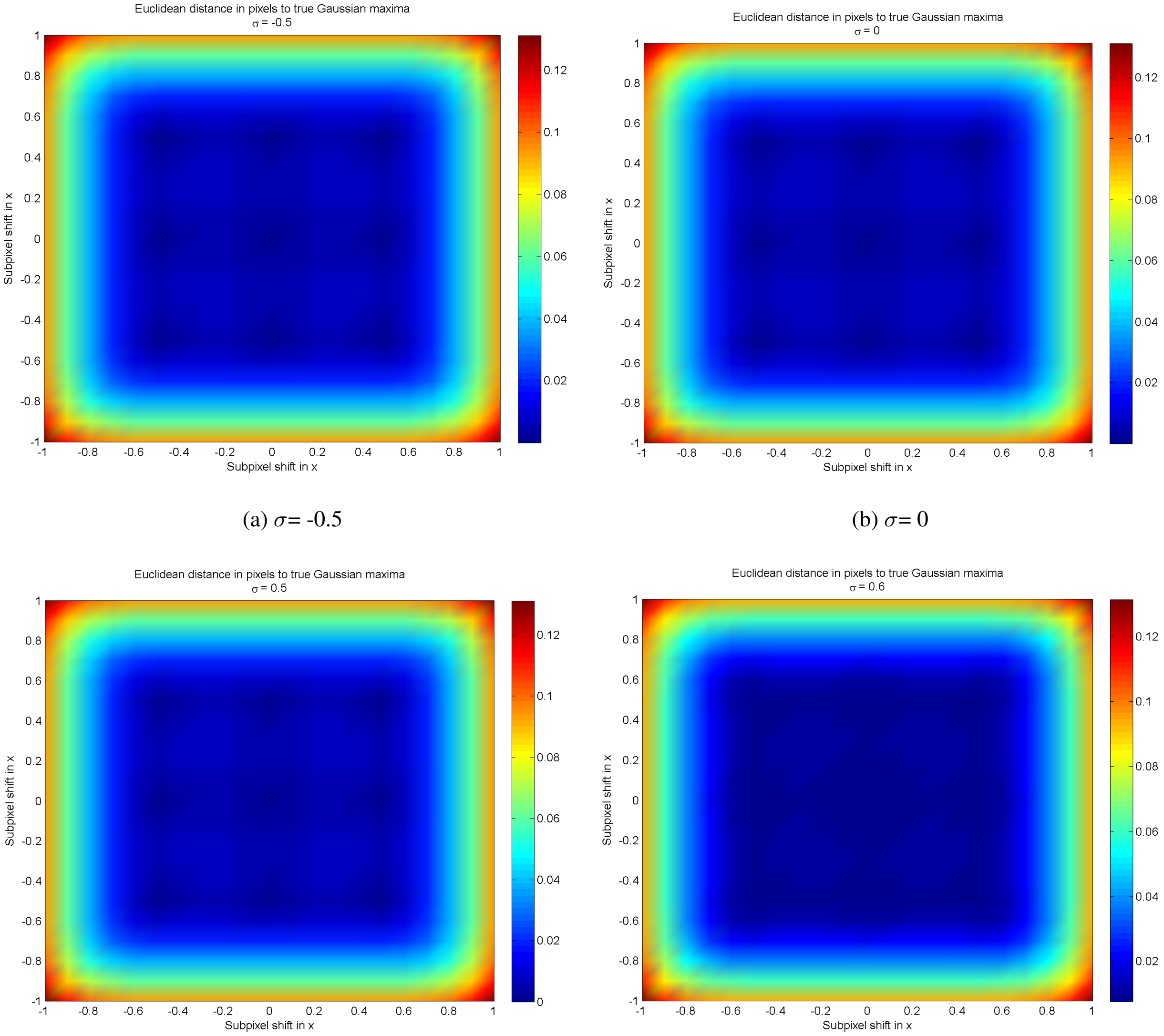
2. The Matlab SIFT implementation with and without subpixel keypoint localisation, Figure 3.18b.

The Colour SIFT demo release is based on Lowe's Demo SIFT release however contains source code where the subpixel localiser can be disabled. A comparison of Lowe's Demo release of SIFT and Colour SIFT shows that near identical results are produced when subpixel localisation is included. However, when the subpixel localisation is removed the performance of the feature descriptor match quality between transformed images is substantially improved. In the Matlab SIFT implementation a similar result is observed, where the inclusion of subpixel localisation degrades the match quality between the transformed images.

This result implies that the Gaussian approximation for curve fitting which shows an accurate subpixel location fitting in Figure 3.17 does not hold for real image features. Furthermore the Gaussian approximation assumes even spaced divisions in scale space which is an exponential dimension compared to the linear dimensions of x, y image space.

### 3.3.2 Canonical Orientation

When a keypoint is found and localised, SIFT assigns to the keypoint a repeatable orientation which is then used for rotational normalisation, this step give the keypoint descriptor a degree of in-plane orientation invariance. The canonical orientation is found from the direction in which most of the



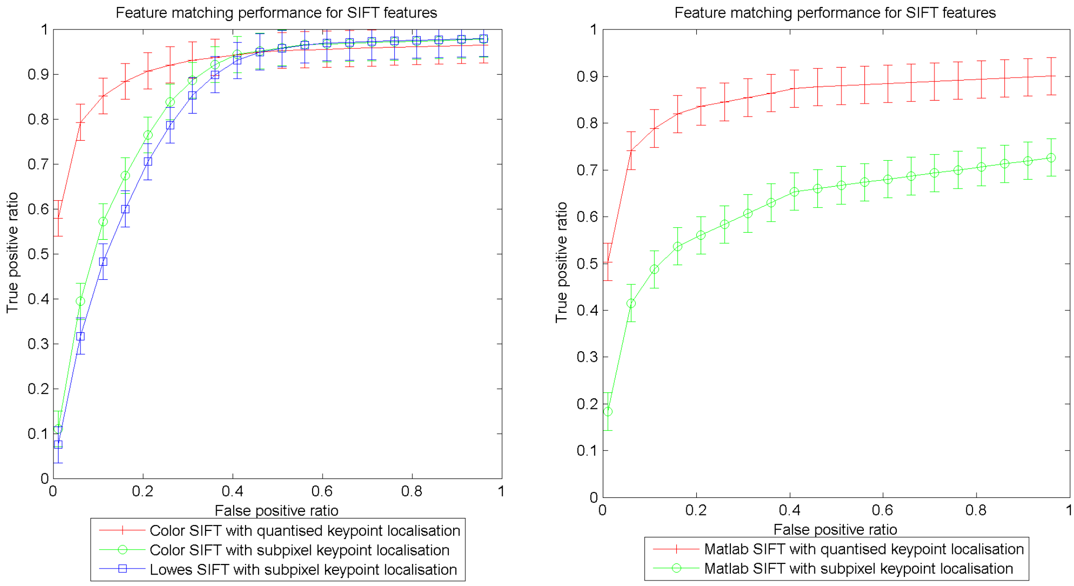
(a) $\sigma$= -0.5

(b) $\sigma$= 0

(c) $\sigma$= 0.5

(d) $\sigma$= 0.6

Figure 3.17: Difference between estimation and actual location of Gaussian maxima, blue represents a small error, red represents a large error



(a) C SIFT Subpixel Comparison

(b) Matlab SIFT Subpixel Comparison

Figure 3.18: Subpixel Results

image gradients surrounding that keypoint are oriented, the sample patch is then aligned with this orientation. The orientation of the patch is quantised to allow a histogram approach to be used in order to determine the dominant orientation of the patch. It has been proposed that by increasing the canonical orientation resolution it is possible to increase the reliability of keypoint matches (Lo, 2009). This section investigate the effects of canonical orientation quantisation.

In standard SIFT the canonical orientation of a keypoint is found by examining the histogram of the image gradients of the area surrounding a keypoint. The support region for the canonical orientation estimation is taken from a 11x11 pixel patch in the Gaussian image pyramid. The 11x11 pixel sample patch surrounding the keypoint represents a patch of 1.5x the sigma of the scale of the keypoint at that level in the Gaussian pyramid. The gradients from the area surrounding the keypoint are then weighted with a Gaussian so that the gradients closest to the keypoint location have greatest influence on the canonical orientation. The sigma of the weighting Gaussian is 1.5x the scale of the keypoint at the level in the Gaussian pyramid. The weighted gradients surrounding the keypoints are then histogrammed, non maximal suppression is then applied to eliminate any bins in the histogram which are not peaks; the highest peak is assigned as the canonical orientation. As there may be multiple peaks which dominate the orientation histogram under differing lighting conditions; any peak with a magnitude within 80% of the main peak is also assigned as a keypoint with its own canonical orientation.

2.5D SIFT uses an increased resolution for canonical orientation assignment. This is implemented by setting the quantisation level for the canonical orientation histogram to 1° and filtering the orientation histogram (Lo and Siebert, 2009; Lo, 2009). The higher canonical orientation resolution places the feature descriptor sample patch over the keypoint with a higher angular resolution. It is hypothesised that the higher angular resolution will cause the corresponding keypoint feature descriptors to be better aligned and more similar throughout rotation changes. The filtering is intended to reduce noise in the canonical orientation histogram before assigning keypoints.

#### 3.3.2.1 Results

The optimum performance of SIFT is obtained using a canonical orientation resolution between 5° and 20°, Figure 3.19. Increasing the canonical orientation resolution to 1° significantly decreases the performance of the local feature matcher, as does increasing the resolution past 20°, where 36° represents a significant decrease in performance. These results suggest that the choice of canonical orientation resolution is influenced by the orientation resolution of the feature descriptor, which in these experiments is set to 45°, 8 orientation bins to cover 360° for each receptive field. Using the feature descriptor orientation resolution as the influencing factor sets the lowest usable canonical orientation resolution to the Nyquist frequency of angular divisions every 22.5°. Furthermore, the reduced performance of the increased canonical orientation resolution can be explained by the example where multiple keypoints with differing orientations are found at a single location. When using the higher canonical resolution with the descriptor resolution fixed at 45°, identical descriptors may be extracted for two keypoints at the same location with differing canonical orientation. When matching query descriptors to the database using the log-likelihood matching criteria it is likely that the first and second nearest neighbours will have been extracted from the same location. These keypoints will therefore have identical descriptors and result in the nearest neighbour being rejected as below the log-likelihood threshold, regardless of the distinctiveness of the descriptor to other descriptors in the database.

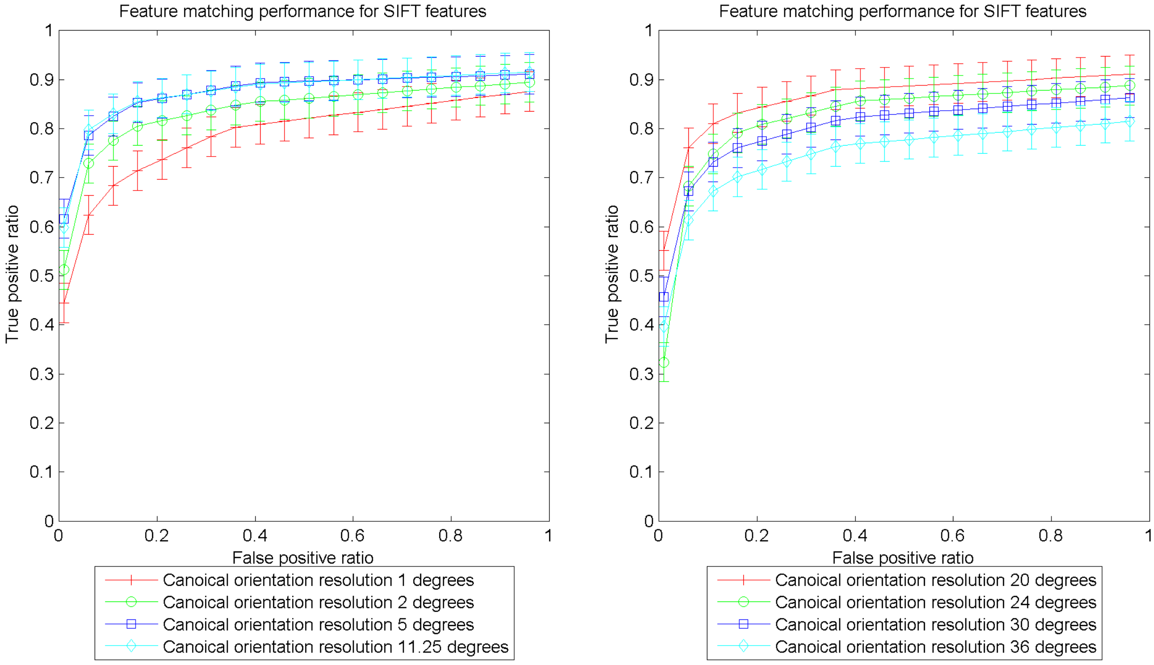


Figure 3.19: Canonical Orientation Performance

### 3.3.3 Receptive Field Configurations

To build a unique descriptor of a keypoint localised in position, scale and orientation, a series of histograms of the image gradients are taken from a sample patch surrounding the keypoint. To form the feature descriptor for a keypoint, the keypoint sample patch is divided into a series of receptive fields; each receptive field creates a weighted histogram of the image gradients it covers, Figure 3.20. The keypoint feature descriptor is created by concatenating each of the resulting receptive field histograms, and normalising to unit length. The choice of receptive field arrangement defines the spatial sampling of the the local area surrounding a keypoint. A number of variations of the receptive field arrangement have been proposed. In this section an evaluation of these has been conducted.

The receptive field configurations investigated in this section are created using a look-up table of weightings to find the influence that each pixel in the sample patch will have on each descriptor element. Each receptive field histogram in the resulting feature descriptor is calculated from the weighted influence of the strength and orientation of the gradients of each pixel in the sample patch. The weighted influence for each pixel in the sample patch is found from the receptive field

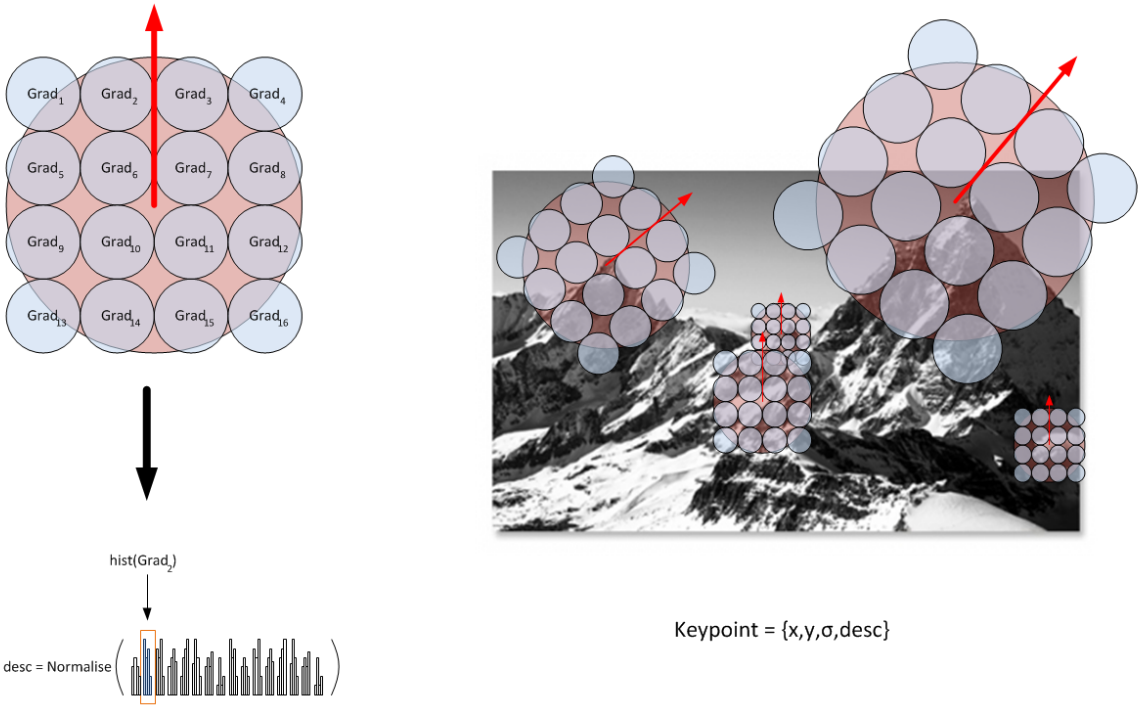


Figure 3.20: Receptive fields in SIFT

look-up table shown in Figure 3.21. Any receptive field configuration can be generated by creating a series of weighting images. The advantages of this approach are that the weighting function for the receptive field configuration only needs to be calculated once and that any modifications made to the sample patch configuration, such as rotations, are also applied to the receptive field configuration.

The receptive field configurations investigated in this section are:

1. **Standard SIFT,** 4x4 rectilinear configuration of receptive fields each with a sigma of 2 pixels and separation of 2 sigma between receptive fields. 16 receptive fields giving a feature length of 128, Figure 3.22.

2. **2.5D SIFT,** 3x3 rectilinear configuration of receptive fields with a overlap of 1 sigma between fields to reduce anti-aliasing, each field has a sigma of 4 pixels to cover the same sample patch foot print as standard SIFT (Lo and Siebert, 2009). 9 receptive fields giving a descriptor length of 72, Figure 3.23.

3. **Modified 2.5D SIFT configuration,** 2.5D SIFT configuration with Standard SIFT receptive field size and non-overlapping fields. 9 receptive fields giving a descriptor length of 72,

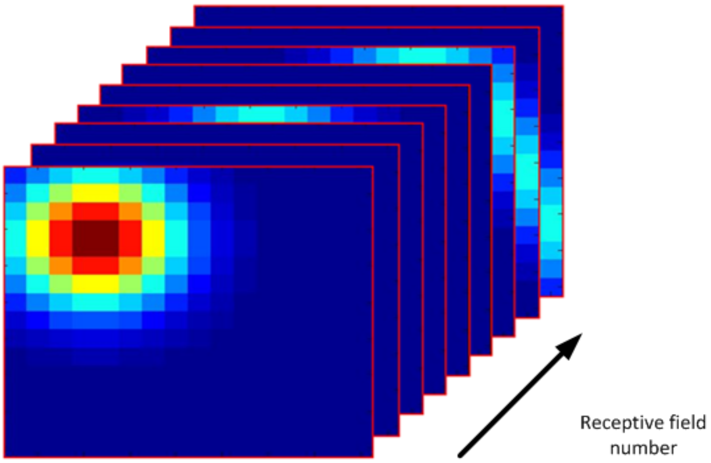


Figure 3.21: Pixel weighting values for each receptive field

Figure 3.24.

4. **Dense sampled configuration,** Combination of 2.5D SIFT and Standard SIFT configurations, giving the same sample patch representation as Standard SIFT with the anti-aliasing invariance of 2.5D SIFT. 25 sample patches giving a feature descriptor length of 200, Figure 3.25.

5. **Foveated receptive field configuration,** the S3-16 configuration has shown a performance comparable to the state of the art while maintaining a descriptor length of 128 (Winder and Brown, 2007), Figure 3.26. As the receptive fields in this configuration differ in size, to ensure that each field will have the same influence in the resultant feature descriptor the sum under the area of each receptive field in this arrangement is equal to 1.

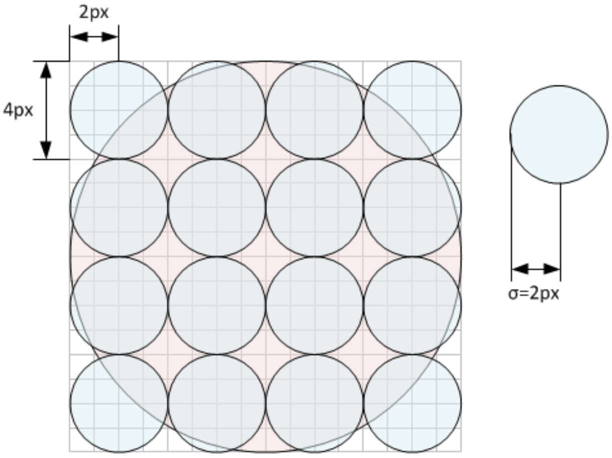


Figure 3.22: Standard SIFT Receptive Field

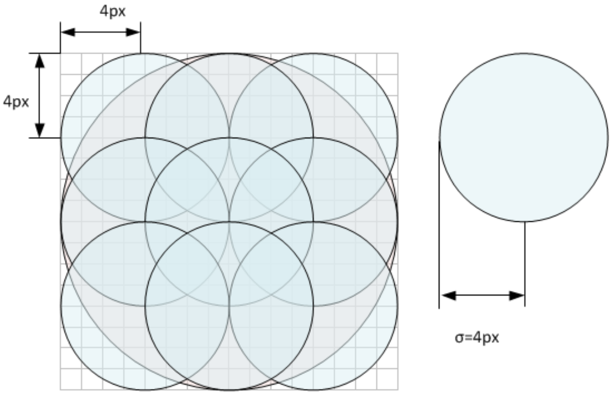


Figure 3.23: 2.5D SIFT Receptive Field

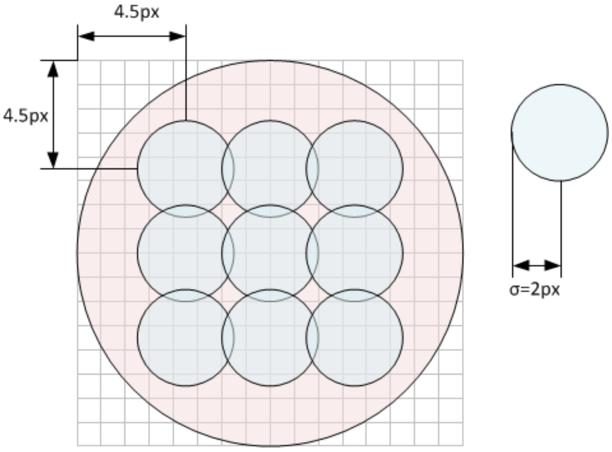


Figure 3.24: 3x3 Receptive Field

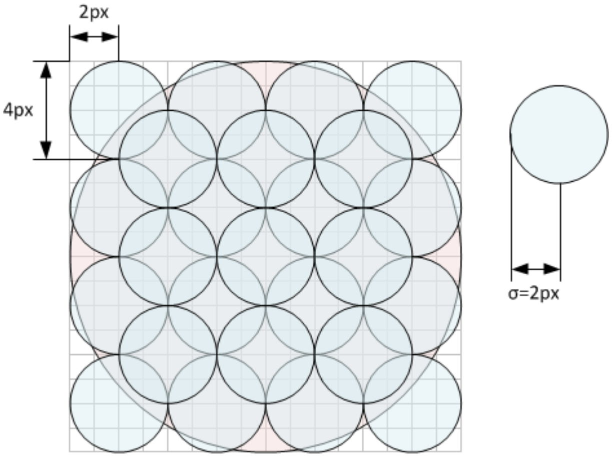


Figure 3.25: 25 Receptive Field configuration

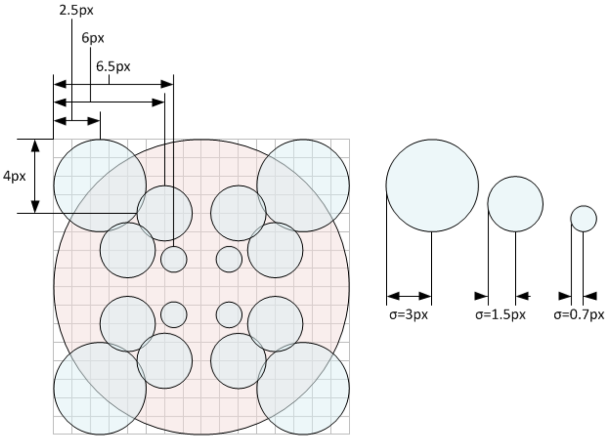


Figure 3.26: S4-16 Receptive Field configuration

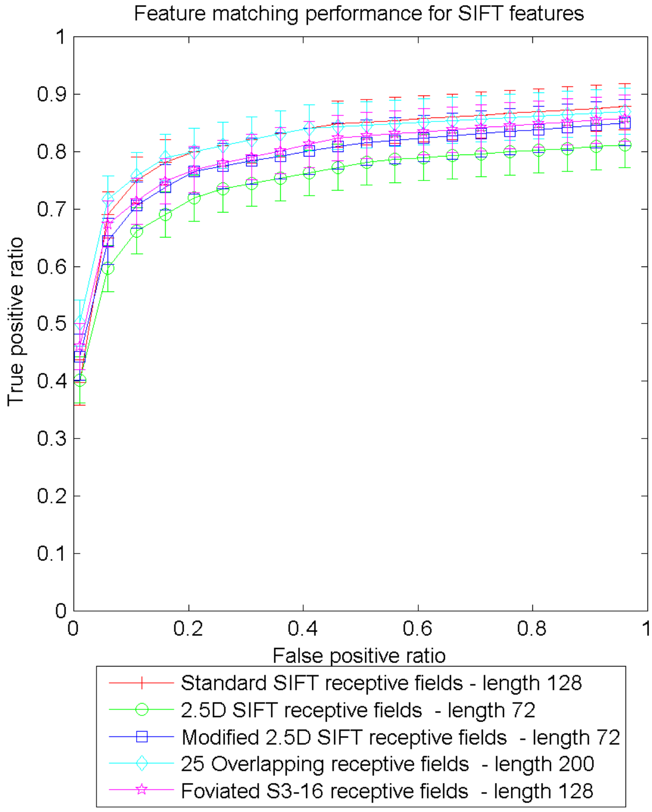


Figure 3.27: Receptive Field configurations Results

#### 3.3.3.1 Results

Figure 3.27 shows the ROC curves with the performance of the receptive field sampling schemes outlined in this section. The experiments conducted here show that the standard SIFT receptive field configuration achieves the highest performance out of the tested receptive field configurations. The dense sampled receptive field configuration achieves similar performance to standard SIFT however has an increased descriptor length. This results indicates no significant performance benefit is gained from including intermediate receptive fields to increase invariance to anti-aliasing. The receptive field configuration used in 2.5D SIFT gives a statistically significant lower performance than the standard SIFT configuration despite 2.5D SIFT receptive fields covering the same area of sample patch.

The foveated receptive field configuration implemented in this study does not give the same performance improvement as seen in the study where it was originally presented (Winder and Brown, 2007). The reduced performance observed in this experiment is not statistically significant, however, may be a result of the fixed sample patch criteria applied to all the sampling approaches presented here. A further study into the locality of information surrounding a keypoint and the effects of foveation in accounting for this using increased sample patch sizes, may show the performance improvements observed for the S3-16 sample patch; however is beyond the scope of this study.

### 3.3.4 Surface representation

To create a feature descriptor for each keypoint, the local feature matching algorithms used here sample a measurement of the image for each pixel in the sample patch. In standard SIFT the image measurement for each pixel is surface gradient magnitude and orientation. However, the range image modality offers more information regarding the underlying structure of a free form object, shape index and curvature are surface measures which use 3D information to create an invariant surface measurement (Koenderink and van Doorn, 1992). A number of matching approaches have adopted this surface measure for 3D object description (Hetzel et al., 2001; Atmosukarto et al., 2010; Guo et al., 2010; Lo and Siebert, 2009; Bayramoglu and Alatan, 2010; Zeng et al., 2010). Shape index and curvature measure a region of 3x3 pixels surrounding a pixel to generate the surface measure, gathering more information regarding surrounding pixels than surface gradient approaches. This section applies shape index and curvature methods to 2D images to establish whether these surface measures can contribute to standard 2D image matching.

The shape index feature descriptor is implemented by replacing the pixel gradient orientation bins in the feature descriptor with shape index bins. The shape index bins have their influence on the overall feature descriptor weighted by the pixel curvature instead of the pixel gradient magnitude, as formulated in 2.5D SIFT (Lo and Siebert, 2009). The Mean, $H$, and Gaussian Curvatures, $K$, are calculated from the first, $f_x$, $f_y$, and second derivatives, $f_{xx}$, $f_{yy}$, $f_{xy}$, of the image gradients at the pixel location $i, j$, see Equations 3.21 and 3.22. The Gaussian and Mean Curvatures are then used to find the principal curvatures, $k_1, k_2$, see Equation 3.23. Using the Gaussian, Mean and principal curvatures the Shape index, $S$, and Curvature, $C$, are calculated for every pixel in the sample patch, see Equations 3.24 and 3.25.

$$H(i,j) = \frac{(1+f_y^2(i,j))f_{xx}(i,j) + (1+f_x^2(i,j))f_{yy}(i,j) - 2f_x(i,j)f_y(i,j)f_{xy}(i,j)}{2\left(\sqrt{1+f_x^2(i,j)+f_y^2(i,j)}\right)^3} \tag{3.21}$$

$$K(i,j) = \frac{f_{xx}(i,j)f_{yy}(i,j) - f_{xy}(i,j)}{\left(1+f_x^2(i,j)+f_y^2(i,j)\right)^2} \tag{3.22}$$

$$k_1, k_2 = H \pm \sqrt{H^2 - K} \tag{3.23}$$

$$S = \frac{2}{\pi}\text{arctan}\left(\frac{k_2+k_1}{k_2-k_1}\right) \tag{3.24}$$

$$C = \sqrt{2H^2 - K} \tag{3.25}$$

#### 3.3.4.1 Results

Figure 3.28 shows the performance of using Shape index and Curvature as an intensity surface measure. On the intensity images used in this investigation the Shape index and curvature SIFT does not perform as well as the standard orientated gradients SIFT. The transformations in viewing angle applied by this experimental setup do not reflect the surface of the object where shape index and curvature may form an invariant description.

## 3.4 Summary and Discussions

In this chapter a performance evaluation methodology is outlined. This was then used to investigate a number of modifications made to SIFT. The main aims of this chapter are to develop existing MATLAB SIFT code and investigate potential changes proposed in the literature which could offer performance improvements. In doing this we introduced a framework for developing the code into the current state of the art for range image point based matching, 2.5D SIFT. Modifications to each stage in the SIFT processing pipeline have been outlined and benchmarked against the unmodified

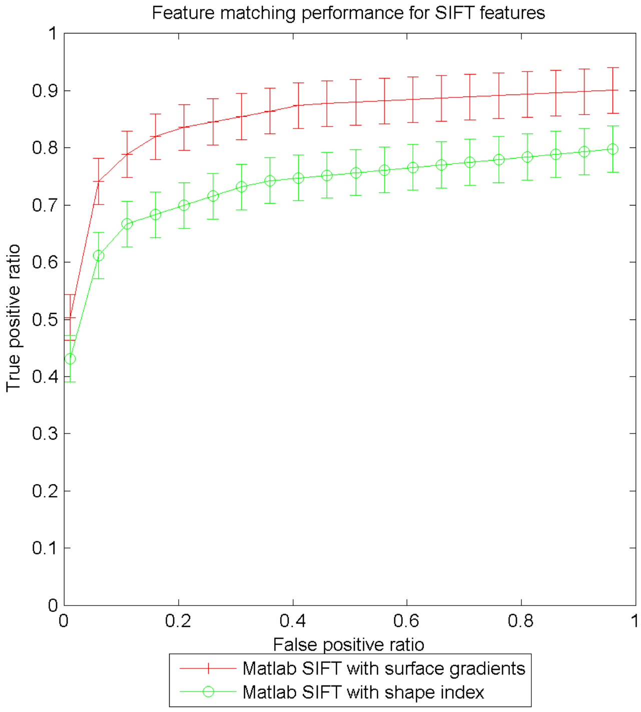


Figure 3.28: 2D Shape Index Results

SIFT MATLAB code. The main outcomes from these experiments are summarised here.

Section 3.3.1 found that contrary to existing implementations of SIFT, subpixel keypoint localisation by fitting a Taylor approximation to the pixels surrounding keypoint locations dramatically decreased the performance of the resulting local feature matching algorithm. The results indicated that the use of a second order polynomial fit to the pixels surrounding a keypoint location is insufficient to characterise the turning point in local patch to subpixel accuracy. This stage in the feature extraction pipeline will therefore be omitted in future implementations of SIFT in this study. The divisions in canonical orientation of a keypoint were found to be proportional to the divisions in angle in the feature descriptor; with standard SIFT the feature descriptor allocates the image gradients into a histogram with bin divisions of $45°$, the most effective canonical orientation estimation divisions should therefore have divisions of about half this angle, $\sim 20°$. A range of receptive field configurations were investigated, these were principally designed to investigate 2 sampling concepts proposed in the literature: receptive field configurations with an overlap of greater than 1 sigma, and a foveated configuration. The optimum receptive field configuration for the sample patch size and configuration used in this study was found to be the original SIFT configuration where the receptive fields cover every pixel in the sample patch and represent these within the half power point overlap of a receptive field in the configuration. Using overlapping receptive fields or adding intermediate receptive fields to the original SIFT configuration was found not to increase the matching performance. These modifications do however increase the feature descriptor length lowering the information density of the feature descriptor. Shape index and curvature are used in 2.5D SIFT as a surface description measure. In this chapter an initial step towards developing the MATLAB SIFT into 2.5D SIFT was taken by implementing shape index and curvature measures in the feature extraction stage. The resulting Shape index and curvature SIFT performance on the 2D images was compared against the surface gradients measures used in SIFT. It was found that this surface measure did not improve the performance of matching on the 2D image database. However it is hypothesised that 2.5D range image data will behave differently from 2D intensity data and the benefits of Shape index and curvature will be evident in the range domain.

The experimental setup used in this chapter is capable of investigating the invariance of local feature matching algorithms to a range of image transformations and image noise. The performance of the individual feature descriptors to repeatable apply the same keypoint feature descriptor at the same image location, scale and orientation in transformed query images, is investigated. Around 400 individual keypoints are extracted from a single image giving a population size of around 8000 keypoints from which to form an algorithm performance measure. ROC curves are used to present the results, showing the trade off between correctly and incorrectly identified keypoints for each algorithm, independent of sensitivity threshold. The experimental setup presented in this chapter is limited to linear 2D image transformations, and unable to apply transformations typical of realistic data such as, out-of-plane changes in viewing angle which cause non-linear motion of pixels relative to each other between viewing instances, and illumination changes resulting from object motion relative to a light source. Furthermore the transformations applied to the intensity images in this chapter do not account for the behaviour of range images to the same transformations applied in the range domain. The next chapter addresses these issues and extends the experimental setup outlined here to the range domain.

# Chapter 4

# 3D Evaluation Approach

This chapter extends the methodology presented in the initial investigation to allow for invariance of local features to 3D out of plane rotations and illumination effects. The data is collected as range and intensity images allowing for co-registered multimodal features to be extracted and aligned between single instances of object views. This chapter presents the methodology adopted for calibrating the 3D stereo capture rig and turntable configuration used in the evaluation of the 3D local features investigated throughout the remainder of the thesis.

## 4.1 Objectives

The objective of this chapter is to establish a methodology whereby the performance of keypoint matching approaches proposed in subsequent chapters may be evaluated. The data collected in the process outlined in this chapter has the following desirable characteristics, which at the time of publication were not currently available in other test data sets:

- Images are collected by a stereo pair of cameras and form co-aligned intensity and range images.
- The observations of the objects represented in both range and intensity images have real world noise characteristics.
- The observations of the objects represented in the images exhibit out-of-plane changes in view angle.

- The observations for intensity images have illumination changes associated with out-of-plane changes in viewing angle.
- Ground truth for the motion of all points on the object surface is available, allowing the repeatability of individual keypoints to be established.

The approach used in the previous chapter tracked the motion of all points on a 2D image, allowing the repeatability of the individual keypoints in isolation of others detected in the image to be evaluated. However, the approach used in the previous chapter was limited to 2D images and as such realistic transformations resulting from changes in 3D viewing position, which causes non-linear changes in the position of points in the image, could not be investigated. Furthermore, the evaluation used in the previous chapter lacked depth information, therefore the performance of keypoint extraction on range image measurements could not be evaluated. To account for these limitations this chapter uses intensity and range images of an object captured in a controlled setting, where the motion of all points on the object surface between observations is known.

The approach taken in this chapter is to use the depth and camera model information from the range image and camera calibration to find the X, Y, Z world space location of points on the object surface. A measurement of the 3D transformation applied to the object between viewing instances is used to form a high resolution estimate of the motion of keypoint location between viewing observations. To establish this ground truth, a computer controlled turntable is used to control the 3D motion of the object and a stereo capture rig is used to find the X, Y, Z location of points on the object. The computer controlled turntable is calibrated in order to create a mathematical model for the transformation it applies to the object. Using this transformation, the location of keypoints from both query and target image are projected into a canonical space where their locations can be compared. With the keypoint locations projected into a canonical space the proximity criteria from the previous chapter can then be used to establish the performance of feature descriptor matching, with the results presented as a ROC curve.

The remainder of this chapter is organised as follows, Section 4.2 gives an overview of the approach taken for 3D location comparison, Section 4.3 details the approach used to calibrate the turntable and form the ground truth, Section 4.4 details the data capture procedure, Section 4.5 outlines the performance evaluation, and Section 4.6 concludes the chapter with a summary.

## 4.2 Overview of approach

This section presents an overview of the 3D evaluation approach outlined in this chapter. Figure 4.3, shows the stereo camera and turntable configuration used to image and actuate the the objects. Figure 4.2, shows a general overview of the system implemented in 3 stages in execution order, these are: system calibration, data capture and keypoint matching evaluation. The remainder of this section gives an overview of these stages:

1. **Calibration,** The stereo pair of cameras are calibrated using the C3D calibration routine (Ju et al., 2003). The turntable is calibrated using an approach outlined in Section 4.3.

2. **Data Capture,** Once calibrated the cameras and turntable form a system where the elements cannot be moved with respect to each other. Using this arrangement, a stereo pair of intensity images of the objects are captured at 5° intervals for a full 360° range of out-of-plane view changes in the turntable axis of rotation, which is closest aligned with the yaw axis, see Figure 4.1. From the captured intensity images and the camera calibration information from step 1, a range image for each observation is constructed using C3D (Ju et al., 2003). The co-aligned intensity and range image pairs for all observations form the observation database. This process is detailed in Section 4.4.1.

3. **Target Database Construction,** From the captured observation database a smaller database of target models can be constructed. The target models are created from observation instances in the observation database whose separation is 60° . From the target observations keypoints are extracted and the instance is removed from the observation database. Using the turntable calibration data, the position of the extracted keypoints is then transformed into the canonical space for the object.

4. **Query Database Construction,** From the remaining instances in the observation database a database of query observations is constructed. From the observation database a query database is formed from 20 randomly selected models which are then used as samples to evaluate the performance of a keypoint extraction approach. Similarly as for the target database construction, keypoint locations from the models in the query database are transformed into the canonical space for the object using the turntable calibration data.

5. **Keypoint Association,** Ground truth for keypoint descriptor matches is formed from a comparison of keypoints locations in the model and the query observations. The location of keypoints extracted from a query observation are compared to the location of keypoints extracted from the closest match in the target model database. The closest match in the target database is the observation instance from the same object, which has the closest angular distance to the query model, see Figure 4.1.

6. **Keypoint Comparison,** keypoint feature descriptors extracted from the query images are compared to the keypoint feature descriptors extracted from the nearest target image. The descriptors are compared using the log-likelihood threshold. The matches for all threshold sensitivity levels are recorded. The keypoint descriptor matches are compared against the keypoint location matches from the previous stage; the keypoint location matches are used as ground truth for the keypoints whose descriptors should also match. Results are recorded as the ratio of correct descriptor matches per image over the total descriptor matches per image, the *true positive ratio*, and the ratio of incorrect descriptor matches per image, the *false positive ratio*, for all log-likelihood sensitivity thresholds.

7. **Display results as a ROC Curve**, the average of the true positive ratios and the false positive ratios for all log-likelihood threshold levels, across the 20 query images are used to form the ROC for the local feature matching algorithm being evaluated. The ROC curve is formed as a plot of true positive ratios against false positive ratios for all log-likelihood sensitivity thresholds.

## 4.3 Turntable Calibration

This dissertation aims at improving the reliability of keypoint feature descriptor matches between observations of an object. In Section 4.5, the performance of the keypoint feature descriptor in the query image is evaluated as its ability to match to the correct keypoint extracted from the target

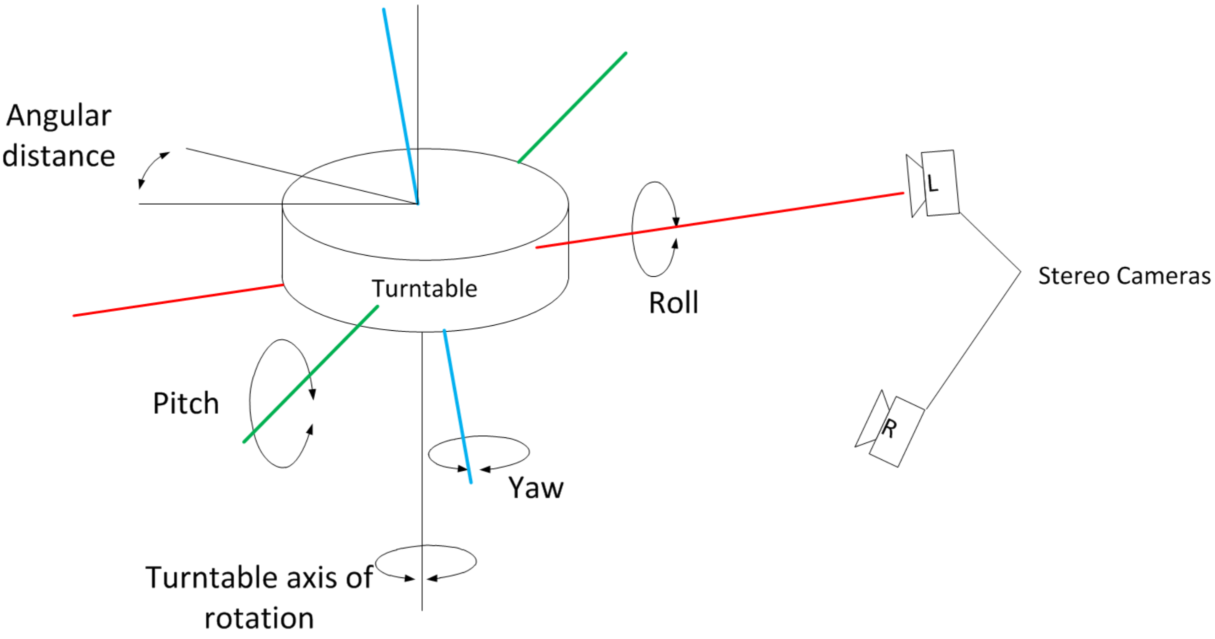


Figure 4.1: Turntable with rotation axes shown with respect to the left camera and turntable axis of rotation. The angular distance shown is the distance moved by the turntable between two observations.

image. However, in order to perform this evaluation the location of keypoints in both the model and the query instances must be transformed into a common canonical space where their keypoint locations may be compared. This ground truth for keypoint locations is achieved using a computer controlled turntable to control the motion of the object between observations. This section outlines the methodology for formulating the turntable calibration data.

### 4.3.1 Processing Range Data

The turntable setup used in this experimental configuration is calibrated using range and intensity image data collected by imaging a calibration target at a number of turntable positions. The protocol for calibrating the turntable outlined here can therefore be applied to any turntable setup where range and intensity image data are available and the angular motion of the turntable is known. This section describes the processing of the range and intensity images of calibration targets into calibration points which are then used in sections 4.3.2 and 4.3.3 for calculating the turntable transformations.

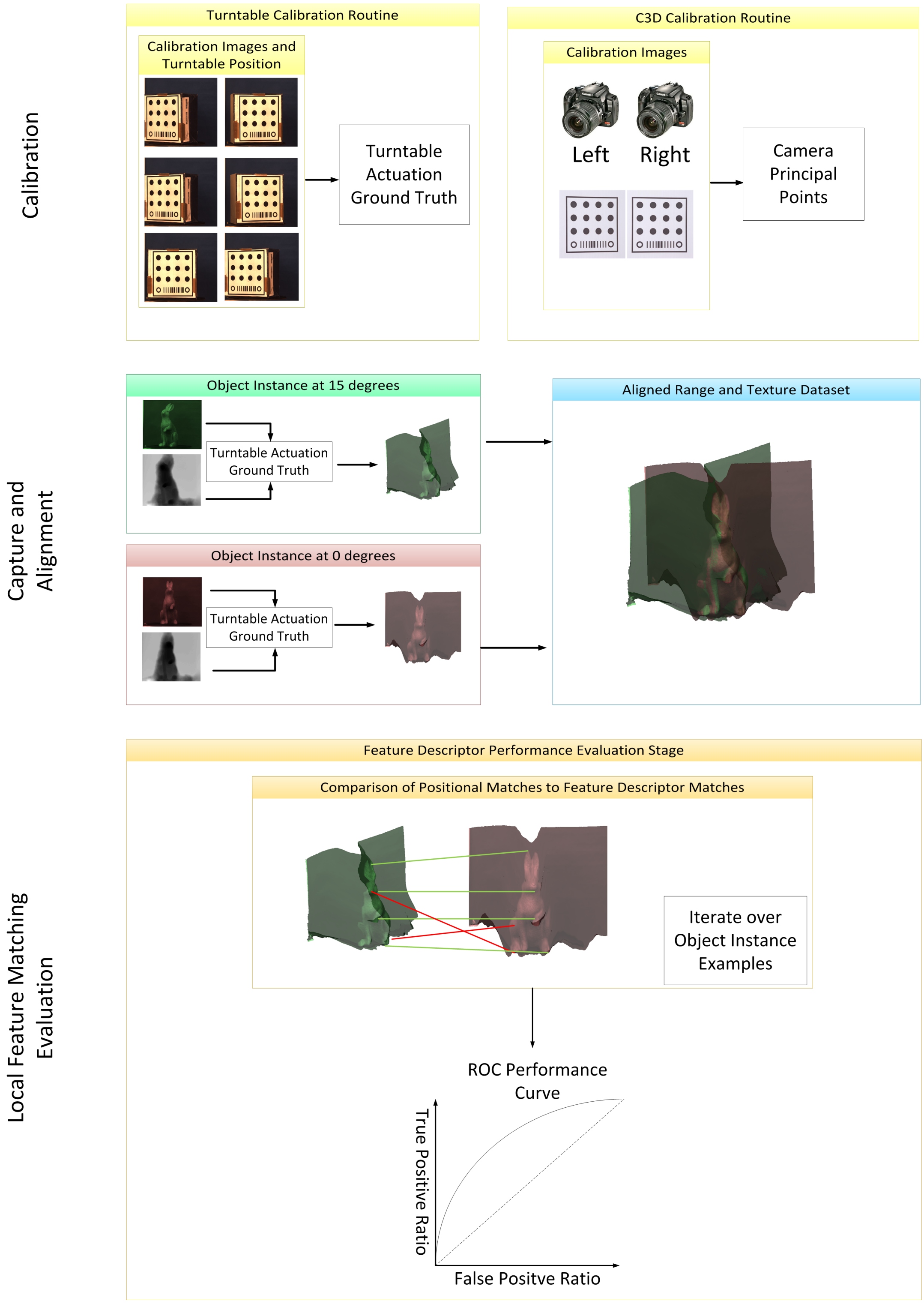

Calibration
Turntable Calibration Routine
Calibration Images and Turntable Position
Turntable Actuation Ground Truth
C3D Calibration Routine
Calibration Images
Left
Right
Camera Principal Points
Capture and Alignment
Object Instance at 15 degrees
Turntable Actuation Ground Truth
Object Instance at 0 degrees
Turntable Actuation Ground Truth
Aligned Range and Texture Dataset
Local Feature Matching Evaluation
Feature Descriptor Performance Evaluation Stage
Comparison of Positional Matches to Feature Descriptor Matches
Iterate over Object Instance Examples
ROC Performance Curve
True Positive Ratio
False Positve Ratio

Figure 4.2: Benchmark Overview

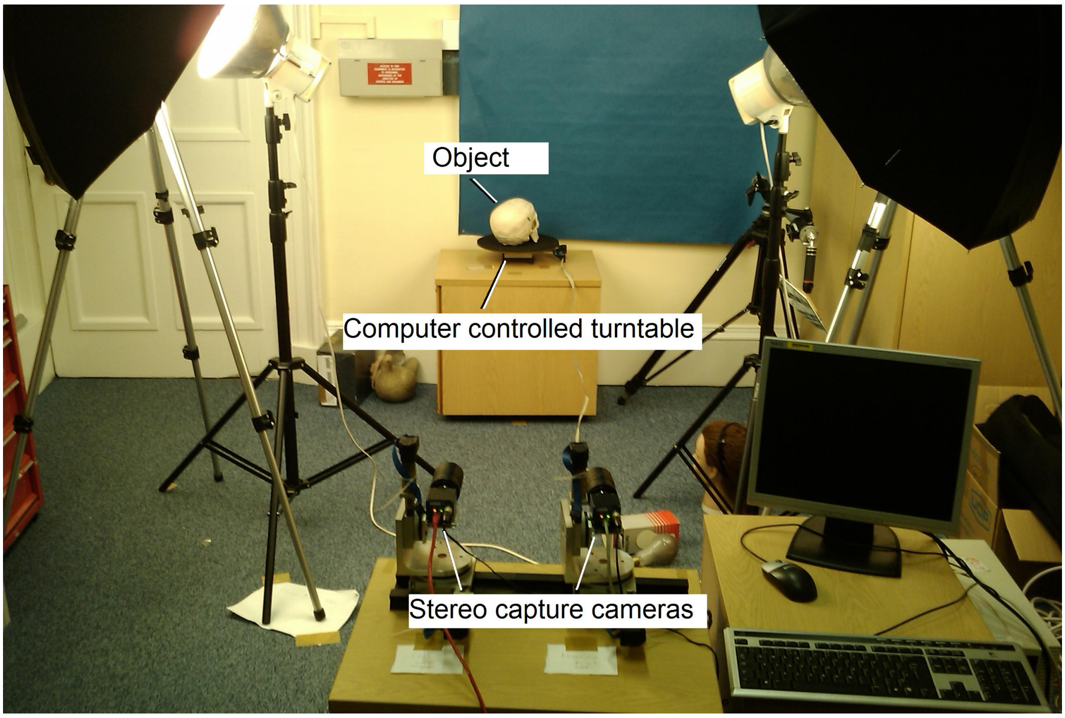


Figure 4.3: Data Collection Setup

#### 4.3.1.1 Calibration points

A set of reliable and well-localised calibration points are detected in the intensity image of the calibration target. These are a set of known positions which are consistently localised in consecutive images of the calibration target. Calibration points are used in C3D for camera calibration (Ju et al., 2003), therefore the code which performs the subpixel localisation and labelling of the calibration points was available for use in this application. In this experimental configuration the calibration points are found using the C3D circle centre finder; however other well localised calibration points such as the checker patterns used in many OpenCV applications may be used (Gabor, 2011).

#### 4.3.1.2 Range Images

Range maps used in this setup are built in C3D from a stereo pair of images. In each of the stereo-pair a local pixel-wise match is found between the left and right images. The location of pixel-wise matches are determined from the cross correlation of local patches which are refined through a scale space to find a pixel match for all locations in the stereo pair of images (Siebert and Marshall, 2000). To aid the cross correlation local patch matching process, a Gaussian noise field is added to the calibration target, Figure 4.4. The Gaussian noise field simulates a speckle pattern projected on to an object in some 3D capture systems (Dekiff et al., 2010; Siebert and Marshall, 2000). To aid the detection of the circles, the Gaussian noise is not applied around the edges of the circles. To further reduce errors in the resulting range data an average over 10 intensity images of 5M pixel resolution was used.

The range images of the calibration targets are built in C3D using the combination of the location of pixel matches from the disparity map, and prior knowledge of the perspective centres of the cameras, gained from camera calibration (Cyganek and Siebert, 2011; Hartley and Zisserman, 2005). The range images are co-aligned with the left camera intensity image.

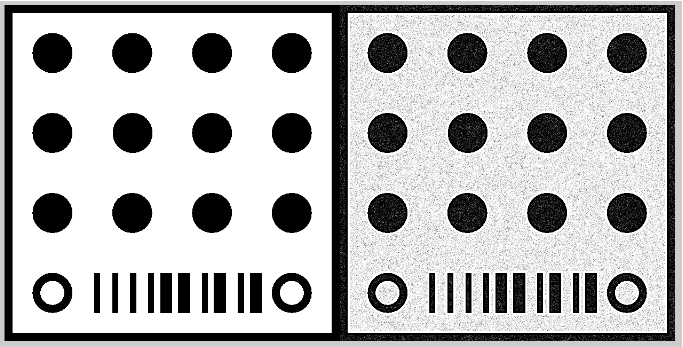


Figure 4.4: Turntable Calibration Target

#### 4.3.1.3 Correcting Circle Centre Range

To further increase confidence in the range values of the calibration points, the calibration points have their range values fit to plane using least means square which takes account of the range of all pixels covering a region on the calibration target. The 4 corner calibration points create a polygon segmenting a planar region on the calibration target where all values within may be used to calculate the equation of a plane in 3D. The range values of the calibration points may then be corrected to lie on the least squares fitted plane.

The centre of the target is segmented using the 4 corner calibration points, see Figure 4.5. The 4 calibration points at corners of the calibration target trace a polygon, inside which all range values and their pixel locations are stored in a 3xn dimensional matrix, where n is the number of samples in the polygon region. To fit a least means square plane to these points a covariance matrix is formed from the stored 3xn matrix, $\text{Cov}(loc_{3xn})$, and the principal components of the covariance matrix are then computed from the Eigenvector, $x$, and Eigenvalue, $\lambda$, decomposition of the covariance

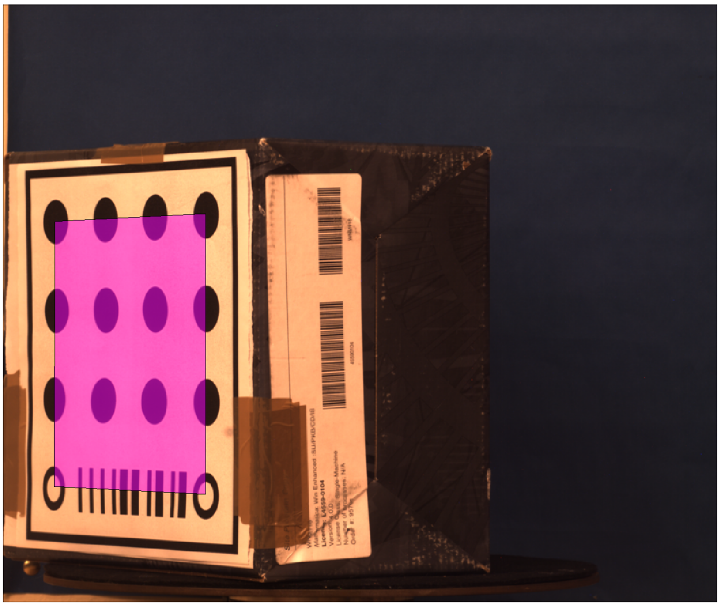


Figure 4.5: Valid region showing the pixel locations used to estimate the equation of the calibration target plane

matrix, see Equation 4.1.

$$[\lambda, x] = \mathrm{Eig}(\mathrm{Cov}(loc_{3\times n})) \tag{4.1}$$

From the formulated eigenvalue matrix, the lowest eigenvalue indicates the direction in which the data shows the lowest variation. The eigenvector with the lowest corresponding eigenvalue is taken as the normal to the plane on which all points of the calibration target lie. To complete the formulation of the equation of the calibration target plane a point on the plane is required. The point $p_0$ is the mean of the 3xn matrix of all points within the valid region, see Figure 4.6.

$$n = x[\min(\lambda)] \tag{4.2}$$

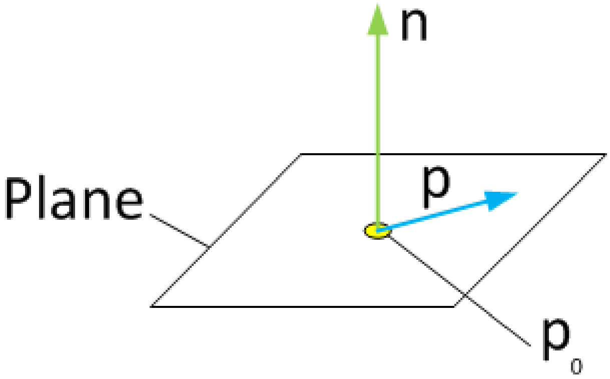


Figure 4.6: Plane geometry

$$plane : n.(p - p_0) = 0 \tag{4.3}$$

The range values, of the calibration points before correction are replaced with the corresponding Z value for the X and Y position on the least means square plane, see Equation 4.4, the value shown here as $Z$, in Equation 4.5.

$$plane : \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} \cdot \begin{bmatrix} X - p_{0x} \\ Y - p_{0y} \\ Z - p_{0z} \end{bmatrix} = 0 \tag{4.4}$$

$$Z = p_{0z} - \frac{n_x(X - p_{0x}) + n_y(Y - p_{0y})}{n_z} \tag{4.5}$$

#### 4.3.1.4 Pixel Scaling

The specification of the printed calibration target defines that the physical separation between adjacent circle centres is set to 4.8cm, in both X and Y directions. From this measurement the scaling of X and Y in the observed images can be determined as a function in the turntable calibration routine. As a requirement for the camera calibration routine the calibration points are localised to subpixel

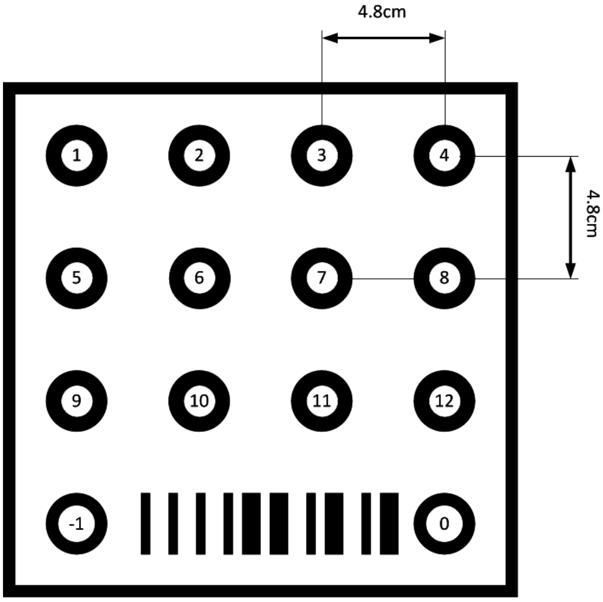


Figure 4.7: Calibration target dot numbering

accuracy. As all calibration points are accurately localised to subpixel accuracy in X and Y in the intensity image, the scaling in X and Y may be calculated by use of any pair of orthogonal and adjacent calibration points. The distance between each of the dots in a pair is known to be 0.048m. In this section two pairs of dots in a calibration target image are required to find the scaling in X and Y, dots numbered 3 and 7 are chosen as the first pair and dots numbered 3 and 4 are chosen as the second pair, see Figure 4.7. The scaling in $x$ and $y$ are given by the values $a$ and $b$, and are measured in metres/pixel.

$$0.048^2 = a^2(x_3 - x_7)^2 + b^2(y_3 - y_7)^2 + (z_3 - z_7)^2 \tag{4.6}$$

$$0.048^2 = a^2(x_3 - x_4)^2 + b^2(y_3 - y_4)^2 + (z_3 - z_4)^2 \tag{4.7}$$

The equations 4.6 and 4.7 are then solved for $a$. The value of $a$ can then be substituted into the equation 4.6 to find the value of $b$.

$$a = \sqrt{\frac{0.048^2(\frac{(y_3-y_4)^2}{(y_3-y_7)^2} - 1) - \frac{(y_3-y_4)^2}{(y_3-y_7)^2}(z_3 - z_7)^2 + (z_3 - z_4)^2}{\frac{(y_3-y_4)^2}{(y_3-y_7)^2}(x_3 - x_7)^2 - (x_3 - x_4)^2}} \tag{4.8}$$

### 4.3.2 Hough Transform

The Hough Transform is a feature extraction technique whereby image data may be characterised as a histogram of parameters. This geometric interpretation of an image can be used as representation for describing an image or the composition of features within an image. The Hough Transform was originally implemented as a means for finding features such as lines and ellipses in images as a basic low-level feature for image description (Duda et al., 1995). This was later extended to form the Generalised Hough Transform. The Generalised Hough Transform is capable of representing arbitrary shapes, by performing a composition analysis where locations in the image vote on a consistent representation of the object (Ballard, 1981). In the turntable calibration routine a consistent representation of the changes in parameters between observations from differing turntable rotations is required. In order to achieve this representation, a 3D Hough Transform is used to vote on the common transformation between observations. The 3D Hough Transform has been reported by Khoshelham (Khoshelham, 2007) and subsequently and independent to this work extended by Tombari and Di Stefan (Tombari and Di Stefano, 2010) to histogram the transformation parameters as a set of absolute angles. However, both these approaches apply an R-table Hough Transform as described by Ballard (Ballard, 1981), whereas the approach described in this section implements the Hough transform as a more efficient linearised transformation matrix. This section discusses implementations of the 2D Hough Transform as a linearised transformation matrix approach as a way to show how this approach may be extended into the 3D domain.

#### 4.3.2.1 2D Hough Transform

SIFT uses an equivalent of the Hough Transform in 2 stages of its matching routine. The first use is a coarse level culling of erroneous keypoints through histogramming of the parameter differences between the model and the query keypoints (Pope and Lowe, 2000). The second use is the parametrisation of smaller sets of keypoints with RANSAC applied to the parametrised form to find keypoints which conform with a consensus (Lowe, 2004). As a mathematical convenience the output of the RANSAC method is treated as similar to the Generalised Hough Transform and is discussed as a means of extending the parametrisation representation into the 3D domain.

In the experimental design described here the transformation matrix for the motion of keypoints on the turntable is extracted from the common parametrisation. Equation 4.9, shows the transformation between a point in the model and query images for extracting the differences in translation, scale changes and rotations, see Figure 4.8a. Histogramming the parameters of the transformation from a single keypoint match is possible in this equation due to the constraints on the rotation matrix, and the known canonical orientation $\theta$.

$$\begin{bmatrix} x'_1 \\ y'_1 \end{bmatrix} = \sigma \begin{bmatrix} \sin(\theta) \\ \cos(\theta) \end{bmatrix} + \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \tag{4.9}$$

When the canonical orientation information is unavailable the transformation may be expressed as an linear transformation between a set of two points, see Equation 4.10. From this equation the change in orientation, $\Delta\theta$, between the two sets of points may be calculated from the variables $c_1$ and $c_2$, see Equation 4.11. The transformation between the two point sets are shown in Figure 4.8b.

$$\begin{bmatrix} c_1 \\ c_2 \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} x'_1 & -y'_1 & 1 & 0 \\ y'_1 & x'_1 & 0 & 1 \\ x'_2 & -y'_2 & 1 & 0 \\ y'_2 & x'_2 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \end{bmatrix} \tag{4.10}$$

$$\Delta\theta = \arctan\left(\frac{c_2}{c_1}\right), \Delta\sigma = \frac{c_1}{\cos(\theta)} \tag{4.11}$$

The second implementation finds the transformation as an Affine Transform. This is implemented as a linearisation between 3 sets of model and query points. The transform described in Equations 4.12 and 4.13 has the advantage that it is capable of calculating the orientation of the point set as a value for each element in the rotation matrix, without requiring the keypoint canonical orientation, see Figure 4.8c.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{4.12}$$

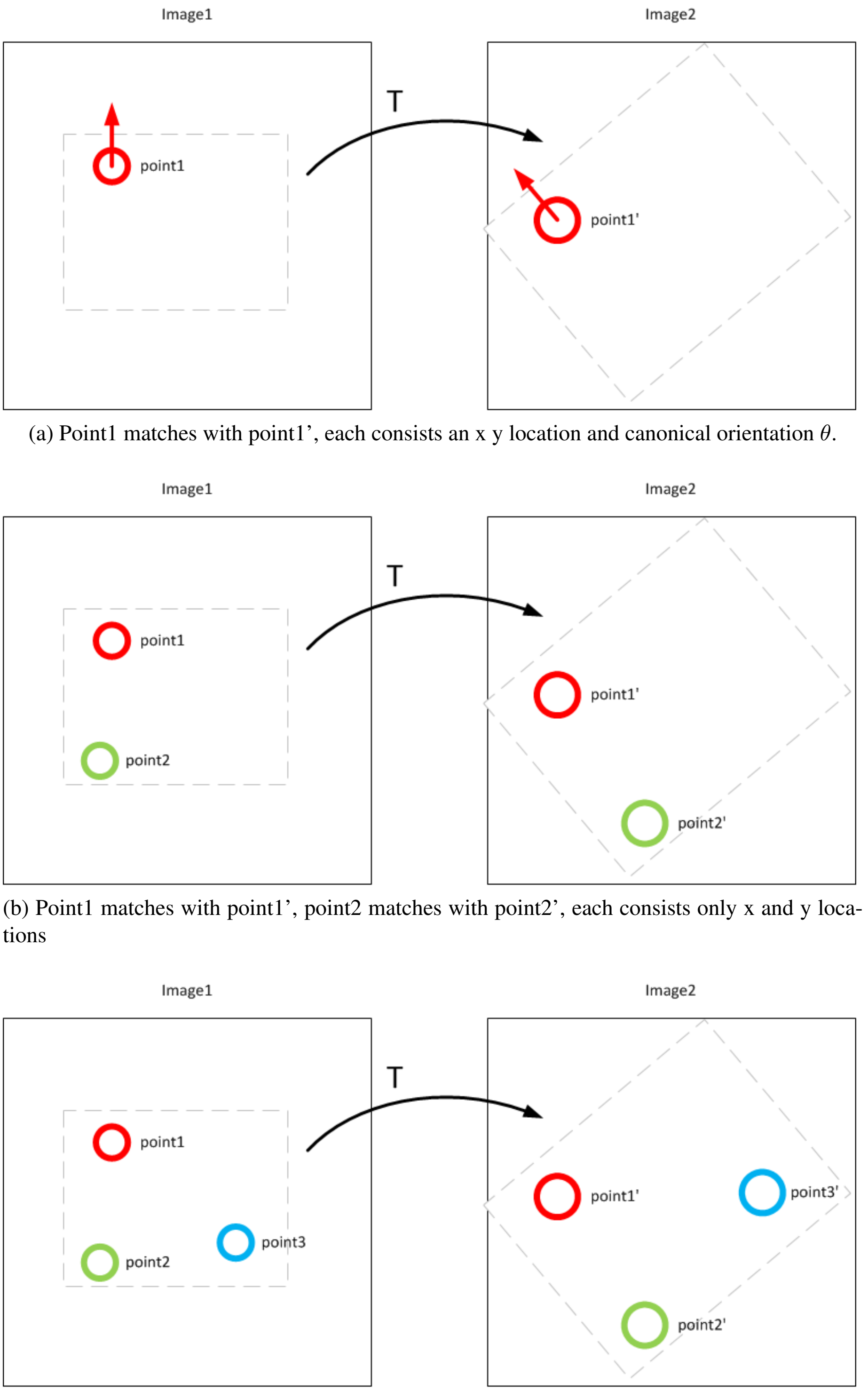
$$\begin{bmatrix} x'_1 \\ y'_1 \\ x'_2 \\ y'_2 \\ x'_3 \\ y'_3 \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_1 & y_1 & 0 & 1 \\ x_2 & y_2 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_2 & y_2 & 0 & 1 \\ x_3 & y_3 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_3 & y_3 & 0 & 1 \end{bmatrix} \begin{bmatrix} m_{11} \\ m_{12} \\ m_{21} \\ m_{22} \\ t_x \\ t_y \end{bmatrix} \tag{4.13}$$

#### 4.3.2.2 3D Hough Transform

The 3D Hough Transform has been described for use in 3D data as an R-table approach (Khoshelham, 2007; Tombari and Di Stefano, 2010). However, here we use an extension of the 2D linearised Hough Transform to find the transformation between the calibration keypoints located on turntable calibration target. In the 2D Hough, 3 keypoints were sufficient to constrain each element in the Affine Transform between two sets of points. However, to constrain the rotation matrix in 3D, 4 keypoints are required. In the turntable calibration routine the points used to constrain the transformation between instances are taken as the 3 calibration points from the corners of the target, the 4th keypoint is determined as the cross product of the 3 calibration points. The 3 calibration points from the calibration target and the 4th synthetic calibration point are shown in Figure 4.9.

$$A = \begin{bmatrix} x & y & z & 1 \end{bmatrix}^T \tag{4.14}$$

$$B = \begin{bmatrix} nx & ny & nz & 1 \end{bmatrix}^T \tag{4.15}$$



(a) Point1 matches with point1', each consists an x y location and canonical orientation $\theta$.

(b) Point1 matches with point1', point2 matches with point2', each consists only x and y locations

(c) Three keypoint correspondences between two images, all consist only of x and y locations

Figure 4.8: Establishing transformation, T, between images using keypoint correspondences

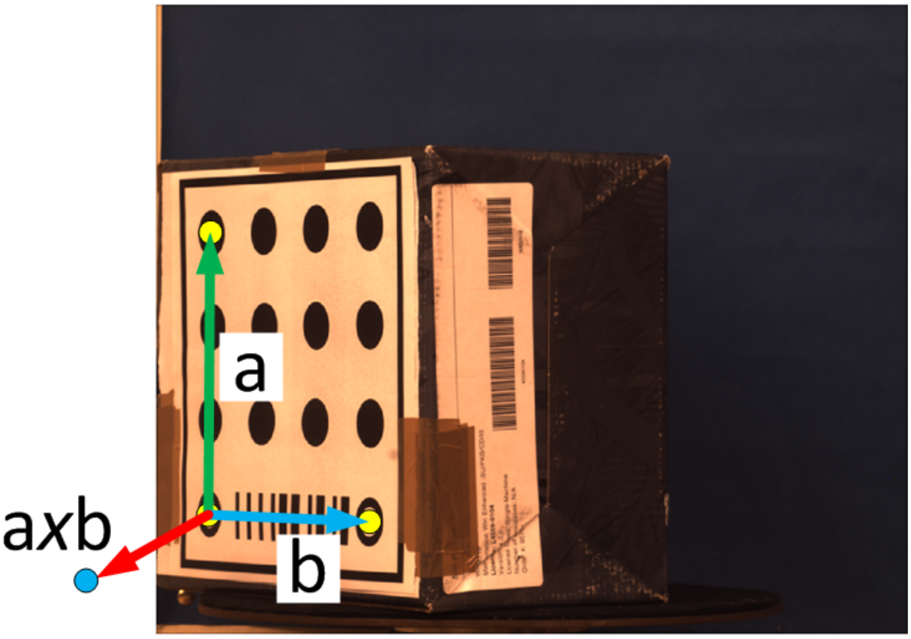


Figure 4.9: Constraint Calibration Points

$$C = \begin{bmatrix} m_{11} & m_{12} & m_{13} & t_x \\ m_{21} & m_{22} & m_{23} & t_y \\ m_{31} & m_{32} & m_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.16}$$

$$B = CA \tag{4.17}$$

The transformation between a set of calibration points from the first target and second target are represented by Equations 4.14 and 4.15. The calibration point set in the first target are represented by $\{x, y, z\}$ and the calibration point set in the second target are represented by $\{nx, ny, nz\}$. The transformation between these two point sets is defined by the matrix $C$, Equations 4.16 and 4.17. The rotation matrix may then be linearised as done in Section 4.3.2.1 to calculate the Affine Transform between two point sets. The linearised form of the matrix $C$ is shown in Equation 4.20, and calculated from Equation 4.21, where Equations 4.18 and 4.19 show the effects on matrices $A$ and $B$ of linearising matrix $C$.

$$A = \begin{bmatrix} x_1 & 0 & 0 & y_1 & 0 & 0 & z_1 & 0 & 0 & 1 & 0 & 0 \\ 0 & x_1 & 0 & 0 & y_1 & 0 & 0 & z_1 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_1 & 0 & 0 & y_1 & 0 & 0 & z_1 & 0 & 0 & 1 \\ x_2 & 0 & 0 & y_2 & 0 & 0 & z_2 & 0 & 0 & 1 & 0 & 0 \\ 0 & x_2 & 0 & 0 & y_2 & 0 & 0 & z_2 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_2 & 0 & 0 & y_2 & 0 & 0 & z_2 & 0 & 0 & 1 \\ x_3 & 0 & 0 & y_3 & 0 & 0 & z_3 & 0 & 0 & 1 & 0 & 0 \\ 0 & x_3 & 0 & 0 & y_3 & 0 & 0 & z_3 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_3 & 0 & 0 & y_3 & 0 & 0 & z_3 & 0 & 0 & 1 \\ x_4 & 0 & 0 & y_4 & 0 & 0 & z_4 & 0 & 0 & 1 & 0 & 0 \\ 0 & x_4 & 0 & 0 & y_4 & 0 & 0 & z_4 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_4 & 0 & 0 & y_4 & 0 & 0 & z_4 & 0 & 0 & 1 \end{bmatrix} \tag{4.18}$$

$$B = \begin{bmatrix} nx_1 & ny_1 & nz_1 & nx_2 & ny_2 & nz_2 & nx_3 & ny_3 & nz_3 & nx_4 & ny_4 & nz_4 \end{bmatrix}^T \tag{4.19}$$

$$C = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{21} & m_{22} & m_{23} & m_{31} & m_{32} & m_{33} & t_x & t_y & t_z \end{bmatrix}^T \tag{4.20}$$

$$C = A^{-1}B \tag{4.21}$$

The form of matrix C has calculated the transformation between the two sets of calibration points as a directional cosine matrix and a translation. However, directional cosine matrices are not unique, and many differing direction cosine matrices can exist for the same 3D rotation of points; this problem is referred to as the gimble problem and is not present in the 2D case. To solve this problem the orientation of a point set in 3D can be described as a set of absolute angles defining a rotation about an axis. The directional cosine rotation matrix is converted to a quaternion matrix which describes the rotation as a unit, 4 dimensional vector, from this the angle and axis of rotation can be easily extracted. The quaternion is then used to find the axis and angle of rotation (Horn, 1987). This approach gives two possible outcomes for the axis and the angle. In the approach taken here the solution is made unique by rectifying the axis and angle so as to give only one possible rotation and rotation axis. Subsequent to the start of this work a similar 3D Hough Transform has been proposed which uses the same approach for describing the resulting directional cosine rotation matrix as a combination of angle and axis of rotation (Tombari and Di Stefano, 2010).

$$M = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix} \tag{4.22}$$

$$\begin{bmatrix} q_1 & q_2 & q_3 & q_4 \end{bmatrix} = \text{Quart}\left(M\right) \tag{4.23}$$

$$angle = 2 \arccos\left(q_1\right) \tag{4.24}$$

$$axis = \frac{\begin{bmatrix} q_2 & q_3 & q_4 \end{bmatrix}}{\sin\left(\frac{angle}{2}\right)} \tag{4.25}$$

The result of the 3D Hough Transform is a 7 dimensional parameter vector for each transformation between observations. These vectors could be then be histogrammed in a quantised parameter space. However, in this experimental configuration the result of the 3D Hough Transformation is stored as an individual point in a continuous space. The resulting rotation applied by the turntable can be found from the mean of a comparison of a number of observations of calibration targets (Aragon-Camarasa and Siebert, 2010). The translation component of the 3D Hough Transform vector, see $\{t_x, t_y, t_z\}$ in Equation 4.20, in the turntable calibration defines the motion between individual observations of the calibration targets; this is however not the centre of rotation of the turntable. The next section details the approach for calculating the centre of rotation of the turntable.

### 4.3.3 Finding Centre of Rotation

The 3D Hough Transform returns the rotation and translation between two sets of 3D calibration points, however the calculated translation does not define the centre of rotation of the turntable. The line about which the turntable rotates can be found by considering the intersection of two planes which lie radial to the turntable rotation. These planes can be defined by their normal, which is tangential to the rotation of the turntable, and a point on the plane, which can be computed as the mean of a calibration point location used to form the tangent.

Figure 4.10 shows a 2D representation of the turntable centre finder. $Vec1$ and $Vec2$ are tangents to the circle formed by taking two observations of the same point on a calibration target for different positions of the turntable. The intersection of the vectors perpendicular to $Vec1$ and $Vec2$ which pass through the mid point of the vectors is the centre of the turntable.

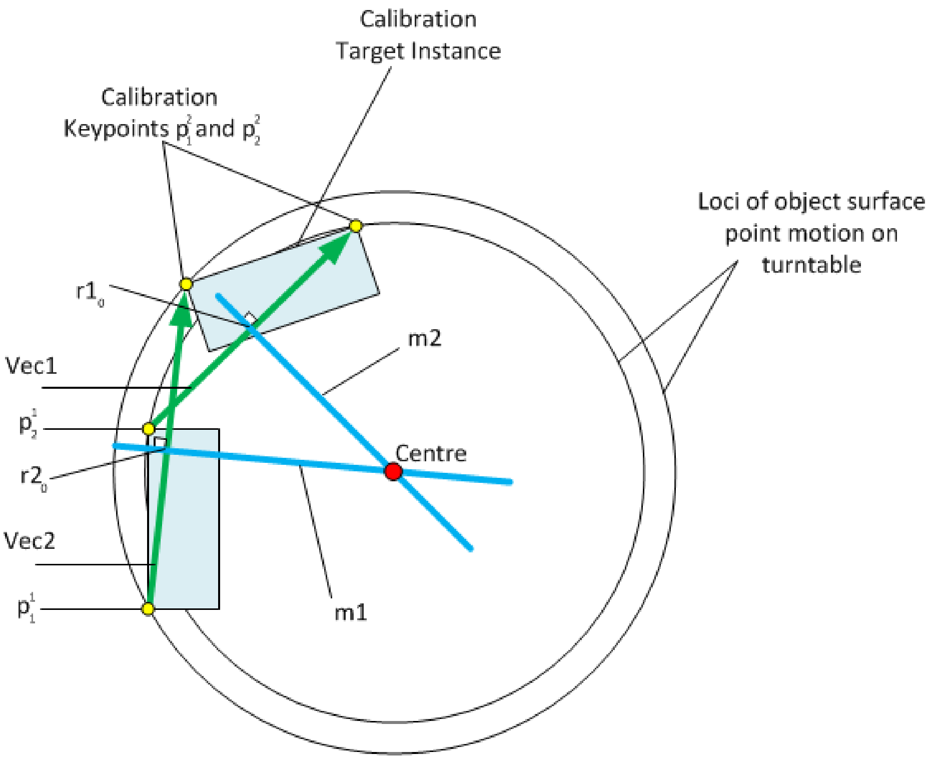

Calibration Target Instance
Calibration Keypoints $p_1^2$ and $p_2^2$
Loci of object surface point motion on turntable
$r1_0$
m2
Vec1
$p_2^1$
$r2_0$
Centre
Vec2
$p_1^1$
m1

Figure 4.10: Centre Finder

$$target1 = [p_1^1, p_2^1] \tag{4.26}$$

$$target2 = \left[p_1^2, p_2^2\right] \tag{4.27}$$

$$[Vec1, Vec2] = [target2 - target1] \tag{4.28}$$

The centre of rotation is calculated from the intersection of planes whose normal is tangential to the rotation of the turntable. $r1_0$ and $r2_0$ lie on each of these planes and have normals of $Vec1$ and $Vec2$ respectively, Equations 4.31 and 4.32. The equation of the plane can then be rearranged, Equations 4.35 and 4.36.

$$r1_0 = \frac{Vec1}{2} + p_1^1 \tag{4.29}$$

$$r2_0 = \frac{Vec2}{2} + p_2^1 \tag{4.30}$$

$$plane1 : Vec1.(r1 - r1_0) = 0 \tag{4.31}$$

$$plane2 : Vec2.(r2 - r2_0) = 0 \tag{4.32}$$

$$Vec1_x x + Vec1_y y + Vec1_z z = Vec1.r1_0^T \tag{4.33}$$

$$Vec2_x x + Vec2_y y + Vec2_z z = Vec2.r2_0^T \tag{4.34}$$

$$m1 = Vec1.r1_0^T \tag{4.35}$$

$$m2 = Vec2.r2_0^T \tag{4.36}$$

The rearranged plane equations can then be solved for the intersection. Equation 4.38 where the point $c$ gives a point on the line which defines the axis of the centre of rotation of the turntable, Equation 4.37.

$$c = \begin{bmatrix} X \\ Y \\ 0 \end{bmatrix} \tag{4.37}$$

$$\begin{bmatrix} Vec1_x & Vec1_y \\ Vec2_x & Vec2_y \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} m1 \\ m2 \end{bmatrix} \tag{4.38}$$

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} Vec1_x & Vec1_y \\ Vec2_x & Vec2_y \end{bmatrix}^{-1} \begin{bmatrix} m1 \\ m2 \end{bmatrix} \tag{4.39}$$

To increase the confidence in the point chosen as the centre of rotation of the turntable, a least squares fit for the intersection of many planes is calculated, Equation 4.42. A combination of all pairs of calibration targets imaged at 5° intervals is used for additional confidence.

$$A = \begin{bmatrix} Vec1_x & Vec1_y \\ Vec2_x & Vec2_y \\ \vdots & \vdots \\ Vecn_x & Vecn_y \end{bmatrix} \tag{4.40}$$

$$b = \begin{bmatrix} m1 \\ m2 \\ \vdots \\ mn \end{bmatrix} \tag{4.41}$$

$$(A^T A)^{-1} A^T b = c \tag{4.42}$$

### 4.3.4 Calibration Summary

This section has described the calibration routine used in this experimental configuration to establish the ground truth for motion of the object surface between imaging observations. The result of the calibration routine is a deterministic linear transformation for the alignment of every point on a range image surface between observations. The steps taken to achieve this have been to first locate calibration keypoints in 3D, then use the location of these to calculate the turntable motion, these steps are shown in Algorithm 1.

**Algorithm 1** Calibration Routine

```
1. Locate calibration keypoints in 3D
    (a) Locate calibration points in texture images
    (b) Build range images
    (c) Constrain calibration range
    (d) Find x-y image pixel scaling
2. Calculate turntable transformation
    (a) Calculate turntable rotation
    (b) Calculate centroid of rotation
```

#### 4.3.4.1 Accuracy

Three factors affecting the accuracy of the experimental configuration to align pixels on the range image surface between observations have been identified. These potential sources of error are listed here and discussed in this section, these are:

- Control inaccuracies in the position of the turntable.
- Inaccuracies in turntable calibration resulting from the calibration routine.
- Errors in the capture of the range surface.

An error in the position of the turntable can be caused by mechanical friction, or the build quality of the turntable actuating equipment. To mitigate for this potential source of error, the turntable employs an optical feedback encoder in a control loop to allow the actuating angle to be refined to match the reference angle. To allow for the turntable to settle into a steady state, 1 minute is allowed between turntable actuation and image capture.

The error in turntable calibration routine can be characterised by the comparison of points found between calibration target observations. To characterise this the calibration points from multiple observations are transformed to the canonical position. The standard deviation of the difference in position between calibration points was measured at less than a millimetre. For use in the experimental configuration the standard deviation of the error in calibration point locations in the canonical space must be sufficiently smaller than the catchment region defined in Section 4.5. As catchment regions are of many pixels each covering more than a millimetre, the turntable calibration error value of less than a millimetre is therefore an acceptable error.

The error in range image calculation may result from imaging errors or errors in camera calibration. In this experimental evaluation C3D returned a calibration error of less than 0.5 pixels. The resultant noise in the range images is discussed in Section 4.4.

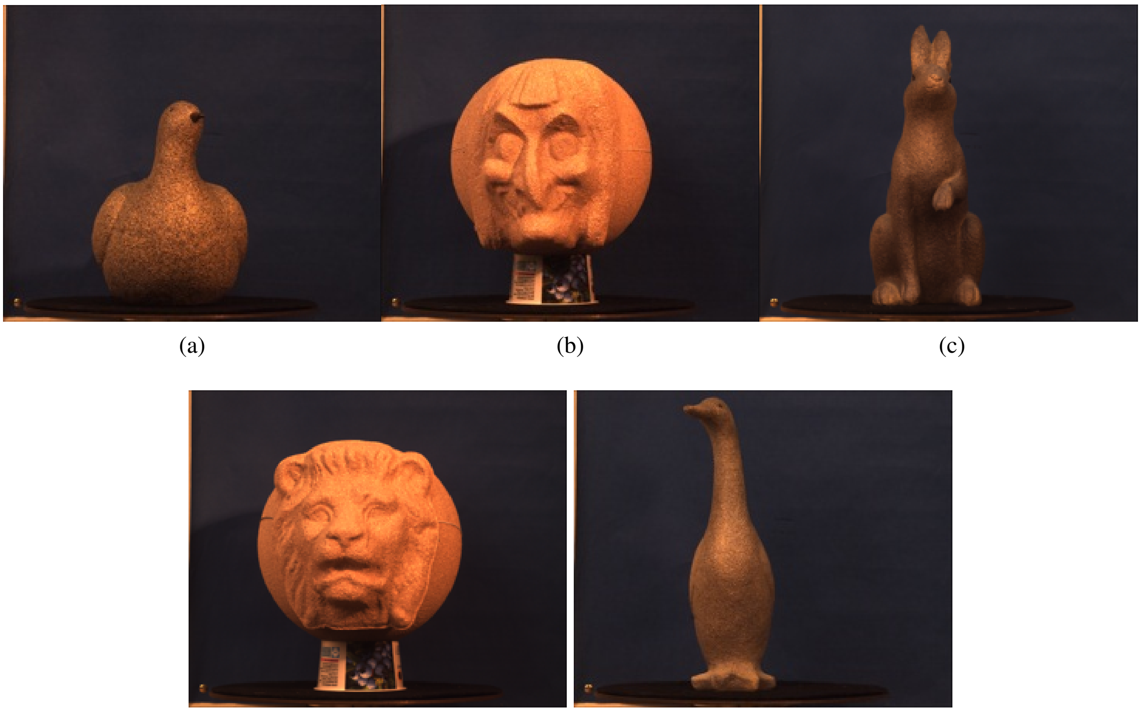
## 4.4 Capture setup

Stereo pair images of each object observation were captured using a camera configuration located 2.2m from the object. The stereo pair camera configuration comprised of cameras with 50mm lenses, and baseline separation of 35cm. This camera configuration is sufficient to assume the weak perspective for both intensity and range images captured within the operating region of the cameras. From the stereo pair intensity images a range image for each object is built using the software package C3D (Ju et al., 2003). The range image is formed with respect to the left camera intensity image. Stereo pair image were captured at 5° intervals to cover the full 360° view in the yaw axis of rotation of the object. The position of the object is controlled by the calibrated turntable, introduced in the previous section. Object observations were imaged using 5 Mega pixel resolution intensity cameras. To reduce image noise 10 intensity images were captured for each object position and from these an average image formed. The image noise in the average image, $Im_{noise}$, is assumed to be Normally distributed. This can be estimated by comparing the pixel variance in each individual image, $Im_{pix}$, to the variation in the average image. Using central limit theory the resultant image noise in the average image, $Im_{noise}$, is estimated as the average pixel variance in a single image, $var(Im_{pix})$, divided by the square root of the number of images averaged, $\sqrt{N_{im}}$ , Equation 4.43. This is calculated to be 0.4680%.

$$Im_{noise} = \frac{var\left(Im_{pix}\right)}{\sqrt{N_{im}}} \tag{4.43}$$

### 4.4.1 Range Image Dataset Collection

The performance of the evaluation configuration outlined in this chapter relies on object observations with high quality range images. This requirement can be met by selecting objects whose surface texture aid the cross correlation matching approach used in C3D to formulate the range images (Siebert and Marshall, 2000; Ju et al., 2003; Urquart, 1997). Surfaces which typically form



(a) (b) (c)

(d) (e)

Figure 4.11: Models Captured

high confidence matches between stereo pair images are highly textured surfaces with no specular reflection (Gupta et al., 2010b; Cyganek and Siebert, 2011). As such the selection of objects used were 5 free-form rigid body objects with speckle-like surface texture, Figure 4.11. The selected objects also have the additional advantage that the speckle-like surface texture forms unique intensity local patterns for locations on the objects which the keypoint feature descriptors investigated in this dissertation should be optimised to characterise.

### 4.4.2 Colour Segmentation

The turntable position transformation between observations is only applied to the portion of the image which is actuated by the turntable. However, intensity and range image captures of an object also include a background which is not actuated by the turntable. Local features may be detected on this background and form correct matches between observations, however the location ground truth will reject these as a turntable position change is assumed. To mitigate for this effect colour separation is applied to remove any keypoints detected on the background.

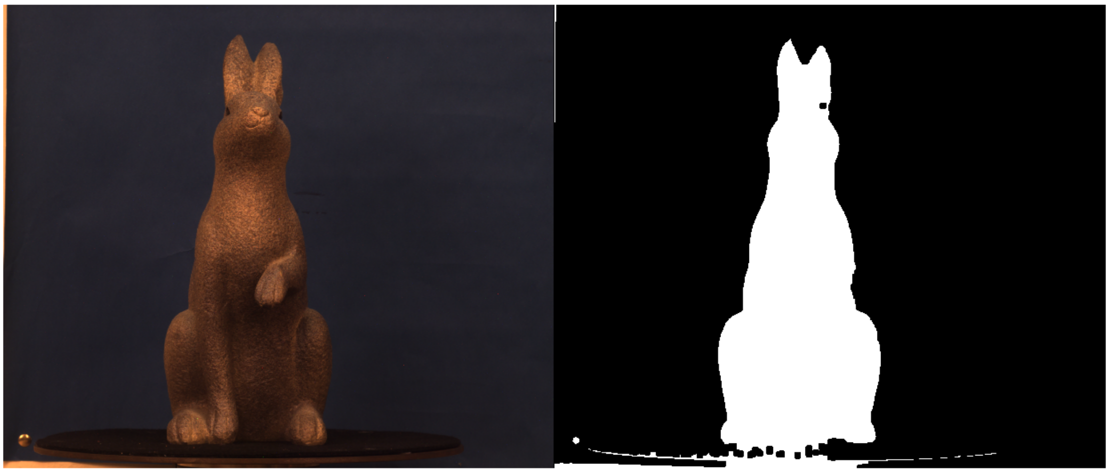


Figure 4.12: Colour Separation Mask

In this experimental configuration a blue background is used to create a mask by colour separation. The resultant colour separation mask is then smoothed by applying a dilation of 2 pixels and then eroded by 2 pixels. Keypoints detected in this invalid region defined by the mask may then be removed from the extracted keypoints. Figure 4.12 shows the left colour and intensity image, and the resulting colour separation mask.

## 4.5 3D Benchmarking Tool

The 3D evaluation methodology presented in this section adopts a similar approach as used in Section 3.2. In this section intensity and range images are collected using the same calibrated cameras and turntable configuration described in Sections 4.3 and 4.4. The experimental methodology presented here allows the effects of 3D view points changes on feature descriptor discriminability to be investigated while maintaining an exact correspondence of the range surface between observation, allowing for a ground truth of keypoint location matches to be established. In addition to the effects of out of plane pose changes, illumination changes are also present as the object surface changes its pose with respect to the light sources. These variations between successive images of the object, in conjunction with the available range data will allow for an evaluation of proposed pose invariant local features, using both intensity and range data. The remainder of this section details the experimental setup and implementation issues. Algorithm 2 shows the order of operation of the

**Algorithm 2** Benchmarking tool

```
Sort object database: find target images, every n degrees
Collect Target Keypoints
For target images

    Read range, texture and colour images
    Extract features: SIFT
    Colour separate valid keypoints
    Transform position (and pose) to canonical space
    Add to keypoints to database, store model id

Collect Query Keypoints
For Queries = 1 to 20

    Select random model
    Read range, texture and colour images
    Extract features: SIFT
    Colour separate valid keypoints
    Transform position (and pose) to canonical space
    Match Nearest Query and Target Keypoints
    For sensitivity = 0 to 1
        Nearest Neighbours
        ROC Statistics
        ROCData(Queries, sensitivity) = ROC point

For sensitivity 0 to 1

    ROC point mean = mean(ROCData(:, sensitivity))
    ROC point std = standardDeviation(ROCData(:, sensitivity))

Store ROC points
```

experimental design, key sections are shown in bold and expanded on in subsequent subsections.

### 4.5.1 Keypoint Comparison

Keypoints localised and described by a SIFT variation are compared in two stages in this experimental setup. The known location of corresponding keypoints between target and query images is established using a catchment region in location and scale surrounding the query keypoint in the 3D canonical space. All target keypoints which are within this catchment region and from the same model are associated together and stored in the position match tuple `keyMatch`, Algorithm 3.

The keypoint feature descriptor are matched together using the nearest neighbour log-likelihood distance measure between a query keypoint and all target keypoints. The log-likelihood distance measure estimates the distribution of incorrect matches by a ratio of the nearest match to the second

**Algorithm 3** Match Nearest Query and Target Keypoints

```
    For keyNum = 1:length(queryKeypoints)

        posDist = sqrt((targetPos - repeat(queryPos(keyNum)))^2)
        scaleDist = abs(log2(targetScale) - log2(queryScale(keyNum)))
        posMatch = posDist < posThresh * queryScale(keyNum)
        scaleMatch = scaleDist < scaleThresh
        modelMatch = targetId == queryId
        keyMatch{keyNum} =  posMatch & scaleMatch & modelMatch
    return keyMatch
```

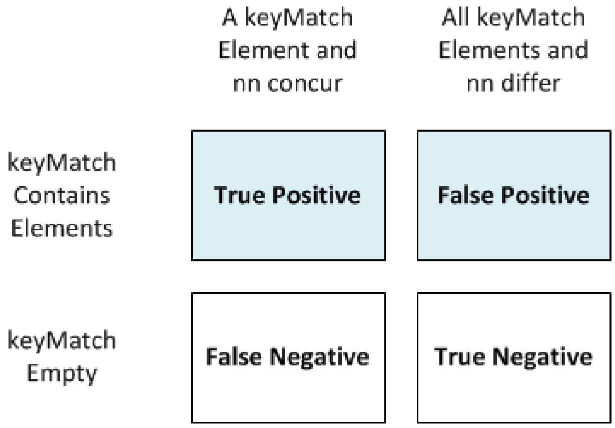
**Algorithm 4** Nearest Neighbours

```
    For keyNum = 1:length(queryKeypoints)
        featDist = sqrt((targetFeat - queryFeat)^2)
        firstNn = min(featDist)
        firstNnId = targetId(firstNn)
        featDist(~(targetId == firstNnId)) = 1
        featDist(firstNn) = 1
        secondNn = min(featDist)
        if firstNn/secondNn < sensitivity
            nn(keyNum) = firstNn
        else
            nn(keyNum) = 0
    return nn
```

nearest match. In the target database there may exist multiple instances of the same keypoint. To ensure that any duplicate keypoints are not considered for the second nearest neighbour all feature descriptors from any target model other than the nearest match are set to the maximum distance value of 1, the second nearest neighbour is then taken from the same model, Algorithm 4.

### 4.5.2 Keypoint ROC Analysis

The ROC curves used to compare the performance of local matching features are determined with reference to the query keypoints. For every query keypoint the neighbouring target keypoints in all target instances of the same object are found and associated in Algorithm 3. To generate the ROC curves the benchmark applies the rule shown in Figure 4.13 to every query keypoints. If any query target keypoint pair match in position and feature descriptor space, the keypoint is labelled as true positive, otherwise if a position match exist which is not verified by the feature descriptor matching



| | A keyMatch Element and nn concur | All keyMatch Elements and nn differ |
|---|---|---|
| keyMatch Contains Elements | **True Positive** | **False Positive** |
| keyMatch Empty | **False Negative** | **True Negative** |

Figure 4.13: Confusion Matrix for Evaluation ROC Curve

stage a false positive label is applied, Algorithm 5.

## 4.6 Summary and Discussions

An evaluation approach for the comparison of keypoint correspondences between object views has been introduced. The evaluation has been designed to form a deterministic transformation for the motion of keypoint locations between viewing instances, allowing the comparison of all keypoints localised on the object surface. In order to calibrate the turntable, an efficient approach for the analysis of the composition of distinctive locations on a calibration target has been introduced, the 3D Hough Transform.

The use of a stereo capture and calibrated turntable configuration has allowed for co-aligned range and intensity images to be captured at accurately controlled view poses. The use of co-aligned range and intensity image data allows not only for the 3D alignment of locations on the object surface, but also for an evaluation of the performance of keypoints from each modality independently, or the performance of keypoints from the combination of modalities to be investigated. The objects collected have a highly textured surface to aid stereo matching, giving a high confidence in the stereo match. It is anticipated that this highly textured surface will create a distinctive and unique image sample patch for each keypoint location, allowing for the matching of feature descriptors under the best conditions to be evaluated.

The following chapter aims to investigate the possible cross modal combinations of local fea-

**Algorithm 5** ROC Statistics

```
true = 0
false = 0
positive = 0
negative = 0
For keyNum = 1:length(nn)
    if length(keyMatch{keyNum}≠0 & nn(keyNum)≠0
        match = sum(keyMatch{keyNum} == nn(keyNum))
        true = true + match
        false = false + ~match
        positive ++
    elseif length(keyMatch{keyNum}==0 & nn(keyNum)≠0
        false ++
        negative ++
    elseif length(keyMatch{keyNum}≠0 & nn(keyNum)==0
        positive ++
    elseif length(keyMatch{keyNum}==0 & nn(keyNum)==0
        negative ++
truePositive = true/positive
falsePositive = false/positive
return truePositive falsePositive
```

tures extracted from the intensity and range images and evaluates their performance using the experimental setup outlined in this chapter.

# Chapter 5

# Single Observation Pose Invariant Local Features

This chapter follows on from Chapter 3, where variations on local feature matching algorithms in 2D were investigated, to investigate the potential advantages offered by range domain or multimodal combinations of information from range and intensity domain images. In order to investigate the effects of local feature matching in these modalities a calibrated capture configuration, outlined in Chapter 4, was used to collect a database of real free form objects imaged at differing out-of-plane viewpoints. This chapter utilises the co-aligned range and intensity images to investigate modifications which can be applied to keypoints from single observations to improve their matching performance.

## 5.1 Objectives

This chapter introduces and evaluates a variety of invariant local feature matching algorithms based on the SIFT MATLAB code used in previous chapters. The main focus of the chapter is on single observations of local features and investigates what modifications can be made to the SIFT processing pipe-line in order to best utilise information from range and intensity domain images. Therefore, three stages in the SIFT processing pipe-line where the application of information from the range domain may improve performance:

- **Keypoint localisation**, The range domain represents the underlying surface structure of the object imaged and may offer features which are more consistently localised through view point changes. Furthermore, co-aligned range and intensity images offer the potential to investigate the performance of a combination of features localised in both range and intensity domains, see Section 5.2.

- **Sample patch arrangement**, In Chapter 3 static receptive field arrangements were investigated. This chapter extends this investigation using range domain information. Range domain information offers an invariant measure of surface structure at a keypoint location. This can be used to dynamically change the local keypoint sample patch between keypoint observations to reflect the change in underlying surface structure resulting from a change in 3D view point, see Section 5.3.

- **Surface description**, Keypoint feature descriptors rely on an invariant surface measure calculated from a surface patch surrounding a keypoint location. Co-aligned range and intensity patches offer a greater range of possible local surface measures, in addition to offering cross modal combinations of descriptors resulting from measures of each domain, see Section 5.4.

The ROC curves from the modified SIFT implementations are compared against the ROC performance curves for the unmodified SIFT MATLAB code. The unmodified SIFT MATLAB code serves as a standard bench mark isolating the effects of each modification, Section 5.5. Section 5.7 concludes this chapter with an overview of the effects of each of the proposed local feature matching algorithms and the influence on performance of each change.

### 5.1.1 Overview

This chapter investigates the existing SIFT pipe-line and changes which can be made to accommodate the characteristics of the range imaging modality. The chapter addresses three issues: the principal actions of the SIFT processing pipe-line, how the action of these are affected in the range image and potential changes which can be made to stages in the processing pipe-line to accommodate information from the range imaging modality. The SIFT processing pipeline is shown in Figure 5.1. The principal actions of the local feature extraction are:

1. **Keypoint Localisation**

   (a) **Scale space pyramid** - Allows for features to be located at differing scales. However, scale space image pyramids cannot be applied directly to range images. Therefore, this chapter introduces a range scale space pyramid which accounts for the characteristics of the range imaging modality.

   (b) **Interest point detector** - The approach used to locate keypoints can be applied to the range imaging modality opening the potential for cross modal feature localisation and multimodal feature description, from co-registered range and intensity images.

2. **Local Patch Invariance Estimation**

   (a) **Canonical orientation assignment** - In-plane orientation changes are addressed in the SIFT approach outlined by Lowe (Lowe, 2004, 1999), these are extended by Mikolakczyk and Schmid to create a measure of affine pose of features (Mikolajczyk and Schmid, 2005). However, the range imaging modality offers further information regarding the 3D pose of a local sample patch which should remain invariant under full 3D out-of-plane view point changes.

   (b) **Receptive field correction** - Before extracting a measure of the local surface surrounding a keypoint location, a geometric transformation to the local surface sampling patch may be applied. The geometric transformation applied to the sample patch allows the area covered by a sample patch from the same keypoint in differing observations to sample the same underlying object surface. In standard SIFT this sample patch correction addresses only the in-plane orientation changes, however additional 3D pose information can be used to apply a warping to a given image sample patch to fit it to the underlying 3D object surface.

3. **Feature Extraction**

   (a) **Receptive field arrangement** - The receptive field structure were investigated in Chapter 3, the conclusion of this study was that a rectilinearly arranged 16x16 pixel patch with Gaussian receptive fields achieved the highest performance. Therefore, this receptive field arrangement has been selected and is used throughout this chapter.

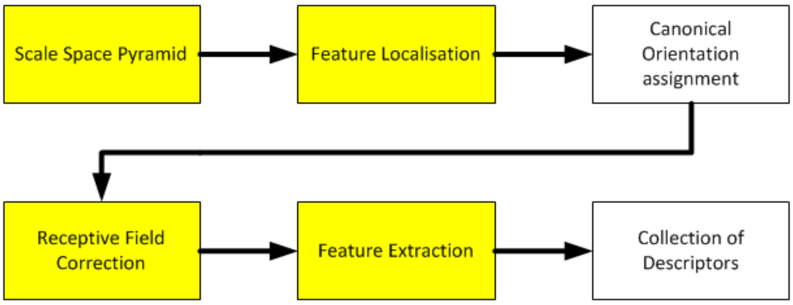

Scale Space Pyramid
Feature Localisation
Canonical Orientation assignment
Receptive Field Correction
Feature Extraction
Collection of Descriptors

Figure 5.1: 2.5D SIFT Descriptors

(b) **Surface measurement** - The range imaging modality has characteristics that differ from the intensity modality, where the approach of using surface gradients has been applied. In this chapter surface gradients and a number of other surface measures are proposed and investigated.

## 5.2 Multimodal Keypoints

Keypoints may be localised in either range or intensity image modalities with their corresponding image patch feature descriptors extracted independently from the other imaging domain. This cross modal keypoint localisation and local feature description approach offers the potential to gain a high number of keypoint locations seeded in the texture rich 2D intensity domain, and form a corresponding descriptor in the illumination and pose invariant range domain. Conversely, a greater performance benefit may be achieved using the lower resolution range image to form keypoint locations which are more localisable and repeatable, which, with feature descriptors extracted from the 2D intensity domain form a well localised and distinct feature descriptor for an image location.

This section aims to outline the preprocessing stages which must be applied to the range images, the characteristics of keypoints localised in each domain and the naming conventions for the SIFT types investigated in the results section, section 5.5.

### 5.2.1 Range Scale Space Pyramid

In order to create local features in the 2.5D range domain, the range image must be described with equal pixel divisions in the X, Y and Z axes; this relationship between surface structures at every scale should remain the same. In order to achieve this every pixel in the native scale range image measures the distance to the object in terms of number of pixels, this gives an equal divisions in X, Y and Z axes for the native scale range image. To maintain these equal axes throughout the scale space pyramid the range image pixel values must be scaled accordingly. This section describes the scaling process.

Range image scale space pyramids are calculated differently from intensity image scale space pyramids. Akaunduz and Ulusoy detail a scale space for range images (Akagunduz and Ulusoy, 2007; Abate et al., 2007). In this section a practical implementation of the range scale space pyramid approach used in this dissertation is outlined. The resulting range image scale space pyramid has the desirable property that the 3D geometry of the imaged surface remains invariant through levels in scale space.

The initial range image pixel values, $Im_{range}$, are given in meters. To correctly process the range images as 3D point clouds, where the X, Y and Z axes are equal, the range image pixel measurements, $Im_{range}$, must be converted from metres to pixels, $Im\hat{_{range}}$. In order to achieve this conversion, a scalar value, $div_{pixel}$, which represents the size of each pixel in metres is calculated. The value of $div_{pixel}$, is calculated by projecting the image into the active region and measuring the size of a single pixel. The projection of the image is found from the distance from the camera perspective centre to the active region $\bar{Z}$, and the optical viewing angle in the x direction of the left camera $\theta_x$, Figure 5.2. The size of each pixel is taken as the proportion of the whole image which a pixel occupies; $n_x$ is the number of pixels in the x image direction, see Figure 5.2. The size of x and y pixels in meters were found in the C3D camera calibration routine to be equal, therefore the pixel size in the y direction is assumed to be the the same as the pixel size in the x direction.

$$\bar{X} = 2\bar{Z}.\tan(\theta_x) \tag{5.1}$$

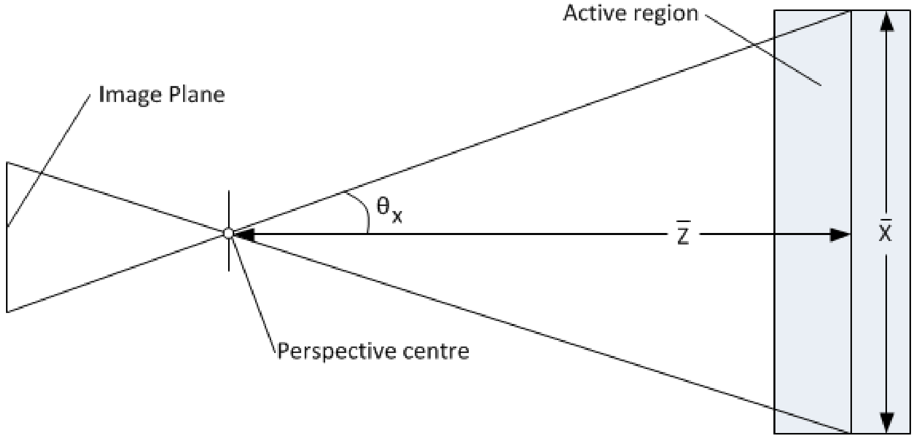


Figure 5.2: Projection of the image into the active region of the cameras. Measurements in this space allow the range images, recorded in metres, to be converted to equivelent pixels sizes.

$$div_{pixel} = \frac{\bar{X}}{n_x} \tag{5.2}$$

$$\hat{Im_{range}} = \frac{Im_{range}}{div_{pixel}} \tag{5.3}$$

In the stereo capture configuration used in this dissertation there is a small operational region of the object viewing space where both cameras are focused, in this region the camera model for the left camera, and subsequent range, image assume weak perspective. The difference in range at the edges of the range image between this configuration and a true orthographic camera model was calculated to be 0.3mm.

To create a range image scale space pyramid, equal divisions in image pixel values and range values are required for each level in the scale space pyramid. Equal divisions in X, Y and Z at every level in scale space maintains the surface shape throughout scale space. In order to achieve this, the range values for each new level in the scale space pyramid are divided by the subsampling factor used to down sample between scale levels in the pyramid. The subsequent levels in the scale space pyramid are formed by Gaussian filtering the range image with a 2D Gaussian, $\sigma$ equal to the subsampling ratio, decimating the range image by the subsampling ratio and dividing the new range image range values by the subsampling ratio.

$$\hat{Im_{range,pry}} = \sigma . G_{\sigma} * \hat{Im_{range,pyr-1}} \tag{5.4}$$

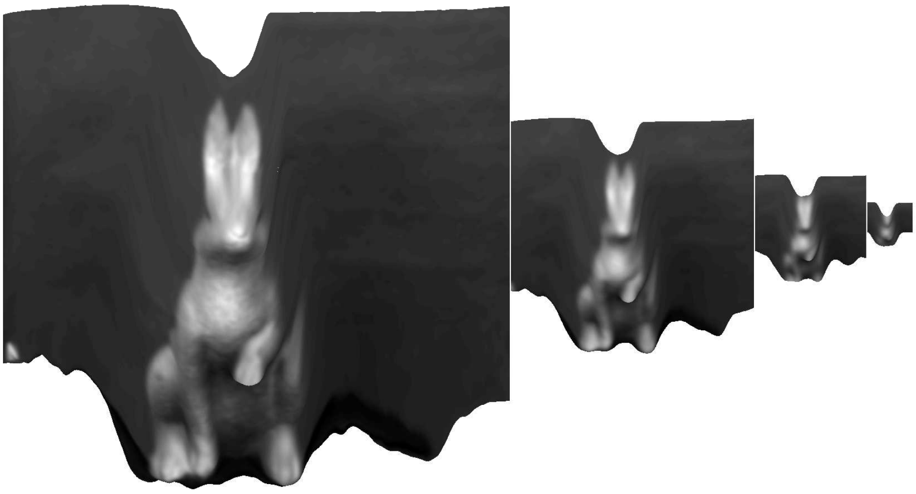


Figure 5.3: Range Scale Space Pyramid

A scale space image pyramid with a subsampling factor equivalent to $\sqrt{2}$ pixels is used, this maintains a smooth variation in scale space without discarding potential keypoints (Cyganek and Siebert, 2011). The range image scale space pyramid covers 4 octaves in scale each with 2 intervals.

### 5.2.2 Characteristics of Keypoints

The regions surrounding keypoints localised in intensity and range domain images have differing characteristics. To highlight these differences, a typical example of a range and intensity pair of images is described here. Figure 5.4 shows a comparison of keypoints localised in each domain, sub-figure c shows the keypoints from both domains on the one image. By inspection, it can be seen that typically range and intensity domain keypoints are localised at differing image locations with only a small number of keypoints locations overlapping between the domains. The intensity domain typically produces twice the number of keypoints locations as the range domain, Table 5.1. Figure 5.5, shows a comparison of the locations of the keypoints in each of the image domains with regions of semantic interest from an instance of facial landmarking Ferrario et al. (1998). Keypoints near semantic points of interest are highlighted in red. This comparison shows that in this example,

| | Number of Keypoints |
|---|---|
| Texture | 275 |
| Range | 508 |
| Multimodal | 783 |

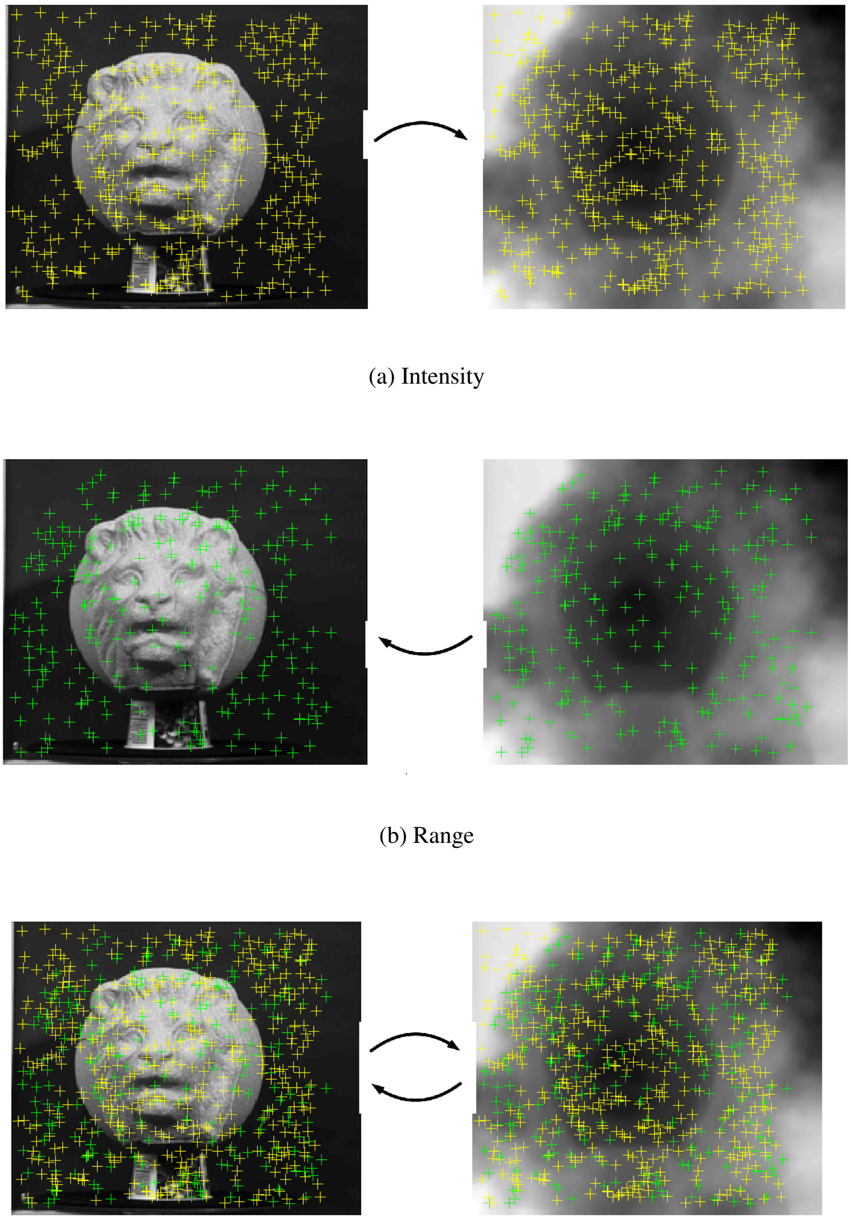
Table 5.1: Number of keypoints per modality

keypoint localisation in the range image detected more keypoints close to repeatable landmarks selected by human clinicians, when compared to keypoints localised in the intensity domain. This result indicates that the quality of the keypoints is expected to be higher for the range imaging domain. Furthermore the semantically important keypoint, localised in the range domain with the Harris Corner detector approach, show a higher degree of symmetry, where keypoints detected on the left side of the object are also detected on the right, than for the corresponding intensity keypoint localisation approach.

### 5.2.3 Multimodal Keypoint Combination

To create multimodal keypoints, keypoints extracted from the range and intensity domains need to be combined together in a optimal fashion. There are four possible combinations of modalities for cross modal features in addition to the combination of multimodal features, giving a possibility of nine different combinations of keypoints, see Table 5.2. Cross modal features can be formed through localising in one domain and extracting a surface description from the corresponding image location in the other domain. However, the combination of multimodal localisation and description presents two challenges:

- How can combined keypoints be localised in both domains to create a multimodal localised keypoint?
- How can a multimodal feature descriptor resulting from separate surface measurements in range and intensity domains be created?



(a) Intensity

(b) Range

(c) Multimodal

Figure 5.4: Domain localised features

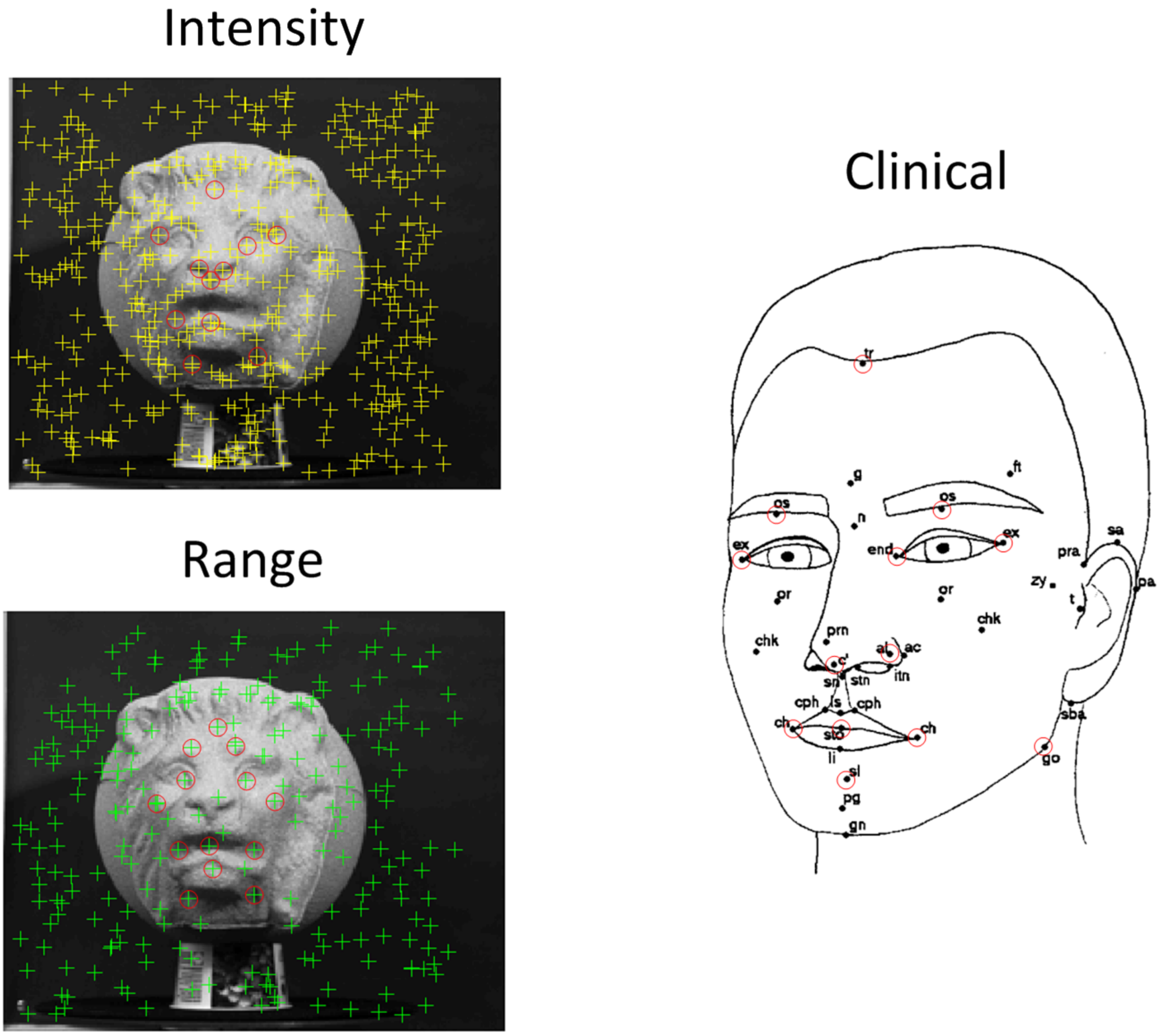


Figure 5.5: Comparison With Semantic Interest Points

| | Texture Localised | Range Localised | Multimodal Localised |
|---|---|---|---|
| Intensity Descriptors | SIFT_2D_2D | SIFT_3D_2D | SIFT_2D3D_2D |
| Range Descriptors | SIFT_2D_3D | SIFT_3D_3D | SIFT_2D3D_3D |
| Multimodal Descriptors | SIFT_2D_2D3D | SIFT_3D_2D3D | SIFT_2D3D_2D3D |

Table 5.2: Cross Modal SIFT

To create a multimodal localised keypoint, a collection of keypoints formed from localising and describing in both the range and intensity domain are created. To form a multimodal localised keypoint from this collection of keypoints, the two collections are concatenated, Equation 5.5.

However, as feature descriptors must must have a vector length of 1, this approach is not appropriate for creating multimodal feature descriptors. In order to create a multimodal feature descriptor, keypoints have feature descriptors extracted from the same image location in each domain. The resulting feature descriptors are then concatenated and renormalised to unit length, Equation 5.7. In this work keypoint feature descriptors extracted in each of the range and intensity modalities have feature descriptor vectors of equal number of elements, therefore equal importance is assigned to each modality during the matching stage. All keypoints have their canonical orientation assigned from the modality from which they have been localised. All combinations of multimodal keypoints can be created following the rules set out in this subsection.

$$Keypoints_{loc2D,loc3D} = \left\{ \begin{array}{c} x_{2D}, y_{2D}, \sigma_{2D}, feat \\ x_{3D}, y_{3D}, \sigma_{2D}, feat \end{array} \right\} \tag{5.5}$$

$$norm_{L2} = \sqrt{\text{sum}\left\{feat_{2D}^2, feat_{3D}^2\right\}} \tag{5.6}$$

$$Keypoints_{loc2Ddesc2D3D} = \left\{ x_{2D}, y_{2D}, \sigma_{2D}, \frac{\{feat_{2D}, feat_{3D}\}}{norm_{L2}} \right\} \tag{5.7}$$

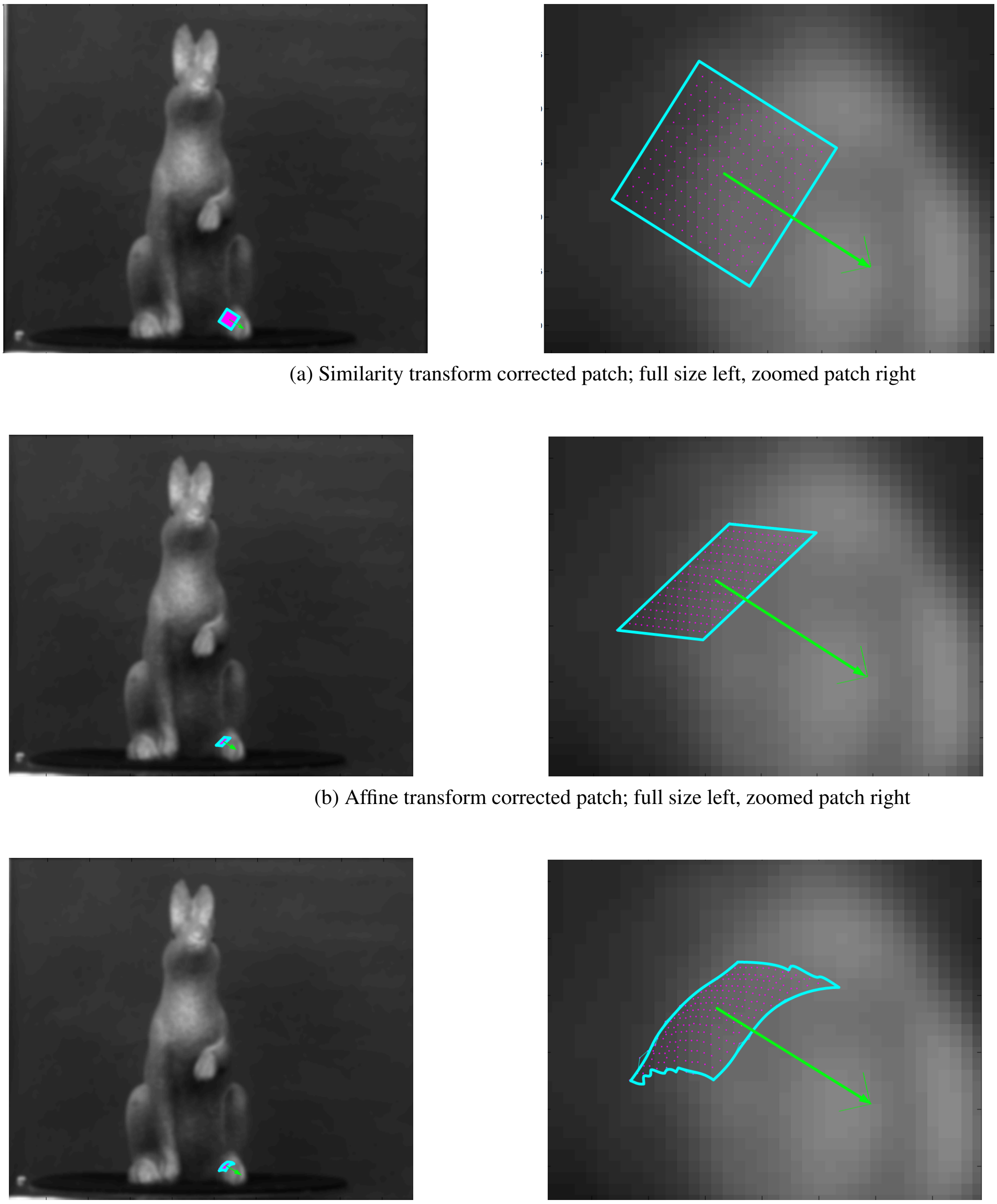
### 5.2.4 Naming conventions

This section has presented a multimodal feature localisation and description approach. The differences between range and intensity images have been presented and the implementation issues addressed. The results for the outlined descriptors are presented in Section 5.5.1, the naming conventions of the descriptors outlined in this section are shown in Table 5.2.

## 5.3 Sample Patch Warping

SIFT offers invariance to in-plane rotational changes by calculating a dominant orientation of 2D intensity gradients in an image sample patch and correcting the orientation of the sample patch to align with this orientation. The orientation estimation and sample patch orientation correction stages ensure that the sample patch will have a normalised orientation which consistently aligns with the dominant orientation irrespective of the camera orientation. However, the invariance offered through this approach is limited to in-plane orientation, or out-of-plane orientation changes of planar structures where the perspective warp introduced by the imaging device is small. To account for a greater degree of freedom in image transformations, the orientation estimation and sample patch correction stages in the local feature matching algorithms processing pipe-line has been extended in order to find and correct for the dominant affine approximation for the local feature sample patch. However, the range of invariance offered by both these approaches is unable to handle the non-linear 3D motion of points due to changes in view point when imaging 3D free form objects. Additionally, the estimation of surface orientation is based on a measure of the intensity image which may not remain invariant through view point changes.

The range image modality offers a consistent surface topology between views which can be used to infer the connectivity of sampled points on the object surface. 2.5Dpc SIFT uses this additional 3D local pose information to correct for changes in image orientation and apply affine transformations to the sample patch to compensate for changes in view point (Lo and Siebert, 2009). By adopting the range image representation, the concept of applying corrective transformations to the sample patch can be extended further to apply a full projective transformation to a keypoint sample patch. In this section the 3 corrective transformations which can be applied to keypoint local sample patch are introduced, these are the similarity transform, the affine transform and the projective transform, shown in Figure 5.6.



(a) Similarity transform corrected patch; full size left, zoomed patch right

(b) Affine transform corrected patch; full size left, zoomed patch right

(c) Projective transform corrected patch; full size left, zoomed patch right

Figure 5.6: Sample patch corrections

### 5.3.1 Similarity

The similarity transform allows for a rotation of the sample patch in the imaging plane to align with a repeatable orientation. In this dissertation, the SIFT method of calculating the in-plane orientation of the sample patch is used. The orientation of the sample patch is found from the weighted mean of the 2D image gradients surrounding the keypoint, Equation 5.8.

$$\theta = \sum wght(x_n, y_n) \mathrm{arctan}\left(\frac{L\left(x_n, y_n+1\right) - L\left(x_n, y_n-1\right)}{L\left(x_n+1, y_n\right) - L\left(x_n-1, y_n\right)}\right) \tag{5.8}$$

The new sample patch is rotated to align with the keypoint orientation $\theta$ and centred on the keypoint location, $k_x$, $k_y$, equation 5.10 and is defined by sampling the intensity image or range image at the points $\hat{s_x}$, $\hat{s_y}$; $s_x$ and $s_y$ is the original sample grid, spanning the range of -16 to +16 pixels in the x and y directions.

$$M_{in} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.9}$$

$$\begin{bmatrix} \hat{s_x} \\ \hat{s_y} \\ 1 \end{bmatrix} = M_{in} \begin{bmatrix} s_x \\ s_y \\ 1 \end{bmatrix} + \begin{bmatrix} k_x \\ k_y \\ 1 \end{bmatrix} \tag{5.10}$$

### 5.3.2 Affine

In an approach similar to that applied by Lo and Siebert (Lo and Siebert, 2009), the affine transformation applied to the sample patch is calculated from the slant and tilt of the surface patch. The estimates for the values of slant, $\phi$, and tilt,$\tau$, for each keypoint are taken from the components of the surface normals in the x, y and z directions, $N_x$, $N_y$ and $N_z$. From these the affine transformation matrix can be calculated and applied to the sample grid, Equation 5.17. This approach differs from Affine Invariant regions where the transformation matrix is the result of a factorisation of 2D intensity image gradients (Mikolajczyk et al., 2005; Mikolajczyk and Schmid, 2004).

$$\phi = \mathrm{arctan}\left(\frac{\sqrt{N_x + N_y}}{N_z}\right) \tag{5.11}$$

$$\tau = \arctan\left(\frac{N_x}{N_y}\right) \tag{5.12}$$

$$M_{in} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 & 0 \\ \sin(\theta) & \cos(\theta) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.13}$$

$$M_{tilt} = \begin{bmatrix} \cos(\tau) & \sin(\tau) & 0 & 0 \\ -\sin(\tau) & \cos(\tau) & 0 & -sin(\phi) \\ 0 & 0 & 1 & 1 - cos(\phi) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.14}$$

$$M_{slant} = \begin{bmatrix} \cos(\tau) & -\sin(\tau) & 0 & 0 \\ \sin(\tau)\cos(\phi) & \cos(\tau)\cos(\phi) & \sin(\phi) & -\sin(\phi) \\ -\sin(\tau)\sin(\phi) & -\cos(\tau)\sin(\phi) & \cos(\phi) & 1 - \cos(\phi) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.15}$$

$$M_{3D}(\phi, \tau, \theta) = M_{tilt} M_{slant} M_{in} \tag{5.16}$$

$$\begin{bmatrix} \hat{s_x} \\ \hat{s_y} \\ \hat{s_z} \\ 1 \end{bmatrix} = M_{3D}(\phi, \tau, \theta) \begin{bmatrix} s_x \\ s_y \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} k_x \\ k_y \\ 1 \\ 1 \end{bmatrix} \tag{5.17}$$

### 5.3.3 Projective

This work expands on the corrective transformations applied in the literature, which have been described in previous subsections, by proposing a transformation which is capable of correcting for the motion of points on a 3D free form surface under full out-of-plane rotation. The projective transformation is a non-linear transformation applied to the sample points, based on the range image surface structure. To account for self occlusion from facets of the object surface, the range image is represented as a point cloud which is rotated and resampled so as each keypoint can viewed perpendicular to the range image surface at the keypoint location. The rotation is applied using a similar approach to the affine transformation, whereby the in-plane orientation is applied first followed by the out-of-plane transformation. Points in the point cloud which are occluded are then removed using the "Fast Hidden Point Removal Algorithm" (Katz et al., 2007). The local keypoint sample patch is created by projecting a rectilinear sample grid on to the resampled range surface, Equation 5.20. The resultant sample point locations are rotated back into the range image co-ordinate frame and form the "projective transformation sample patch". The resulting sample patch can be applied to sample the range or intensity domain.

$$\begin{bmatrix} \hat{X} \\ \hat{Y} \\ \hat{Z} \\ 1 \end{bmatrix} = M_{3D}(\phi, \tau, \theta)^{-1} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{5.18}$$

$$\begin{bmatrix} \hat{k_x} \\ \hat{k_y} \\ \hat{k_z} \\ 1 \end{bmatrix} = M_{3D}(\phi, \tau, \theta)^{-1} \begin{bmatrix} k_x \\ k_y \\ k_z \\ 1 \end{bmatrix} \tag{5.19}$$

$$\begin{bmatrix} \hat{s_x} \\ \hat{s_y} \\ \hat{s_z} \\ 1 \end{bmatrix} = \text{Resample}\left( \text{HPR}\left( \begin{bmatrix} \hat{X} \\ \hat{Y} \\ \hat{Z} \\ 1 \end{bmatrix} \right), \begin{bmatrix} s_x \\ s_y \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \hat{k_x} \\ \hat{k_y} \\ 1 \\ 1 \end{bmatrix} \right) \tag{5.20}$$

### 5.3.4 Validation

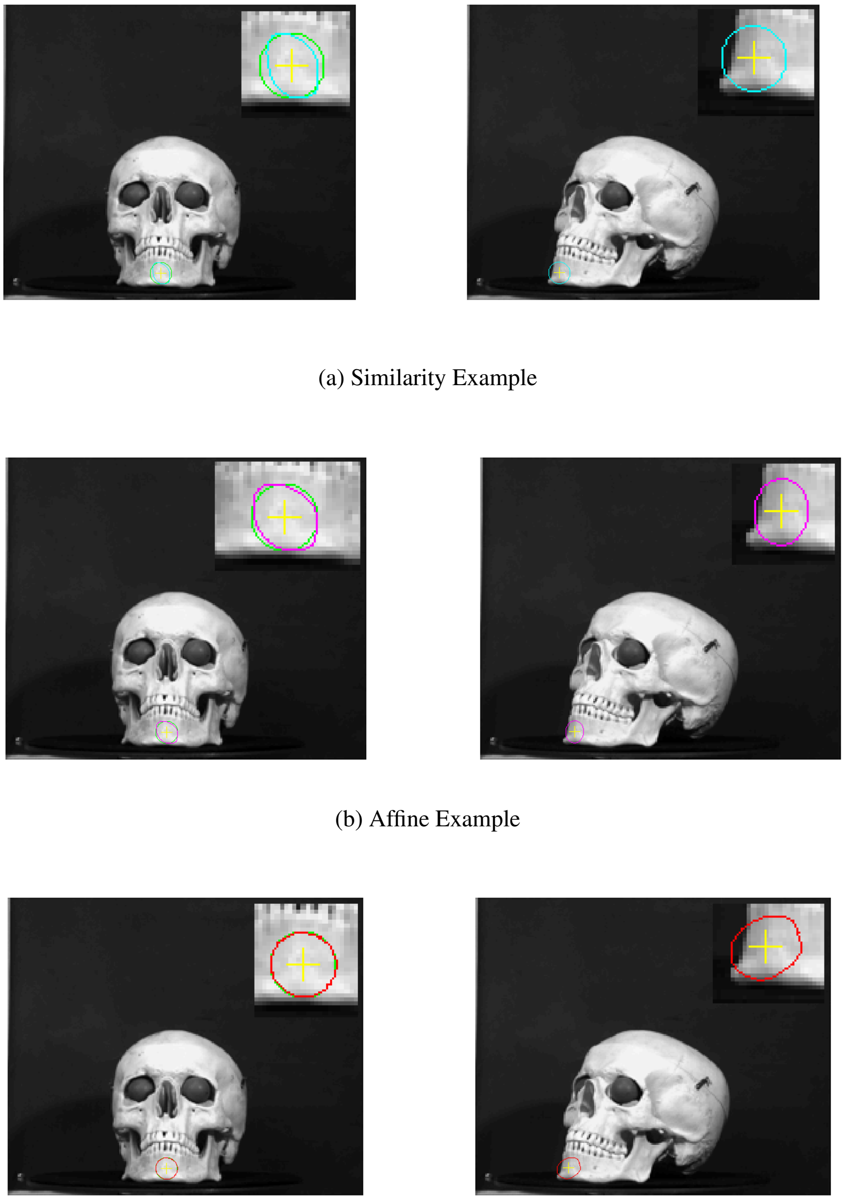
To validate the proposed sample patch warping approaches and investigate the effects of these warping transformations under ideal conditions, the changes in sample patch foot print are shown for each method. Figure 5.7, shows the calculated sample patch foot print for each warping approach in the right image at the keypoint location shown as a yellow cross. The sample patch foot print from the right image is then projected into the left image using the range image for X, Y, Z co-ordinates and the turntable rotation ground truth established in Chapter 4.3. In the left image the keypoint is viewed close to the surface normal, whereas the keypoint observed in the right image is imaged with a surface normal of 30° out-of-plane rotation in the yaw axis. The projected sample patch foot print from the right image can be compared to the circular sample patch footprint in the left image. For all examples in Figure 5.7, the keypoint sample patch footprint in the left image should be close to being circular; the circular footprint is shown in green. From inspection, the projectively corrected example gives the best performance with a consistent sample patch throughout views and a sample patch close to circular in the 0° object instance, see Figure 5.7c.

The difference in the sample patch foot print between object observations was calculated by calculating the average displacement of all points on the footprint circumference, see Figure 5.8. These displacements are plotted against out-of-plane rotation angle, see Figure 5.9. The diameter of the sample patch used was 16 pixels. This validation experiment found that the projective corrected sample patch maintained a performance of similar object surface coverage up to 35° which the other approaches could only achieve up to 10°.

### 5.3.5 Naming conventions

This section has presented an approach for a variety of local image sample patch corrections, from standard in-plane orientation correction to full projective corrected sample patches. The implementation issues have been addressed, showing the exact calculation of the resulting patch for a given local pose estimation. The validation in Subsection 5.3.4, has shown the abilities of each sample patch correction approach to approximate an invariant foot print throughout a range of out of plane orientation changes. The results for the outlined descriptors are presented in Section 5.5.2, the naming conventions of the descriptors outlined in this section are shown in Table 5.6.

The results section investigates using the pose component from the 2D image scene to form the estimation for the transformation applied to the sample patch, column 1 of Table 5.6, in the projective case the transformation uses the range data to estimate the sample patch layout. Column 2 investigates the effect of using the 3D information for pose estimation and sample patch correction only, keypoints are localised and described using the intensity image. Column 3 uses 3D information for all stages of keypoint localisation, pose estimation and description.



(a) Similarity Example

(b) Affine Example

(c) Projective Example

Figure 5.7: Sample Patch Change Between Views

| | Similarity | Affine | Projective |
|---|---|---|---|
| 2D Loc 2D Desc 2D Pose | SIFT_2D_2D | SIFT_aff_2D_2D_2D | SIFT_pro_2D_2D_2D |
| 3D Loc 3D Desc 2D Pose | SIFT_3D_3D | SIFT_aff_3D_2D_3D | SIFT_pro_3D_2D_3D |
| 2D Loc 2D Desc 3D Pose | SIFT_2D_2D | SIFT_aff_2D_3D_2D | SIFT_pro_2D_3D_2D |
| 3D Loc 3D Desc 3D Pose | SIFT_3D_3D | SIFT_aff_3D_3D_3D | SIFT_pro_3D_3D_3D |

Table 5.3: Sample Patch Corrected SIFT

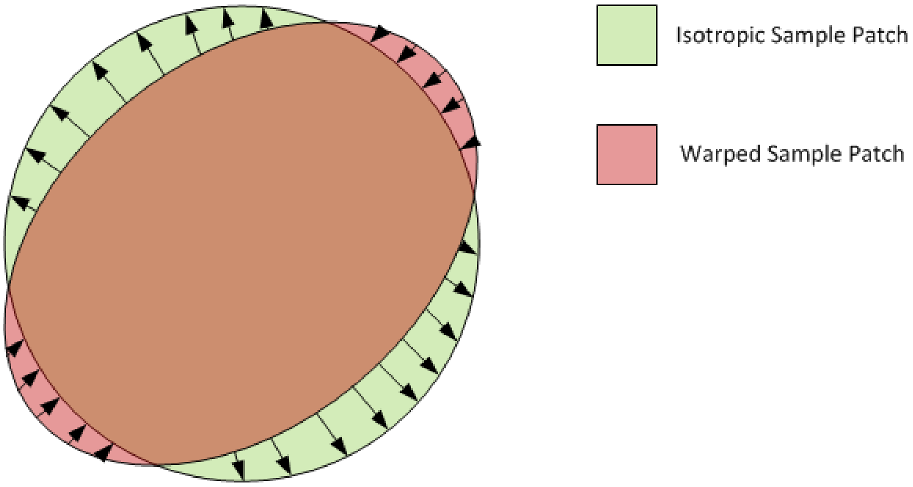


Figure 5.8: Difference Between Sample Patches

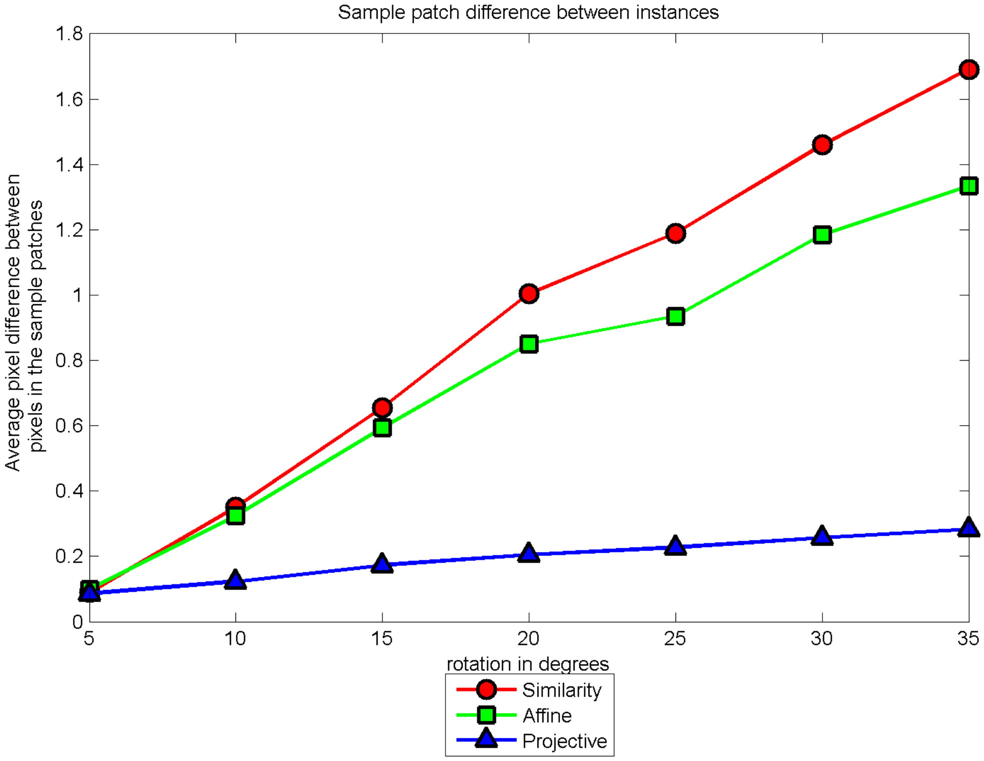


Figure 5.9: Sample Patch Difference Between Instances

## 5.4 Surface Descriptor

To create the feature descriptor vector used to match keypoint instances, a histogram of surface measures encoding a measure of the underlying object surface is formed from a sample patch. For each pixel in the sample patch, a surface measure is computed and used to vote in a weighted histogram which forms the feature descriptor for a receptive field. The feature descriptor for a keypoint is formed from the concatenation of the histograms for each receptive field. The choice of surface representation used to create the receptive field histograms has a large impact on the repeatability of the feature descriptor. The choice of surface measure should therefore be chosen to increase invariance to the transformations which a keypoint can express, while increasing the discrimination between non-corresponding keypoints. In this section a number of surface representations for characterising local patches on 2.5D range images are proposed in order to establish the most effective representation for creating repeatable feature descriptors.

### 5.4.1 Surface Gradients

The surface gradients surface measure is created by applying the standard SIFT feature descriptor extraction directly to the 2.5D range image Lowe (2004). This measure calculates the first derivative of the range image as seen from the viewing angle. The first derivative of the surface is then described as 8 rotation bins covering the 360° for each receptive field of in-plane surface orientation. The vote for the orientation of each surface gradient is weighted by its magnitude. The surface orientations vote into the receptive field histogram based on the strength of the gradient magnitude.

### 5.4.2 Shape Index

Shape index and curvature are frequently cited as a measure of surface invariant to out-of-plane orientation changes (Dorai and Jain, 1997; Koenderink and van Doorn, 1992; Lukins and Fisher, 2006; Lo and Siebert, 2009, 2008; Hetzel et al., 2001; Atmosukarto et al., 2010). In this dissertation shape index and curvature are implemented as a feature descriptor describing the range surface as described for intensity images in Section 3.3.4. Shape index is a function which varies between the limits of -1 to 1, and is capable of describing all surface types which may be expressed by a 3x3 surface patch. The shape index scale is equally partitioned into 8 bins, the contribution of each pixel to the feature descriptor histogram is weighted by the curvature of the keypoint.

### 5.4.3 Element Correction Surface Gradients

The Element Corrected Surface Gradients surface measure uses the standard 2.5D SIFT feature extraction stage applied to the range image surface gradients to create the feature descriptor. However, this approach introduces an additional step to calculate the mean value of the the surface gradients, over the whole 16x16 pixel sample patch, for both x and y directions. The mean gradients are removed from the gradient calculated for each pixel in the sample patch before applying histogramming to create the feature descriptor, Equation 5.22. This step creates a robust, low cost surface metric for the creation of a feature descriptor aimed at improving out-of-plane pose invariance.

$$desc_{SIFT} = \mathrm{Norm}\left(\{\mathrm{hist}(Grad_1), ...\mathrm{hist}(Grad_{16})\}\right) \tag{5.21}$$

$$desc_{EC} \;=\; \mathrm{Norm}\left(\{\mathrm{hist}\left(Grad_1 - \mathrm{mean}(Grad)\right), \ldots \mathrm{hist}\left(Grad_{16} - \mathrm{mean}(Grad)\right)\}\right) \tag{5.22}$$

### 5.4.4 2.5D Local SPIN Image

In addition to the proposed SIFT like features, SPIN image are frequently cited in the literature as local features invariant to changes in pose (Frome et al., 2004; Lai and Fox, 2009; Assfalg et al., 2007). The SPIN image local features implemented in this dissertation are a local feature representation based on the SPIN images outlined by Johnson and Hebert (Johnson and Hebert, 1999; Johnson, 1997). The local 2.5D SPIN images use the same localisation and feature vector length as standard SIFT features however are independent of in-plane orientation changes. The local 2.5D SPIN image feature descriptor approach considers all pixels within an 8 pixel radius of the keypoint centre, these pixels and their range values are represented as an X, Y, Z point cloud which is aligned with the surface normal of the keypoint. The co-ordinates of the aligned keypoints are then converted from Cartesian to polar co-ordinates, see Equation 5.23, giving the angle of location of the points in the point cloud as in-plane rotation angle, radius as distance from surface normal vector, and height from aligned plane tangential to surface at keypoint location, see Figure 5.10.

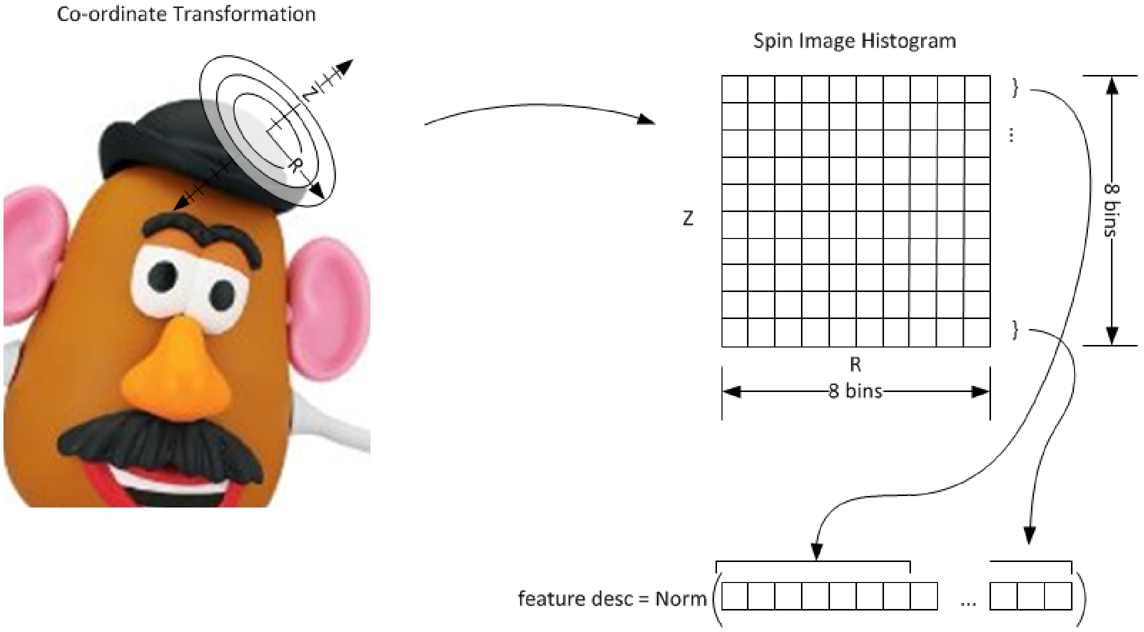


Figure 5.10: 2.5D Local SPIN Image Extraction

$$[\theta, R, Z] = \text{polar}\,(X, Y, Z) \qquad (5.23)$$

The 3D point cloud described in polar co-ordinates is then histogrammed in the $R$ and $Z$ dimensions to form the 2D SPIN image describing the local surface patch. In order to make the feature descriptor comparable to the SIFT descriptors presented in this chapter, the image patch is chosen to be the same as covered by the SIFT descriptor sample patch. In addition the quantisation of $R$ and $Z$ are chosen to give equal weight to each dimension and give a total number of bins equal to the number of elements in the SIFT feature descriptor. The quantisation of $R$ and $Z$ are shown in Figure 5.10. The resulting 2D SPIN image is linearised to form a 1D feature vector describing the image patch. This vector is then normalised to form the 2.5D Local SPIN Image vector.

As the $\theta$ dimension is ignored the resulting feature descriptor has a built-in in-plane orientation invariance. Used in conjunction with the normalisation of the point cloud to align with the surface normal at the keypoint location, this feature descriptor approach offers invariance to all three Euler angles. An overview of the descriptor extraction is shown in Figure 5.10.

| Description | Shortened Reference |
|---|---|
| 2D SIFT | SIFT_2D_2D |
| Surface Gradients | SIFT_3D_3D |
| Shape Index | SIFT_3D_3Dsi |
| Surface Gradients Element Correction | SIFT_3D_3Dec |
| Range SPIN Image | SPIN_3D_3D |
| Texture SPIN Image | SPIN_2D_2D |
| Surface Gradients & Shape Index | SIFT_3D_3Dsgsi |
| Surface Gradients Element Correction & Shape Index | SIFT_3D_3Decsi |
| Range SPIN Image & Shape Index | SPIN_SIFT_3D_3Dsi |
| Surface Gradients Element Correction & Shape Index & Range SPIN Image | SPIN_SIFT_3D_3Decsi |

Table 5.4: Sample patch description variations

### 5.4.5 Surface Measure Combinations

In addition to creating keypoints from a single surface measure of a local sample patch, a variety of surface measures can be combined in an approach similar to that taken in Section 5.2.3 for combining features from differing image modalities, Equation 5.25.

$$norm_{L2} = \sqrt{\text{sum}\left\{feat^2_{measure1}, feat^2_{measure2}\right\}} \tag{5.24}$$

$$feat = \frac{\{feat_{measure1}, feat_{measure2}\}}{norm_{L2}} \tag{5.25}$$

### 5.4.6 Naming conventions

This section has presented a range of approaches for forming a feature descriptor from a local sample patch region of range or intensity images. The range image has been described in terms of shape index, and variations based on the surface gradients; SPIN images for range and intensity images introduced. The implementation of each of these approaches has been outlined, showing the detailed calculation of the feature descriptor vector from the local image sample patch. The results for the outlined descriptors are presented in Section 5.5.3, the naming conventions of the descriptors outlined in this section are shown in Table 5.4.

## 5.5 Results

The performance of the local feature matching approaches outlined in this chapter were evaluated using the experimental setup detailed in Chapter 4. The results in this section are formed by evaluating the performance for matching between random query range and intensity image pairs to target examples of the 3D free form objects separated by 60°. The results for each method are presented as a table giving an example of the number of keypoints, the number of matches and the number of correct matches for the log-likelihood sensitivity level of 0.8. A second table showing the performance of each approach in terms of the percentage increase in the area under the ROC curve when compared with the unmodified SIFT case, and a ROC curve showing the matching quality for the feature descriptors for every sensitivity level. The percentage change in area under ROC curve is defined as the precentage change of the area under SIFT up to a given false positive rate, yellow area, to the area under a comparison ROC cuve to the same false positive rate, green area, see Figure 5.11. An analysis of the results observed in this section is presented in Section 5.6.

### 5.5.1 Multimodal Keypoints

The results for the multimodal keypoints proposed in Section 5.2, are presented here. Table 5.2 gives the naming conventions for the multimodal keypoints results, Table 5.7a shows an average of matching keypoints at the likelihood threshold of 0.8. In this table, overlapping refers to the number of keypoint location correspondences which exist between the query and model images; query keypoints refer to the number of keypoints extracted from the query image; matches refer to the keypoint descriptor matches between the model and query images; correct matches refer to the number of keypoints which have descriptor matches and location correspondences between the model and query images. Table 5.5b shows the percentage change of area under the ROC curve when compared to standard SIFT for false positive rates of 10%, 20% and 100%, Figure 5.12 shows the ROC curves.

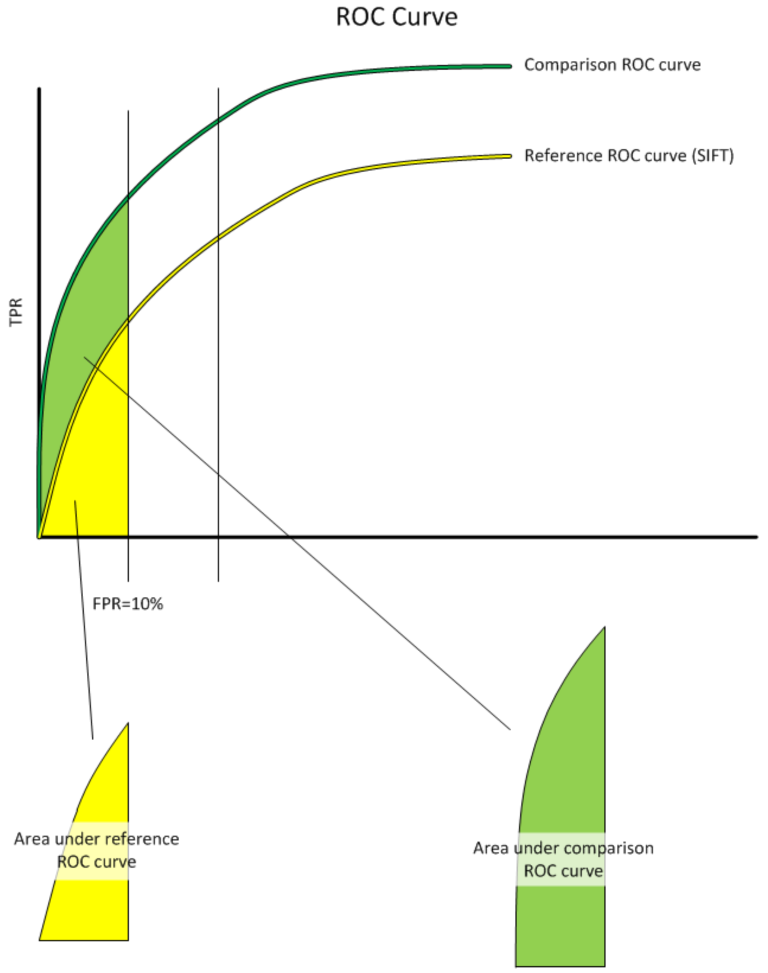

ROC Curve
Comparison ROC curve
Reference ROC curve (SIFT)
TPR
FPR=10%
Area under reference ROC curve
Area under comparison ROC curve

Figure 5.11: Area under ROC curve represented as a percentage change

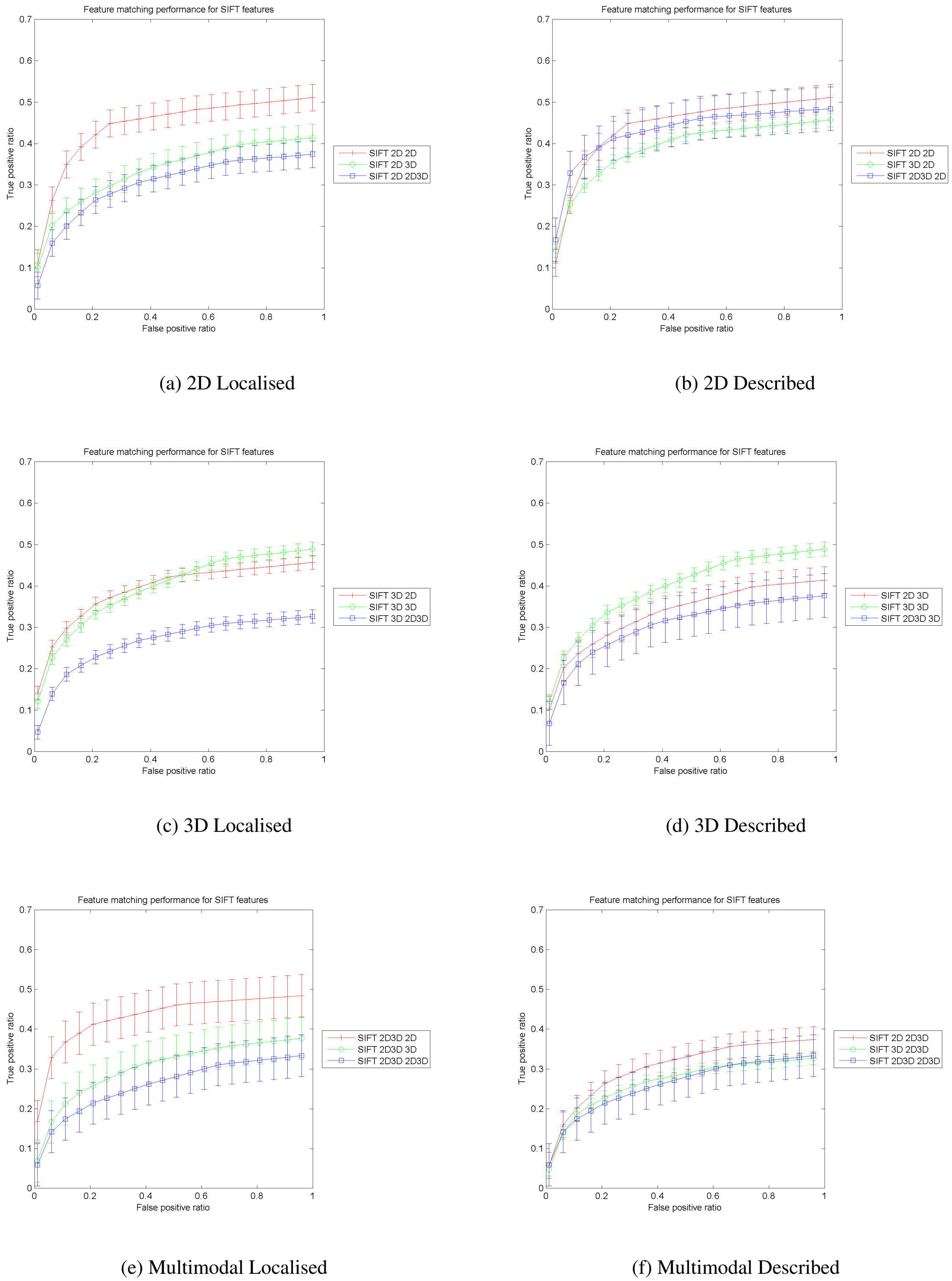
| SIFT description | Overlapping | Query Keypoints | Matches | Correct Matches |
|---|---|---|---|---|
| SIFT_2D_2D (standard SIFT) | 23.3 | 57.7 | 27.0 | 12.0 |
| SIFT_2D_3D | 28.1 | 53.1 | 33.5 | 11.6 |
| SIFT_3D_2D | 13.2 | 44.2 | 14.9 | 5.3 |
| SIFT_3D_3D | 18.5 | 38.9 | 22.1 | 8.1 |
| **SIFT_2D3D_2D** | **50.8** | **102.3** | **44.8** | **25** |
| SIFT_2D3D_3D | 50.5 | 93.3 | 56.0 | 18.8 |
| SIFT_2D_2D3D | 47.7 | 95.0 | 54.9 | 18.8 |
| SIFT_3D_2D3D | 42.3 | 101.4 | 51.5 | 14.3 |
| SIFT_2D3D_2D3D | 46.8 | 97.0 | 55.6 | 15.4 |

(a) Examples of matching at 0.8 log-likelihood threshold for multimodal keypoints

| SIFT description | FPR 10% | FPR 20% | FPR 100% |
|---|---|---|---|
| SIFT_2D_2D (standard SIFT) | 0 | 0 | 0 |
| SIFT_2D_3D | -25.9 | -31.2 | -24.3 |
| SIFT_3D_2D | 1.2 | -10.9 | -11.6 |
| SIFT_3D_3D | -11.8 | -19.4 | -10-6 |
| **SIFT_2D3D_2D** | **22.6** | **7.9** | **-2.6** |
| SIFT_2D3D_3D | -36.6 | -38.9 | -30.9 |
| SIFT_2D_2D3D | -40.5 | -41.5 | -31.4 |
| SIFT_3D_2D3D | -50.3 | -48.8 | -39.9 |
| SIFT_2D3D_2D3D | -46.8 | -49.5 | -40.8 |

(b) Percentage change for area under ROC curve

Table 5.5: Multimodal Keypoints



(a) 2D Localised

(b) 2D Described

(c) 3D Localised

(d) 3D Described

(e) Multimodal Localised

(f) Multimodal Described

Figure 5.12: Multimodal ROC Curves

| SIFT description | Overlapping | Query Keypoints | Matches | Correct Matches |
|---|---|---|---|---|
| SIFT 2D 2D | 23.3 | 57.7 | 27.0 | 12.0 |
| SIFT 3D 3D | 18.5 | 38.9 | 22.1 | 8.1 |
| SIFT aff 2D 2D 2D | 26.0 | 50.7 | 31.6 | 11.6 |
| **SIFT aff 2D 3D 2D** | **28.6** | **42.6** | **25.4** | **17.5** |
| SIFT aff 3D 2D 3D | 15.0 | 44.4 | 18.0 | 4.9 |
| SIFT aff 3D 3D 3D | 13.8 | 30.2 | 25.4 | 5.1 |
| SIFT pro 2D 2D 2D | 27.6 | 67.9 | 22.9 | 5.4 |
| SIFT pro 2D 3D 2D | 17.6 | 54.3 | 19.0 | 6.6 |
| SIFT pro 3D 2D 3D | 17.8 | 37.8 | 18.1 | 4.5 |
| SIFT pro 3D 3D 3D | 14.9 | 23.2 | 28.7 | 3.7 |

(a) Examples of matching at 0.8 log-likelihood threshold

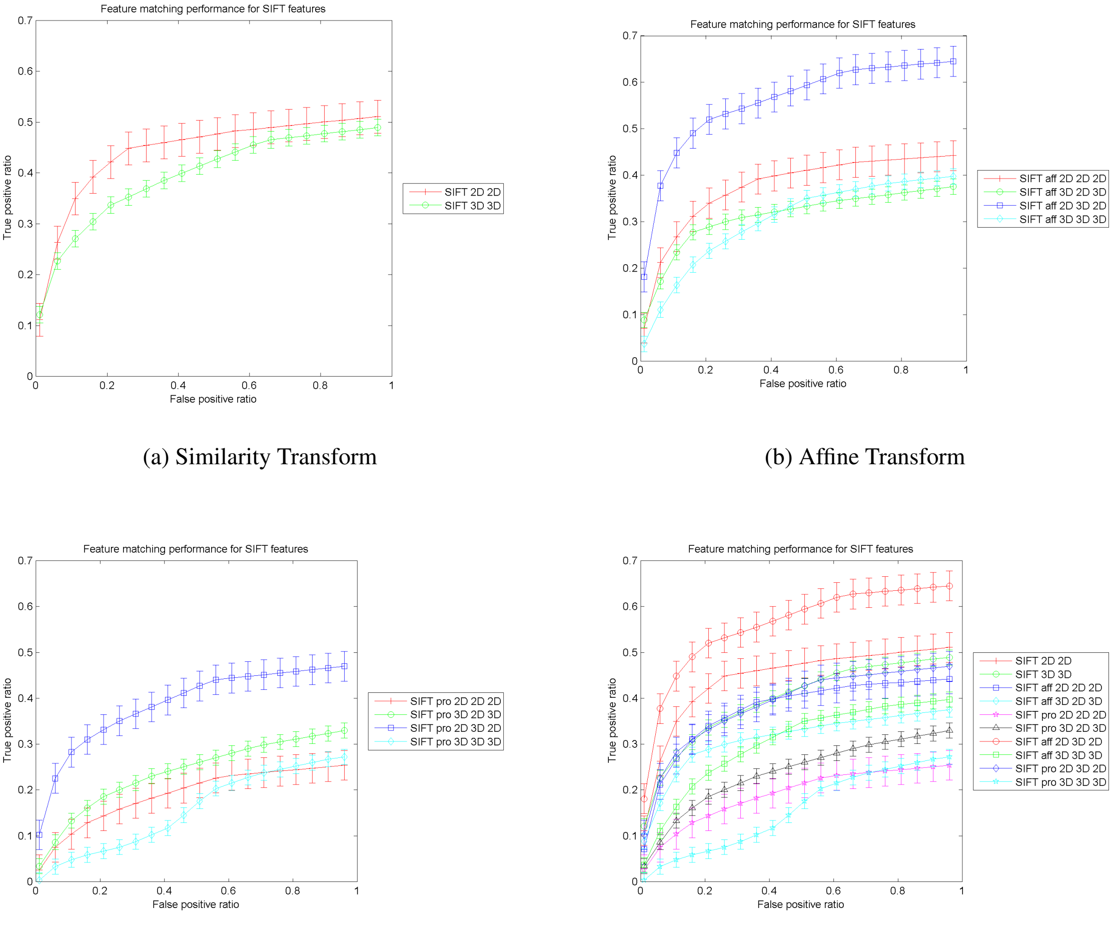
| SIFT description | FPR 10% | FPR 20% | FPR 100% |
|---|---|---|---|
| SIFT 2D 2D | 0 | 0 | 0 |
| SIFT 3D 3D | -11.8 | -19.4 | -10.6 |
| SIFT aff 2D 2D 2D | -22.5 | -21.9 | -15.4 |
| **SIFT aff 2D 3D 2D** | **44.1** | **31.2** | **25.9** |
| SIFT aff 3D 2D 3D | -29.4 | -31.5 | -29.2 |
| SIFT aff 3D 3D 3D | -58.0 | -52.2 | -31.4 |
| SIFT pro 2D 2D 2D | -70.9 | -69.4 | -56.3 |
| SIFT pro 2D 3D 2D | -13.2 | -18.5 | -12.4 |
| SIFT pro 3D 2D 3D | -67.4 | -63.2 | -45.1 |
| SIFT pro 3D 3D 3D | -88.9 | -86.5 | -64.6 |

(b) Percentage change for area under ROC curve

Table 5.6: Sample Patch Corrected Keypoints

### 5.5.2 Sample Patch Warping

The results for sample patch warped keypoints proposed in Section 5.3, are presented here. Table 5.3 gives the naming conventions for the multimodal keypoints results, Table 5.6a shows an average of matching keypoints at the likelihood threshold of 0.8, Table 5.7b shows the percentage change of area under the ROC curve when compared to standard SIFT for false positive rates of 10%, 20% and 100%, Figure 5.13 shows the ROC curves.



(a) Similarity Transform

(b) Affine Transform

(c) Projective Transform

(d) All Transforms

Figure 5.13: Sample Patch Warping ROC Curves

| SIFT description | Overlapping | Query Keypoints | Matches | Correct Matches |
|---|---|---|---|---|
| **SIFT_2D_2D** | **23.3** | **57.7** | **27.0** | **12.0** |
| SIFT_3D_3D | 18.5 | 38.9 | 22.1 | 8.1 |
| SIFT_3D_3Dsi | 16.3 | 45.1 | 17.0 | 7.9 |
| SIFT_3D_3Dec | 19.3 | 45.8 | 23.2 | 9.0 |
| SPIN_3D_3D | 15.6 | 33.9 | 21.7 | 6.9 |
| SPIN_2D_2D | 21.1 | 75.9 | 9.6 | 3.0 |
| SIFT_3D_3Dsgsi | 16.9 | 44.2 | 21.2 | 8.2 |
| SIFT_3D_3Decsi | 19.7 | 52.9 | 18.2 | 9.3 |
| SPIN_SIFT_3D_3Decsi | 14.6 | 50.5 | 9.8 | 5.1 |
| SPIN_SIFT_3D_3Dsi | 7.6 | 57.5 | 7.5 | 2.7 |

(a) Examples of matching at 0.8 log-likelihood threshold

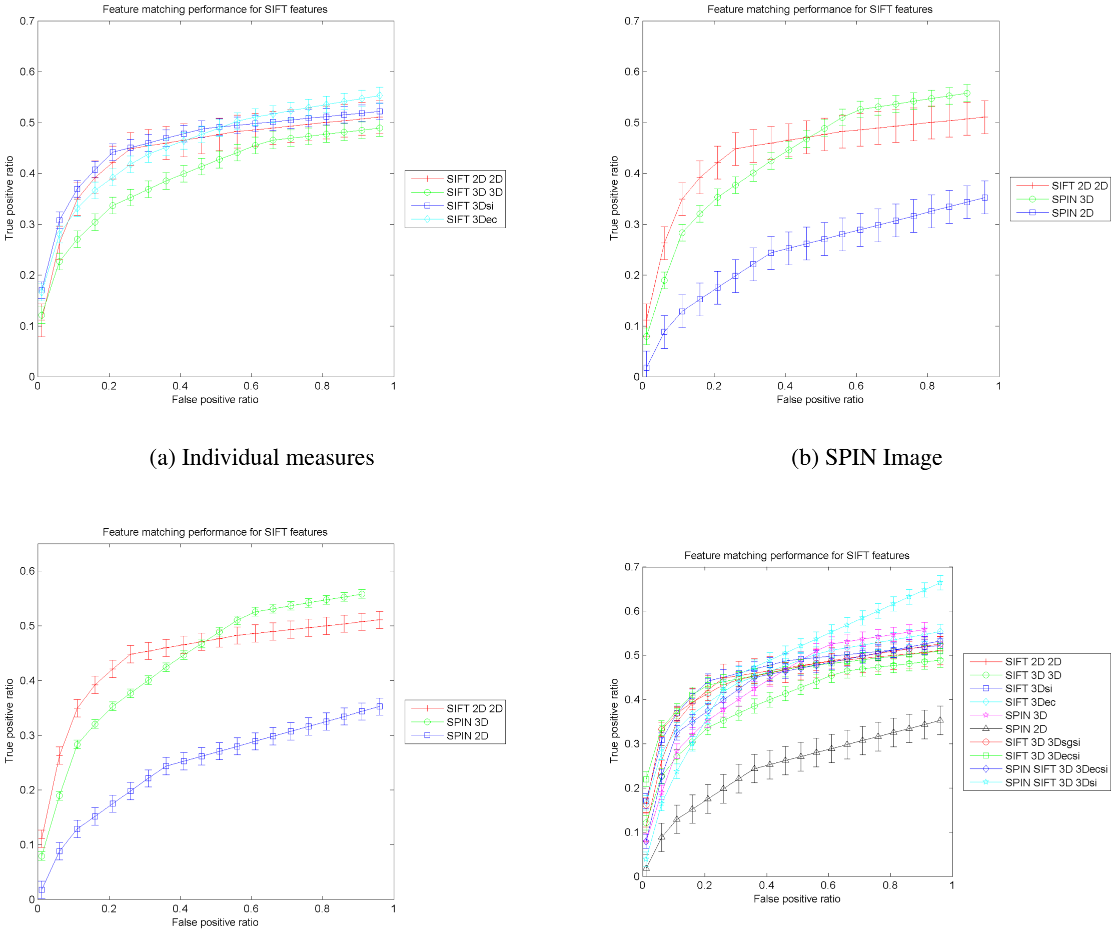
| SIFT description | FPR 10% | FPR 20% | FPR 100% |
|---|---|---|---|
| SIFT_2D_2D | 0 | 0 | 0 |
| SIFT_3D_3D | -11.8 | -19.4 | -10.6 |
| SIFT_3D_3Dsi | 18.3 | 8.2 | 3.5 |
| SIFT_3D_3Dec | 9.1 | 0.1 | 3.1 |
| SPIN_3D_3D | -25.7 | -22.0 | 39.7 |
| SPIN_2D_2D | -68.4 | -58.9 | -44.4 |
| SIFT_3D_3Dsgsi | 28.2 | 11.3 | 2.1 |
| **SIFT_3D_3Decsi** | **38.7** | **15.5** | **2.0** |
| SPIN_SIFT_3D_3Decsi | -12.7 | -7.0 | -1.4 |
| SPIN_SIFT_3D_3Dsi | -35.0 | -22.6 | 8.0 |

(b) Percentage change for area under ROC curve

Table 5.7: Surface Description Variants

### 5.5.3 Surface Description

The results for sample patch warped keypoints proposed in Section 5.4, are presented here. Table 5.4 gives the naming conventions for the multimodal keypoints results, Table 5.7a shows an average of matching keypoints at the likelihood threshold of 0.8, Table 5.7b shows the percentage change of area under the ROC curve when compared to standard SIFT for false positive rates of 10%, 20% and 100%, Figure 5.14 shows the ROC curves.



(a) Individual measures

(b) SPIN Image

(c) Combined Surface Measures

(d) All Surface Measures

Figure 5.14: Surface Description ROC Curves

## 5.6 Analysis

This section analyses the results reported in the previous section. The successes and failures of each of the proposed approaches are discussed and conclusions are drawn regarding the underlying causes of these.

### 5.6.1 Multimodal Keypoints

From the comparison of the multimodal features approaches outlined in Section 5.2, it was found that the approach using multimodal keypoint localisation with intensity domain image patch description produced a significant performance improvement in terms of both achieving the highest number of correct matches and increasing the ROC curve performance when compared to the unmodified SIFT case. This is the only instance of cross modal, or multimodal combinations for localisation or description which produced a marked improvement, all others resulted in significantly diminished performance. The following subsections investigate the localisation and description stages separately to form conclusions.

#### 5.6.1.1 Keypoint localisation

A comparison between the ROC performance at 10% FPR for single domain localised and described keypoints shows that for both range and intensity described features, keypoints localised in the range domain out perform those localised in the intensity domain, see Figure 5.8. However, this performance improvement in terms of ROC curve characteristics is offset by fewer keypoints localised in the range domain. This result indicates that the range imaging modality does produce a more stable keypoint localisation stage, however produce fewer keypoints per image due to the low variation in surface.

#### 5.6.1.2 Keypoint description

Keeping the domain in which the keypoints are localised constant, the best ROC curve performance were achieved using feature descriptors created from image patches collected in the intensity domain. The poorest performance from the keypoint extraction approaches evaluated resulted from the concatenation of feature descriptors extracted from both intensity and range imaging modalities. This result indicates that the keypoint locations which formed highly distinctive features in range

| SIFT Description | FPR 10% | Overlapping |
|---|---|---|
| SIFT_2D_2D (standard SIFT) | 0 | 23.3 |
| SIFT_3D_2D | 1.2 | 13.2 |

(a) Intensity described features

| SIFT Description | FPR 10% | Overlapping |
|---|---|---|
| SIFT_2D_3D | -25.9 | 28.1 |
| SIFT_3D_3D | -11.8 | 18.5 |

(b) Range described features

Table 5.8: Comparison of localisation approaches

and intensity images separately are not the same locations.

### 5.6.2 Sample Patch Warping

The experimental results for sample patch warping SIFT approaches showed that a significant improvement in matching performance was achieved when creating 2D features, with an affine sample patch warping guided by the range image surface gradients. Affine corrected features based on the intensity images is an approach similar to the affine features proposed by Schmid and Mikolajczyk (Mikolajczyk et al., 2005; Mikolajczyk and Schmid, 2004). The comparison between the use of intensity and range information to estimate the image patch pose shows that the intensity image gradients were insufficient for estimating an invariant patch warping on the free form 3D objects used in this experimental configuration.

In every instance where image surface gradients were used to both estimate the surface pose and localise the keypoint, the performance was significantly degraded. This result indicates that keypoints are localised in areas where the local surface gradients are unstable. Therefore, applying a corrective transform to the local image patch with a local pose estimated from the same imaging domain as the keypoint localisation creates inherently unstable feature descriptors.

Extending the out of plane orientation invariance to evenly sample the surface from a perpendicular view point decreased the performance of the resulting local features, contrary to the hypothesis. Furthermore no modifications to the feature extraction and sample patch corrective stage resulted in a performance improvement in the range imaging modality, when compared to the unmodified case: SIFT_3D_3D. The remainder of this subsection investigates the factors which influence the performance of local features using sample patch warping approaches.

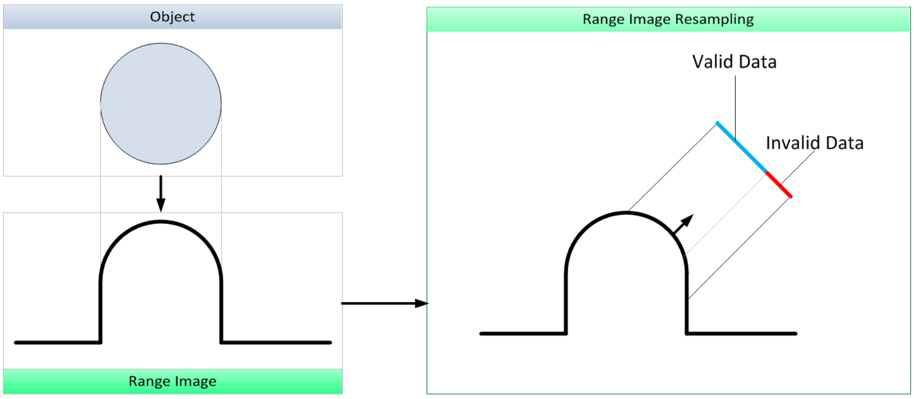


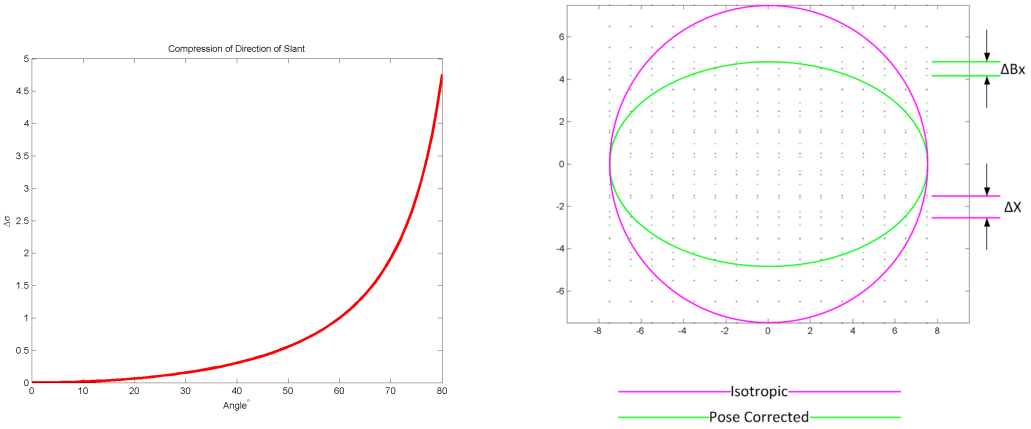
Figure 5.15: Resampling the Range Image

#### 5.6.2.1 Instability of Keypoint Localisation

An error in keypoint localisation for keypoints localised in regions where the image surface demonstrates a large variation cause the pose estimation stage to vary widely between keypoint instances. When using sample patch warping, this error in keypoint localisation manifests as a modification to the sample patch. The error in sample patch footprint exists only where the warping is applied and does not affect the similarity transformed features used in standard SIFT. Therefore reducing the corrective measures applied to the image patch increases the invariance to noise in both keypoint localisation and local pose estimation.

#### 5.6.2.2 Invalid data

Range images do not contain reliable connectivity information for all locations on the imaged object surface. This effect manifests as a smooth transition in the range surface between the visible object surfaces at occlusion boundaries. In the range image this is represented as a flat surface joining two locations, and masks cases where more underlying structure in the 3D object may be present. With projective corrected keypoints, projecting sample points to lie perpendicular to the surface causes many sample points to lie on a “range shadow” of invalid data and therefore decreases the repeatability of the feature descriptor, see Figure 5.15.



(a) Compression against Angle

(b) Sample Patch and Calculation of Compression

Figure 5.16: Compression of Sampling Scale

#### 5.6.2.3 Axis compression

The sample warping processing applied to the local window sample patch, used to create pose corrected features, changes the distance between sampled locations in the sampling patch. This change of distance between sampling points effectively increases the sampling frequency in direction of the dominant slant of the sample patch, causing the keypoint to be extracted at an incorrect scale for the direction of of maximal axis compression. Equation 5.26 defines the effect of axis compression as a change in scale, $\Delta\sigma$, where $\Delta X$ is the unaltered seperation between two sample points, and $\Delta Bx(\tau)$ is the effective sample point seperation for sample patch rotation angle, $\tau$. This effect limits the out-of-plane invariance afforded by affine corrected features to around $\pm 30°$ for the scale space pyramid used in this study with scale divisions of $\sigma = 1.4$ between levels, see Figure 5.16a. To mitigate this effect, a scale space pyramid which accounts for slant and tilt axes compression could produce more stable keypoints for feature extraction using affine pose correction.

$$\Delta\sigma = \frac{1}{\Delta Bx(\tau)} - \frac{1}{\Delta X} \tag{5.26}$$

### 5.6.3 Surface Description

Approaches to construct surface descriptors included an comparison of feature descriptors collected using individual surface measures and combinations of differing surface measures. Shape index gave the most reliable individual surface measure. The surface measure using surface gradients with element correction significantly improved the performance of the surface gradients approach giving a ~20% increase of the area under the ROC curve at FPR 10%, and a ~10% greater area under the ROC curve at FPR 10% when compared to standard SIFT. SPIN images produced poorer results than the baseline SIFT in both ROC curve performance, and number of keypoints. SPIN images performed better in the range image domain than in the intensity. In the range domain a higher true positive ratio than the baseline SIFT was observed when a high level of false positives were tolerated.

The combination of differing surface measures showed that combining surface measures from SIFT-like features gave better ROC curve characteristics than individual surface measures and than the baseline SIFT applied to the intensity domain image. The optimum ROC performance was achieved using a combination of shape index and element corrected surface gradients. However, the performance improvement observed is marginal. Additionally, all the combination approaches investigated use range localised features which resulted in fewer keypoints being detected when compared with intensity localised approaches.

The following subsections investigate the characteristics of the surface measures and improvements which can be made to improve the performances of each of these surface measures.

#### 5.6.3.1 Surface descriptor foot prints

To measure the shape index and curvature for each point in the sample patch, a footprint of the 8 neighbouring pixels surrounding the keypoint is used. The extraction of the surface gradients measure however uses a smaller footprint of 4 pixels, see Figure 5.17. The original performance improvement observed when using shape index may have resulted from an increased measurement stability resulting from the larger footprint. To investigate this effect a modification to element correction surface gradients is proposed which makes use the same footprint as used for calculating shape index, Figure 5.17. Evaluating the performance of the modified surface gradients with element correction shows that this effect is significant and that the resultant approach demonstrates state of the art performance, Figure 5.18.

Combining the modified element correction surface gradients approach with the shape index approach increases the true positive ratio for low tolerance of false positives; however it decreases

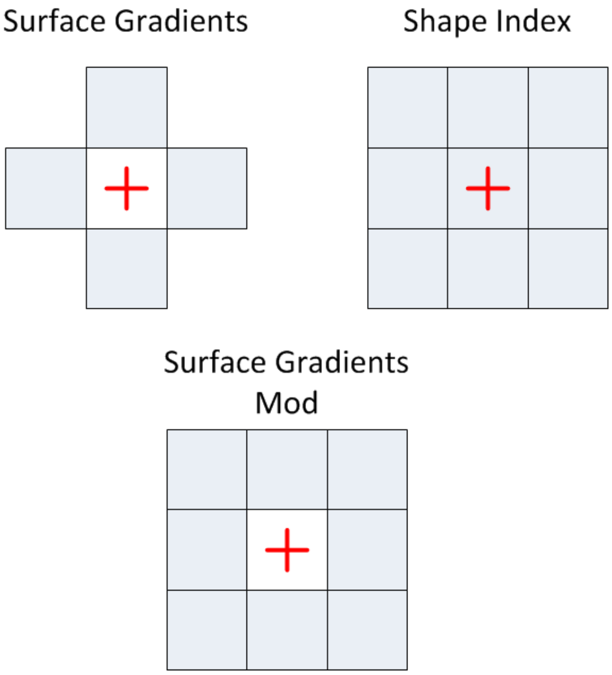


Figure 5.17: Surface Measure Footprint

the true positive ratio for FPR of greater than 5%, Figure 5.18.

#### 5.6.3.2 SPIN image normalisation

The poorest performance of the proposed surface measures resulted from the SPIN image. A potential cause for the degraded performance of the SPIN image approach is the uneven distribution of samples in the voting space used to create the feature descriptor. This uneven distribution of samples in the voting space effectively reduces the information content in the feature descriptor.

The SPIN image aligns the X, Y, Z location of points on the range image surface to the dominant patch pose, then applies a Cartesian to polar co-ordinates transform to represent the range surface as $\theta$, $R$, $Z$. The resultant $R$ and $Z$ values are histogrammed to create the feature descriptor, with even divisions in radius and $Z$, see Figure 5.10. However, as the radius increases, each division of radius covers an increased area giving a greater number of votes to the corresponding elements in the feature descriptor, see Figure 5.19.

Furthermore the scale space approach applied to the range or intensity image ensures that all

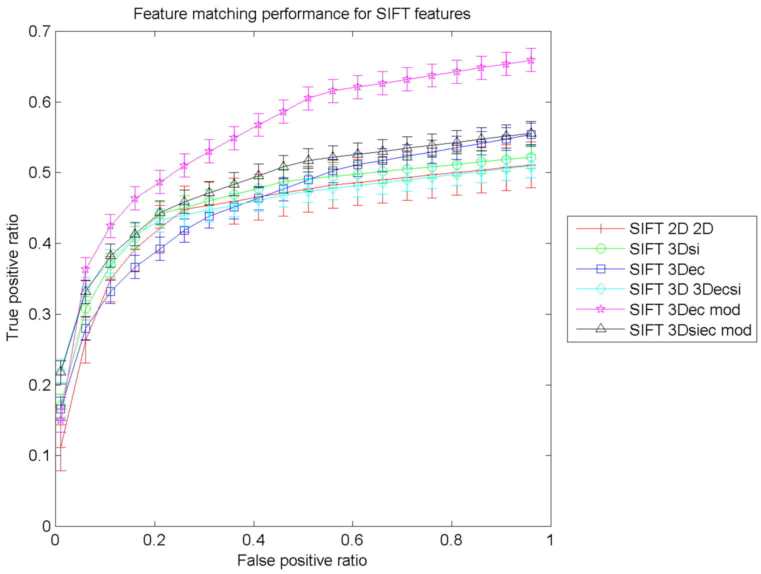


Figure 5.18: Corrected footprint ROC curve

image patches contain a fixed set of spatial frequencies. This low variation in surface topology will cause a consistent uneven distribution of elements in the voting space, where few range values will be binned in the high $Z$ low $R$ bins. Due to this typical distribution of range values in the SPIN image histogram, the SPIN image histogram is not an effective use of a 128 element feature descriptor.

### 5.6.3.3 Range Image Descriptor Normalisation

The normalisation stage which transforms a histogram of surface gradients into a feature descriptor vector of unit length is required in order to perform the comparison between two feature vectors in the matching stage. The normalisation stage also has the additional effect of removing information regarding the level of variation present in the sampled surface. In the intensity domain this normalisation step offers invariance to variations introduced into the image as a result of illumination effects. However, in the range imaging modality the variations removed by the normalisation stage encode underlying characteristics of the range image surface at the keypoint location.

The normalisation stage has an additional effect in the range image: the projection of sample

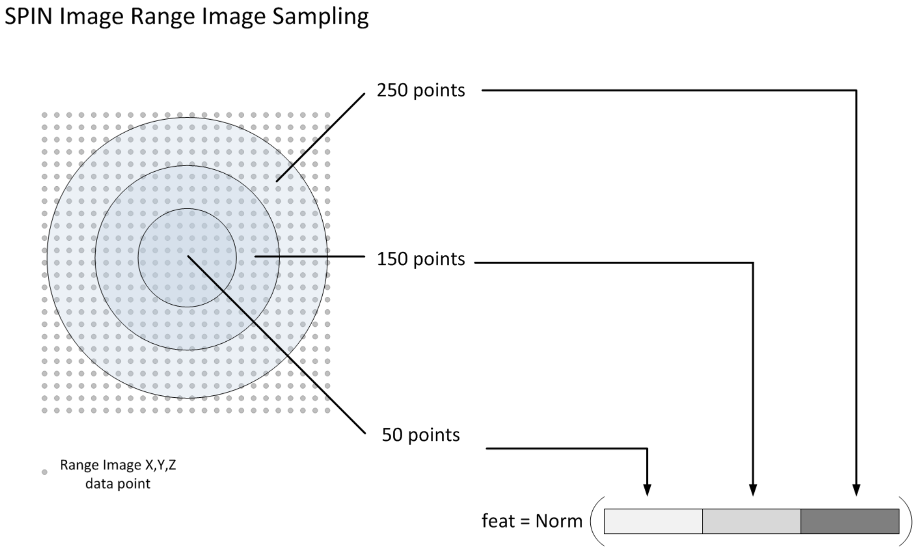


Figure 5.19: SPIN Image Range Sampling

points at an occlusion boundary cause a spike in the variation of surface gradients observed in the feature descriptor before normalisation, see Figure 5.20. It is not possible to know the depth at which the projected sample points will lie when projected past the keypoint location, as the occlusion boundaries in the range image are unknown. Furthermore, the projected sample points will tend to have a large change in range between views. This large change in depth results in a large influence in the normalisation step in the feature descriptor creation. This effect decreases the confidence of keypoints along occlusion boundaries, as occlusion boundary information is allowed to dominate over the informative elements in the feature vector.

## 5.7 Summary and Conclusions

In this chapter a range of modifications to the local feature extraction processing pipeline have been investigated. The MATLAB SIFT code used in previous chapters was further developed to incorporate proposed modifications. The unmodified SIFT code was used as a baseline for comparison in the new experimental configuration outlined in the previous chapter. The study conducted in this chapter focuses at every stage of the processing pipeline on the use of multimodal information from the co-registered range and intensity images. The main outcomes of these experiments are outlined here.

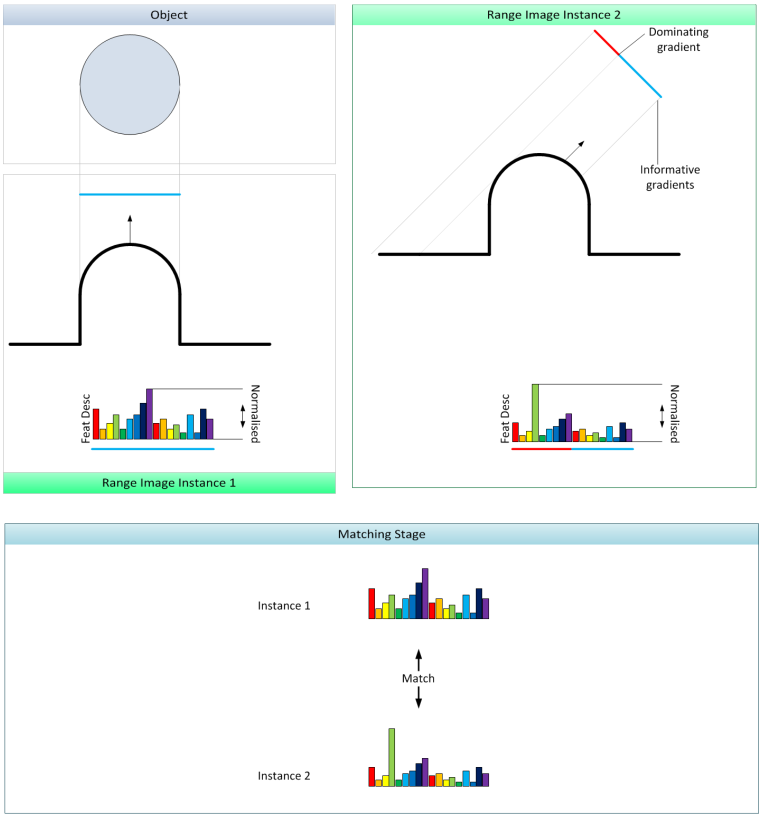


Figure 5.20: Range Image Surface Gradients Sampling

Localisation of keypoints in the range domain creates fewer keypoints than localisation in the intensity domain. The keypoints localised in the range domain formed more distinctive features for a FPR of 10% and below, however intensity domain localised features had a higher recall. Intensity domain described features outperformed range domain descriptions. The optimum multimodal combination of localisation and description resulted from keypoints localised in both the range and intensity domains, and described only in the intensity domain. Multimodal information can be used to apply a warping to the image patch estimated from cross modal information. The effects of applying a warping to the sample patch were investigated using 3 levels of warping: similarity, or no warping; affine, or fitting the sample patch to a plane; and projective warping, resampling the range image from a new viewer position. The invariant surface measure used to estimate the warp to apply to the image patch between keypoint instances can be estimated from either the range or intensity domain, with the description of the image patch formed from either the range or intensity domain independently. It was found that keypoints localised in the intensity domain with an affine warp estimated from the surface gradients of the range image, and an image patch description created from the intensity domain gave the best performance. The high performance observed in using the affine corrected sample patches and not the projective corrected, in part resulted from a constrained sample patch arrangement which is robust to noise in the range domain. In addition keypoints localised in the range domain with the affine warping approach applied showed degraded performance, indicating that these keypoints were localised in regions with high range domain variance and form an unstable local pose estimation.

A range of surface description measures were proposed for describing the characteristics of the image patch extracted in the range domain, these included popular descriptions such as shape index, surface gradients and SPIN images. In addition to these the novel surface description, element corrected surface gradients, was proposed based on surface gradients, this surface measure addresses the issue of out-of-plane view change. The element corrected surface gradients surface measure advances state of the art performance and achieves results comparable to range affine corrected intensity SIFT features. The ROC performance of the optimum descriptors from each section is shown in Figure 5.21.

State of the art performance is advanced through 2 of the approaches outlined in this chapter. However, the modifications made to each of these approaches cannot be combined. Furthermore, partial composition information encoded in the range data is under utilised. The next chapter seeks to use this additional information from the range image to identify multiple instances of features across range image examples, and combine these in a learning phase to model the modes of expression of the feature descriptor.

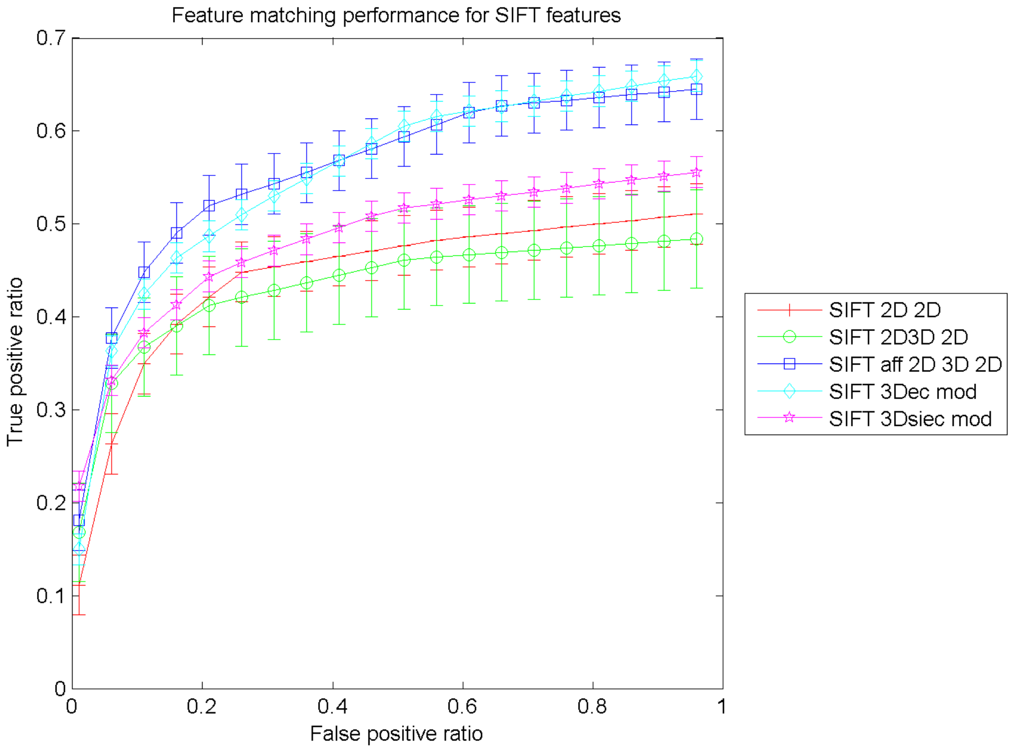


Figure 5.21: Optimum Descriptors

# Chapter 6

# Multi Observation Keypoint Combinations

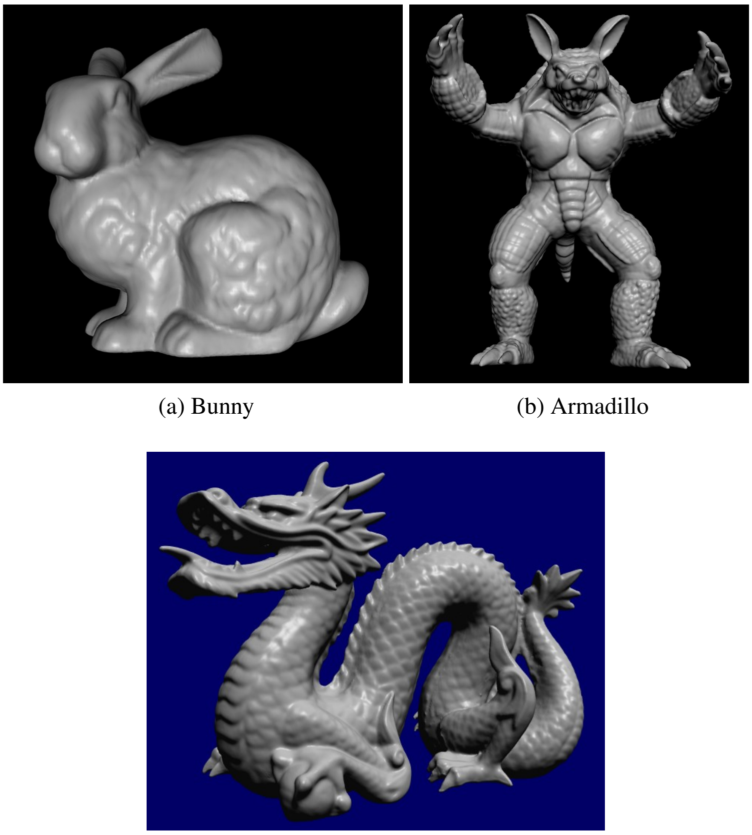
Chapter 5 investigated a number of uses of range data for increasing the reliability of feature matching under changes in viewpoint based on a single observation of the keypoint. This chapter extends the individual features from a single viewing instance by modelling the feature vector space expressed by a keypoint from differing viewing angles. The rational behind this approach is to assume that out-of-plane view point changes can be characterised as a deformation of a keypoint sample patch and therefore can be treated using adaptive object recognition approaches. The composition of the local range surface surrounding each keypoint serves to guide the integration of feature descriptors. This chapter covers the combination of keypoints locations from differing viewing angles to form a set of *"3D local interest points*". From this set of 3D local interest points a number of approaches to integrate keypoint feature descriptors observations from differing viewing angles are investigated.

## 6.1 Overview

The investigation into local features in range and intensity modalities in previous work and chapters has focused on recognition from a single observation instance, using the available local information to form a descriptor which remains invariant across observations. However, 3D data, range data and time series data offer additional information regarding the variety of feature descriptors which a given keypoint can express. For humans this ability to explore the view observation space of a query object can allow the viewer to disambiguate between many multiple object or feature hypotheses (Palmer, 1999). Recent category recognition approaches have achieved much progress in disambiguating between object hypotheses in a range of poses through a similar approach of exploring the observation space to learning the local feature view space based on collections of single instance features present in an image (Yang et al., 2007). This chapter investigates whether it is possible to encode extra information from the view space into the feature descriptor, such that individual low level feature confidence can benefit from information gained across a range of observations.

To create pose invariant feature descriptors in this chapter it is proposed that a measure of mean and variance of feature descriptors observed at a range of views are collected to be used in the matching stage. This approach will allow the variation in the feature descriptor space resulting from changes in viewing angle of a keypoint to be modelled. The modelled variations can be used as tolerable variations in the feature descriptor space, in which the weightings of distances in covariant dimensions between query and target feature descriptors are reduced. Reducing the distances between model and target descriptors has the effect of making feature descriptors collected from a single keypoint on a 3D object appear more similar when matched across a range of views. In order to investigate this approach, common keypoint locations between instances must be established. In addition this chapter estimates the local pose of the surface in order to restrict the range of keypoint poses to be included in the descriptor statistics, thereby limiting the variance of feature descriptors.

Common locations on an object surface between observations can be established through optical flow or 3D surface alignment approaches. In this chapter range images are synthetically created from 3D polygon meshes of objects available from the Stanford Scanning Repository, see Figure 6.1. The transformation between viewing angles used to create the synthetic range images is also used to establish keypoint location correspondences between observation instances. Keypoint locations which are repeatably localised between observations of a specific object are stored as significant 3D interest points. The significant 3D interest points have descriptors collected from a range of observations where the keypoint location is detected. The collected descriptors are used to create a model of the descriptor variation space using a range of learning approaches investigated in Section 6.1.3.



(a) Bunny

(b) Armadillo

(c) Dragon

Figure 6.1: Examples of scanned objects from the Stanford Scanning Repository

The principal contributions of this chapter are to outline a methodology to associate keypoints between views to create 3D local interest points, and to investigate local feature learning in the context of keypoints observed through variations in pose. The chapter is outlined as follows, Section 6.1.4 details an overview of the 3D feature extraction process, Section 6.1.1 finds 3D local interest points, Section 6.1.1 associates keypoints together to create a feature descriptor example collection for the 3D local keypoints, Section 6.1.3 describes a range of statistical approaches which can be applied at the matching stage and Section 6.2 describes the evaluation approach used in this chapter to determine the performance of matching in the descriptor space with differing statistical measures based on the precision recall characteristics of all keypoints observed on an object.

### 6.1.1 3D Local Interest Points

The main aim of this chapter is to investigate the combination of keypoint feature descriptors from differing viewing observations as a means to create a pose invariant keypoint descriptor which

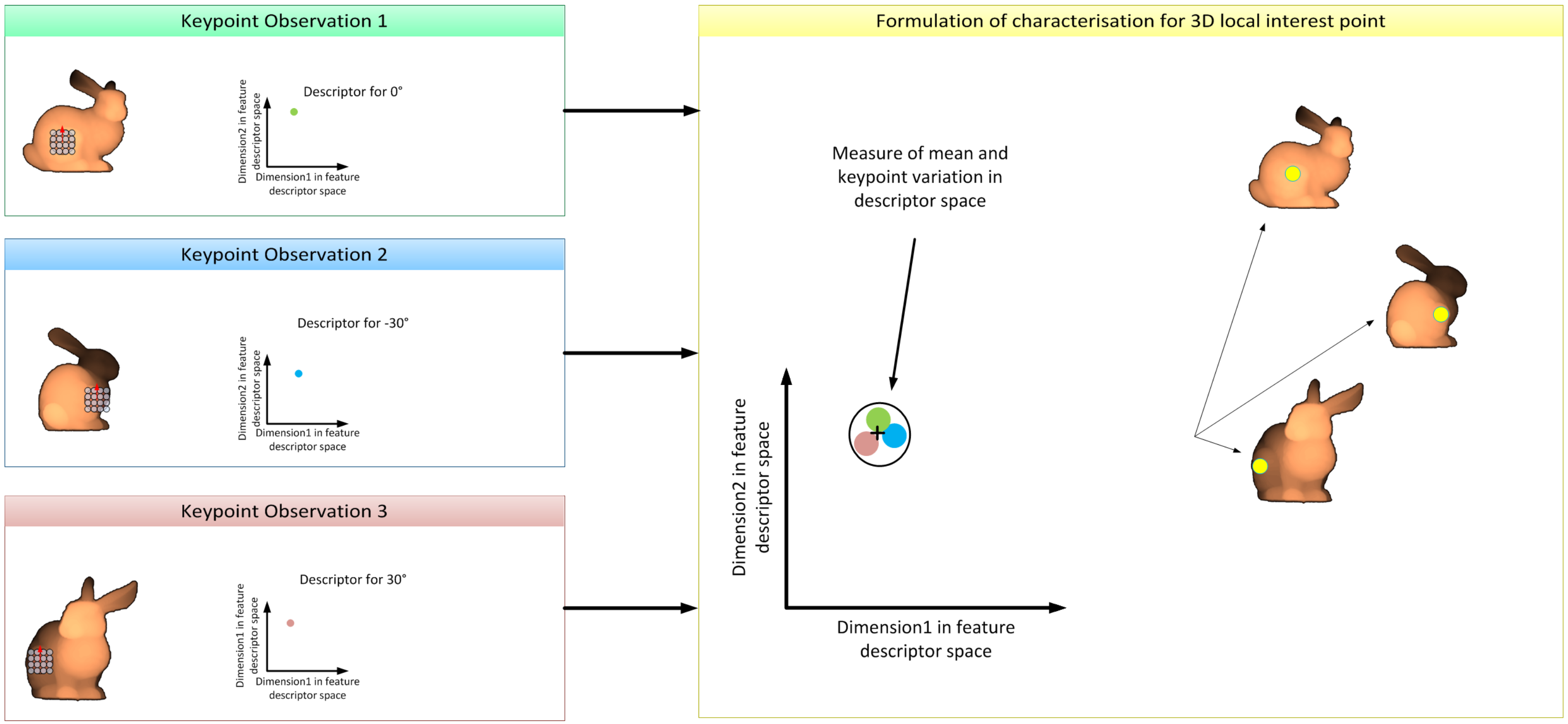

Keypoint Observation 1
Descriptor for 0°
Dimension2 in feature descriptor space
Dimension1 in feature descriptor space
Keypoint Observation 2
Descriptor for -30°
Dimension2 in feature descriptor space
Dimension1 in feature descriptor space
Keypoint Observation 3
Descriptor for 30°
Dimension1 in feature descriptor space
Dimension1 in feature descriptor space
Formulation of characterisation for 3D local interest point
Measure of mean and keypoint variation in descriptor space
Dimension2 in feature descriptor space
Dimension1 in feature descriptor space

Figure 6.2: Multi Instance Feature Overview

generalises to keypoints from a range of observations. In order to investigate this main aim in the chapter, keypoints detected in differing observations must be associated with a set of 3D *Local Interest Points* for an object which are common to all views. In this chapter, the 3D models used to create the range images are available. Therefore, it would be possible to form the 3D Local Interest Points directly from the 3D models using segmentation approaches based on the underlying object surface structure (Dorai and Jain, 1997; Mangan and Whitaker, 1999), or a measure of the composition of points in the point cloud (Lian et al., 2011). However, as this work aims to create a keypoint representation which is matchable to a range of observations, the keypoint localisation and detection for the 3D Local Interest Points should be similar to the localisation and detection applied in a single range image observation. Therefore, this work applies an approach similar to Ohbuchi et al (Ohbuchi et al., 2008), where interest points from all views are collected and integrated together using the standard SIFT localisation.

This section describes the combination of local features detected from multiple viewing angle observations of an object. Keypoint detections from all observations are projected into a common canonical space. Keypoint locations in the canonical space can be seen to cluster around 3D interest points, see Figure 6.3. However, the canonical space also contains spurious detections. These spurious detections are removed using 3D density filtering, Section 6.1.1.2. The remaining keypoints are then clustered to group the individual keypoints into collections of descriptors representing a 3D interest point.

#### 6.1.1.1 Keypoint Detection

Keypoints are detected using the Harris Corner detector interest point locator applied to a scale space pyramid derived from the synthetic range image, see Figure 6.4a. The 3D location of a keypoint is found from interpolating the range image value at the keypoint location. Keypoints which lie not on the object but on the range image background are removed with a mask. The background mask is created from the areas of the range image which do not lie on the object hull, see Figure 6.4b.

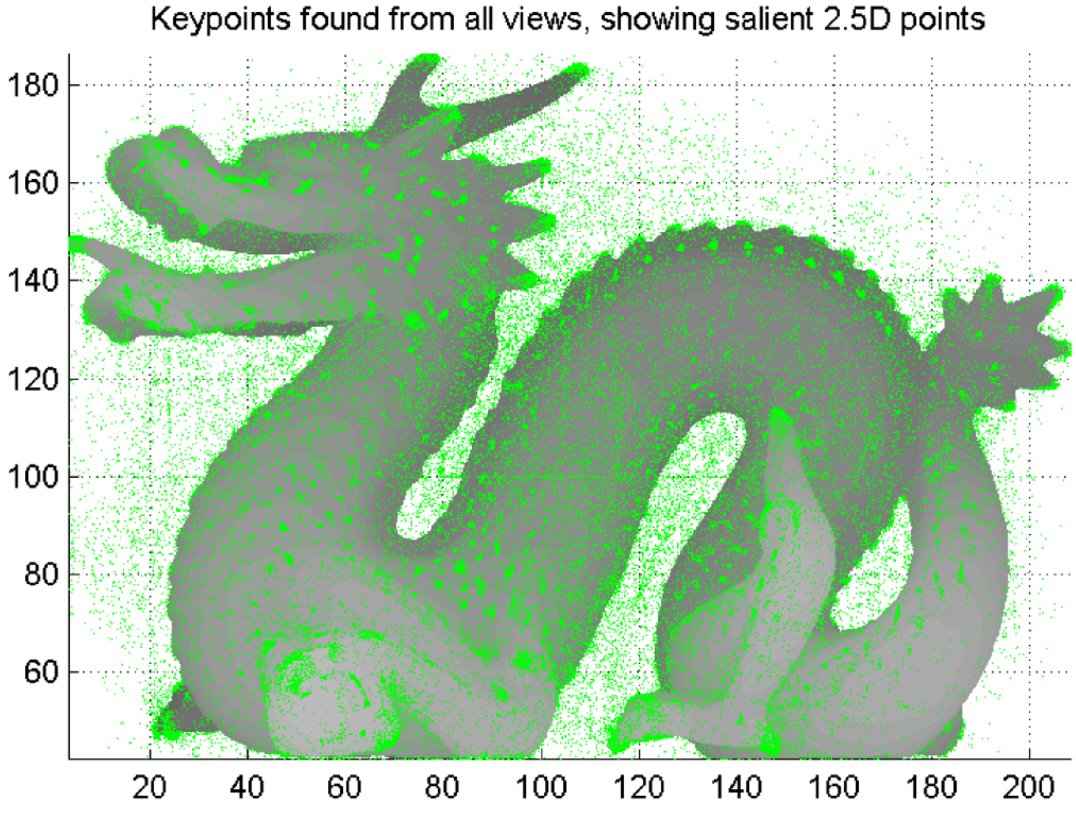
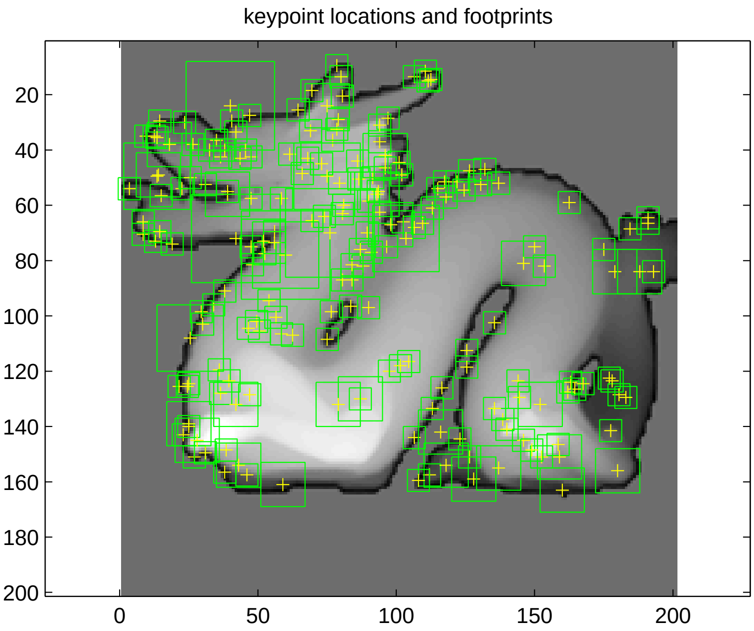
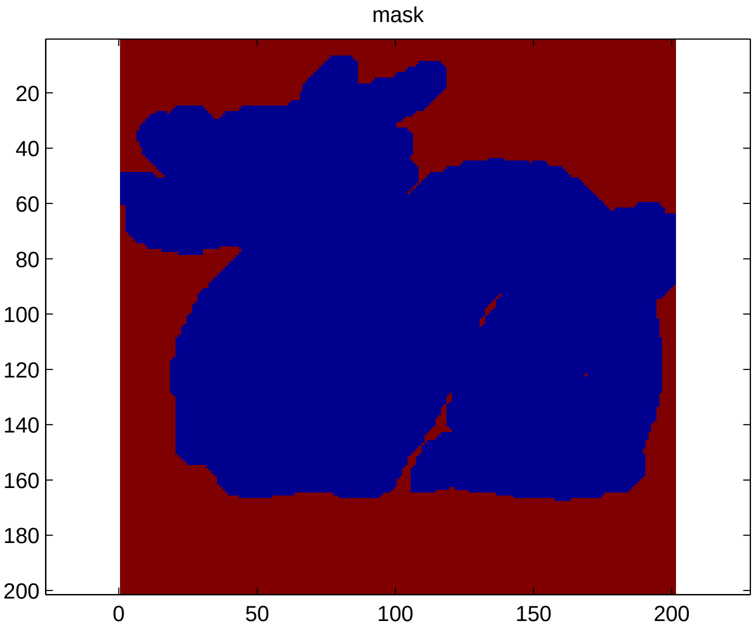


Figure 6.3: Keypoints from all views displayed in the canonical space



(a) Synthetic Range Image with Detected Keypoints and footprints



(b) Synthetic range image background mask shown in red

Figure 6.4: Keypoints Detected in Synthetic Range Images

#### 6.1.1.2 Keypoint Filtering

Keypoint detections from a collection of single observation range images projected into a common object based co-ordinate frame, here termed the *canonical space*, tend to cluster around 3D local interest points, see Figure 6.3. However spurious keypoint locations can also seen to be present. Directly applying the keypoint clustering without removing these noisy keypoint locations will negatively affect the localisation of the 3D interest points as the clustering algorithm attempts to account for spurious keypoint locations. The effect of spurious keypoint locations can be mitigated by removing keypoints localised in areas of low keypoint density.

A voxelised 3D density histogram is created and aligned with the X, Y, Z dimensions in the canonical space. The density histogram ranges from the furthest outlying keypoints in each of the dimensions. The resolution of voxels in the density histogram is set to the width of a pixel in the source range image. The resulting density histogram is then convolved with a 3D Gaussian function with a sigma of 3 pixels in all dimensions. A valid keypoint mask is created by setting a threshold on the density histogram at a level of x keypoints per voxel. Keypoints which do not meet this criteria are discarded.

#### 6.1.1.3 Keypoint Clustering

Two clustering approaches were applied to the filtered keypoint locations to investigate the most suitable for creating 3D local interest points. The approaches investigated were the mean values calculated using Gaussian Mixture Models, GMM, and the cluster centres found using K-means clustering, both approaches were configured to assign 300 interest points per 3D model, see Figure 6.6. The GMM, models the data as a collection of overlapping Gaussian distributions of keypoints (Bishop, 2006). The GMM is calculated iteratively to simultaneously to establish the best fit of Gaussian distributions for all keypoint locations. The mean values of the GMM concentrates the resulting 3D local interest points around the centroid of all keypoint locations, as opposed to areas of high keypoint density. The K-means clustering approach was applied with 10 iterations to approach a convergence. K-means clustering divides the keypoint locations into clusters which tend towards having equal numbers of members. The K-means clustering approach successfully localises areas of high keypoint density, with cluster centres identifying 3D interest locations.

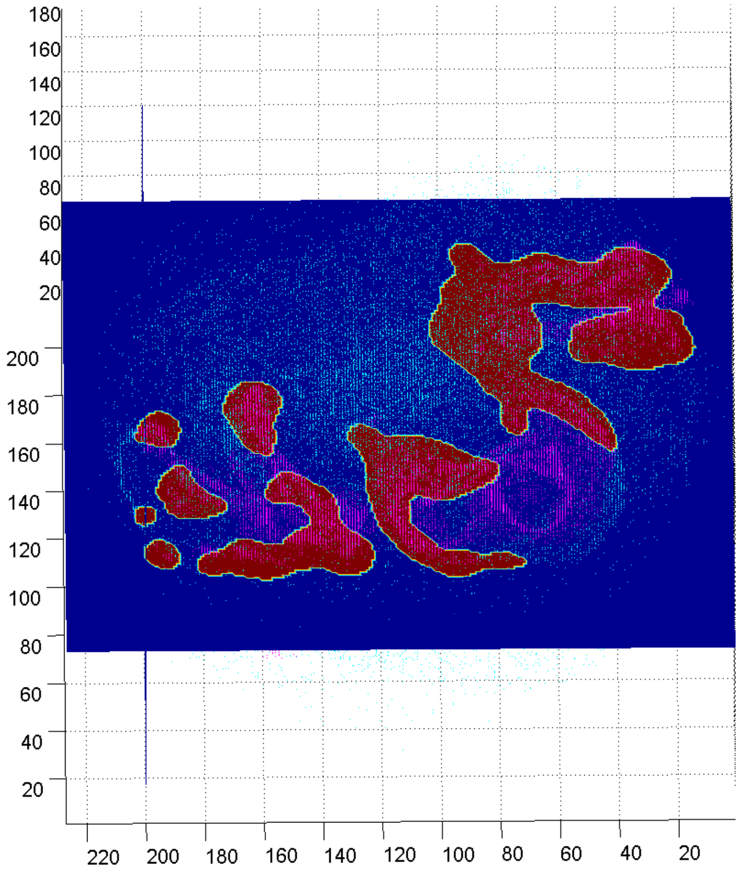


Figure 6.5: Slice through the 3D density histogram; red regions represent valid keypoint locations

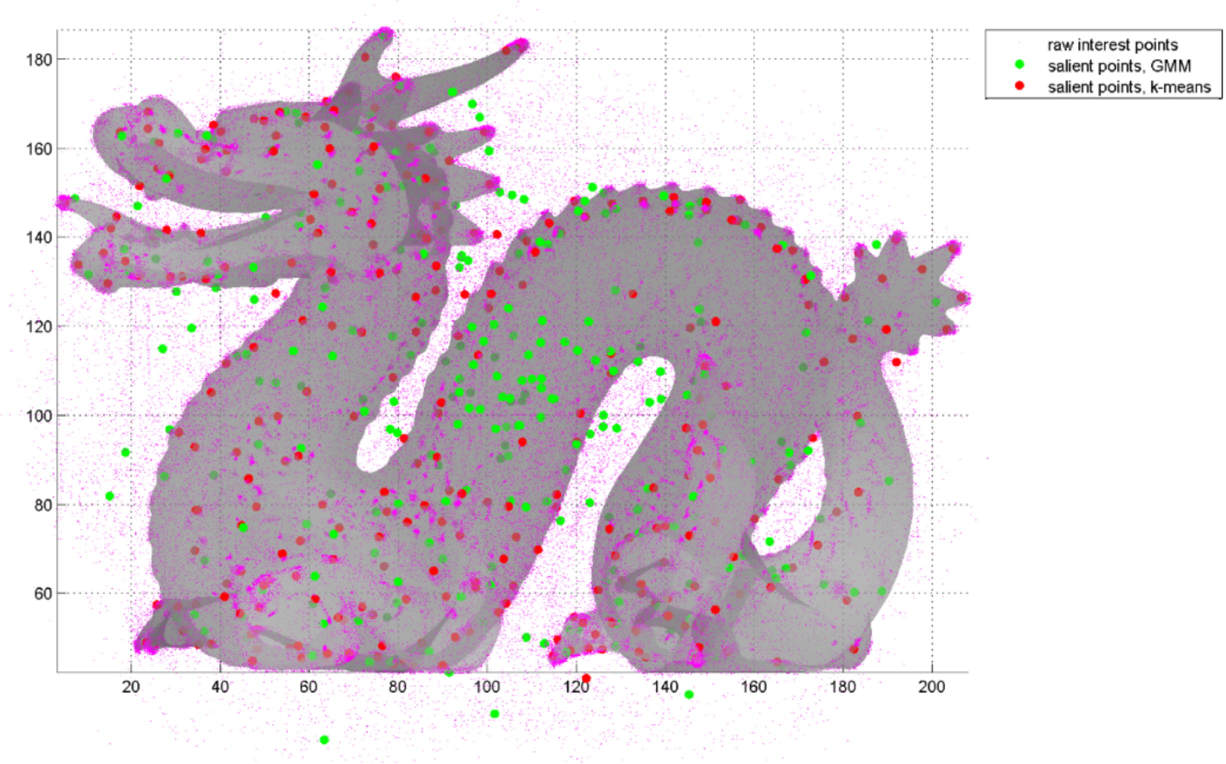


Figure 6.6: Cluster Centres

#### 6.1.1.4 Keypoint Scale

To ensure that only keypoints of the same scale will are included in the 3D interest point cluster, the scale of individual keypoints within clusters are then histogrammed and scale outliers are removed. Removing individual keypoints of differing scale ensures that in the descriptor characterisation stage the keypoint cluster will contain only similar feature descriptors, see Figure 6.7. Keypoints found at larger scales will tend to be fewer in number as there are fewer range pixels available to locate these, to account for this the histogram is weighted by the relative size of a pixel in the detected scale when compared to a pixel in the native scale.

### 6.1.2 Local Pose Estimation

Having established keypoint locations and collections, the pose of each keypoint member in a 3D Local Interest Point should be established with respect to the viewing angle at which it was observed. This section describes the method by which the 3D orientation of the individual keypoint within a range image is estimated. The 3D surface orientation is used to structure the extracted keypoints from individual range image examples in the resulting 3D local interest point descriptor.

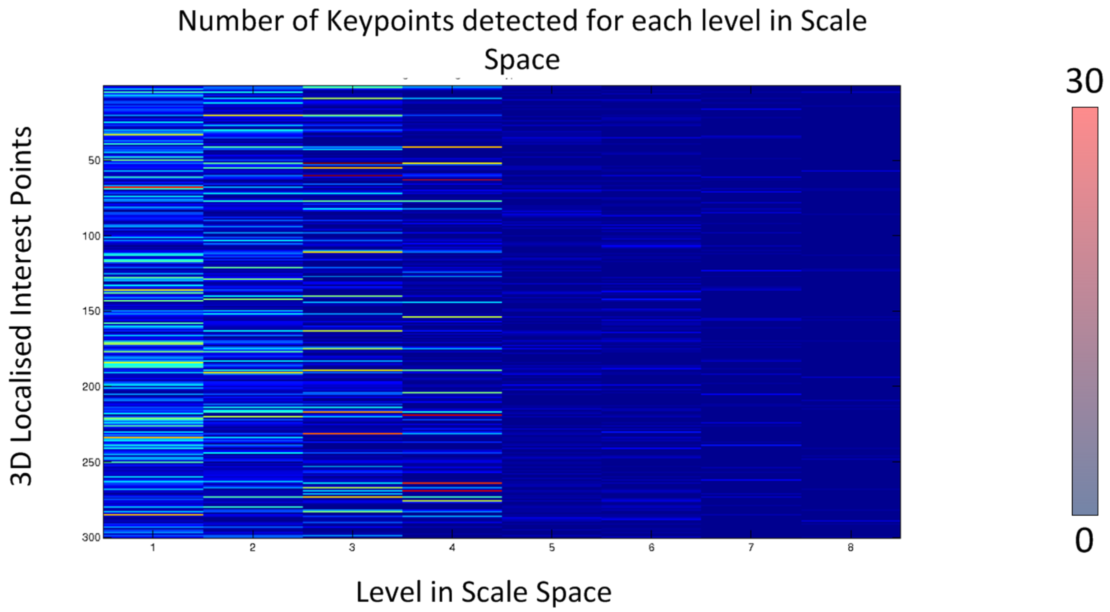


Figure 6.7: Weighted histogram

The remainder of this section describes the methodology.

#### 6.1.2.1 Defining the local rotation matrix

The surface normal for each keypoint example from each range image observation is defined from the first vector in the calculated local rotation matrix. This subsection describes the calculation of the local rotation matrix.

The local rotation matrix is estimated by considering the 10x10 pixel sample patch from the source range image as a point cloud, $Pts$, from which a covariance matrix can be calculated, see Equation 6.1. The covariance matrix for a set of points in $R^3$ forms a set of vectors which defines main axes along which the X, Y, Z points demonstrated the most variation, $eigv$. The length of these vectors is determined by the strength of variation in the given direction. Principal component analysis, PCA, offers a convenient method to separate the unit vectors from their weighted contributions. Taking eigen vectors of the covariance matrix, a matrix of orthogonal unit length vectors can be found which defines the orientation of the surface patch. The eigen vector, $eigv$, corresponding to the lowest eigen value, $\lambda$, defining the surface normal.

$$[\lambda, eigv] = \text{Eig}(\text{Cov}(Pts)) \tag{6.1}$$

$$M_{pose} = eigv \tag{6.2}$$

#### 6.1.2.2 Validation

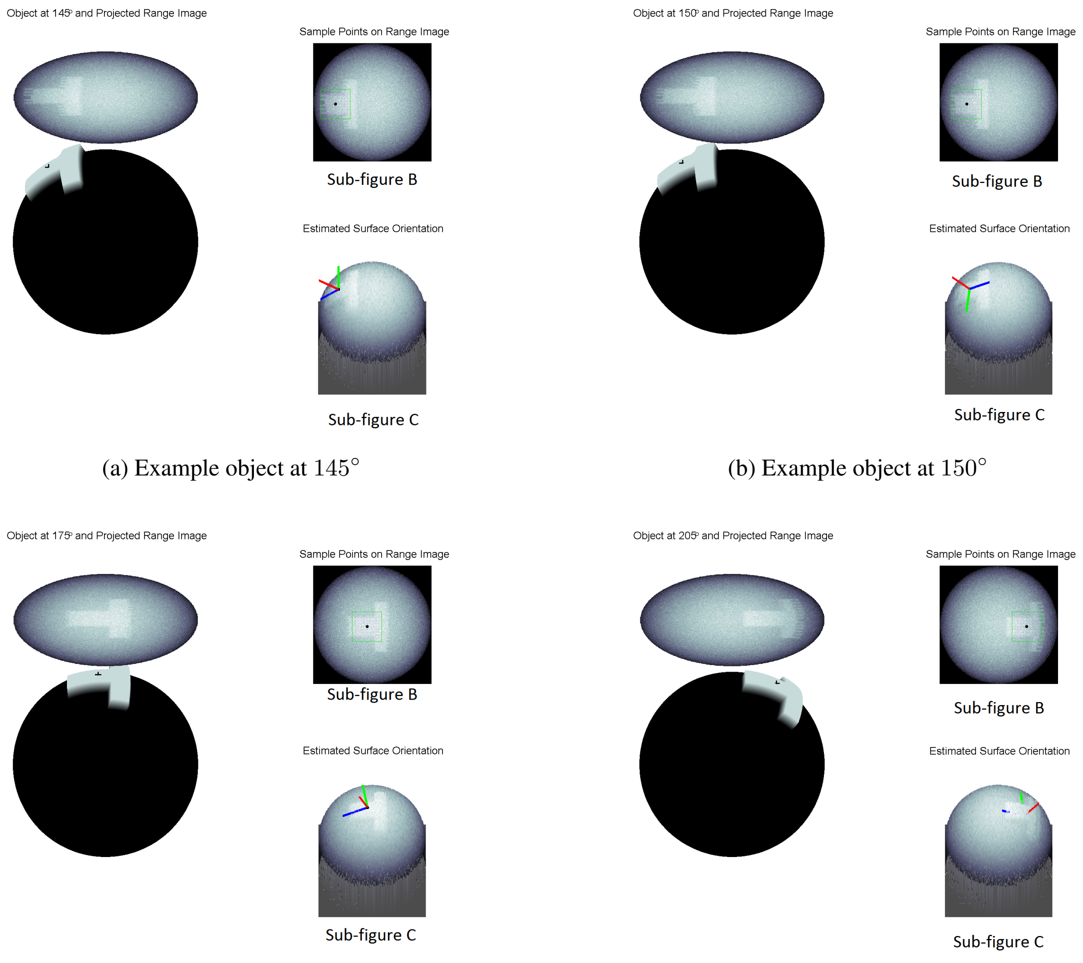
The collection of X, Y, Z points used to estimate the surface orientation clearly has an important influence on the resulting alignment estimation. The points used to create this estimation are extracted in a similar method as used for determining the surface orientation in SIFT. The sample points are taken from a 10x10 patch in the range image scale space pyramid at the scale with which the keypoint was detected by means of the corner detector.

An example of this is shown in Figure 6.8a, where a synthetic 3D object with a distinctive T shape has been projected on to a plane to create a range image, using a similar method as described in the Experimental Setup, Section 6.2. The generated synthetic range image then has Gaussian image noise of 10% of the full scale range deflection applied. Each Sub-figure B shows the sample patch outline in green, and 10x10 sample patch points in magenta. Each Sub-figure C shows the surface alignment axes in red, blue and green, respectively for each eigen vector. Only the surface normal has been rectified to show the consistency of alignment between range images of the object.

From tracking the keypoint location on the object, it was found that a sample patch sampling the object surface could be placed between ±40°. The calculated surface normal was then compared against the actuated surface to show the stability of this approach. The results are shown in Figure 6.9a, Figure 6.9b shows the root mean square error. This result shows an inaccuracy in estimated surface normal of ~10°using the approach outlined in this section.

### 6.1.3 3D Local Interest Point Characterisation

In this chapter keypoints have been extracted from all observations of the object. Their locations and pose at observation have been projected into a common canonical space for the object. This section addresses the central issue of this chapter: is there an approach for feature descriptor integration from multiple observations which will allow a common representation of a 3D local interest point on an object which will match to all keypoints members of the 3D local interest point? In order to address this issue, feature descriptors from a collection of keypoints describing the same object


Object at 145° and Projected Range Image
Sample Points on Range Image
Sub-figure B
Estimated Surface Orientation
Sub-figure C
Object at 150° and Projected Range Image
Sample Points on Range Image
Sub-figure B
Estimated Surface Orientation
Sub-figure C
Object at 175° and Projected Range Image
Sample Points on Range Image
Sub-figure B
Estimated Surface Orientation
Sub-figure C
Object at 205° and Projected Range Image
Sample Points on Range Image
Sub-figure B
Estimated Surface Orientation
Sub-figure C

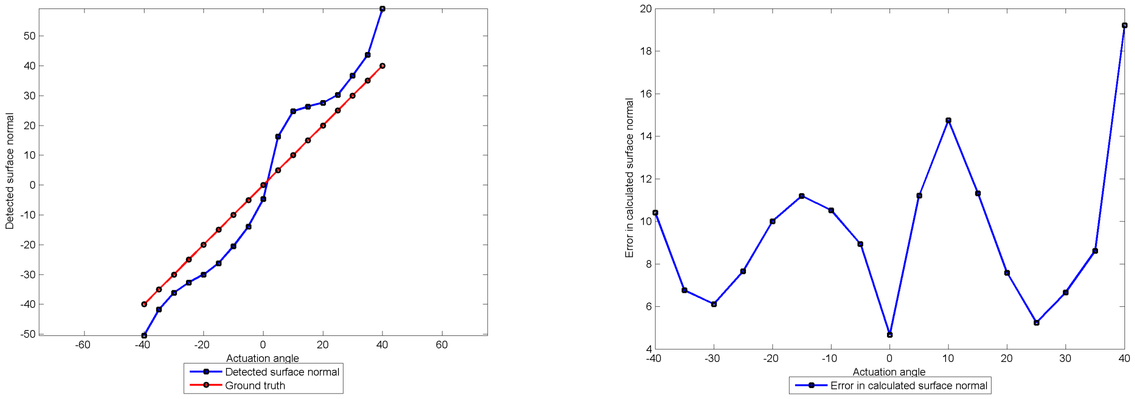
(a) Example object at 145°

(b) Example object at 150°

(c) Example object at 175°

(d) Example object at 205°

Figure 6.8: Pose estimation test data



(a) Calculated surface normal against actuation angle

(b) Error in calculated surface normal

Figure 6.9: Calculated Surface Normals

location must be integrated together to form the 3D local interest point characterisation, therefore there are 2 issues which this section addresses:

- Which keypoints should be grouped?
- What grouping approach should be applied?

In the previous section, the extracted keypoints from multiple observations have been clustered together based on their locations to form the 3D local interest points. However, within these clusters, keypoints from a variety of viewing angles exist. In previous chapters it has been shown, matching with the Euclidean distance, that feature descriptors only remain invariant to around $30^\circ$ of out-of-plane view point change. Therefore, selecting feature descriptors from keypoints based on the range of variation in observation angles should also factor when integrating feature descriptors, in addition to keypoint location and scale.

In this chapter sets of keypoints within 3D local interest points are created by thresholding membership of individual keypoints based on their observation angles. The variation of feature descriptors within the 3D local interest points can therefore be controlled by varying the inclusion of keypoints from surface observations with pose angles further from the surface normal. The formation of these sets of keypoints are shown in Figure 6.10, where Set(10) includes all keypoints which have been observed with a surface normal of $10^\circ$ or less in their range image . The Set(20) keypoints represents all keypoints collected where the local pose angle is $20^\circ$ or less in their range image; the set of keypoints Set(20) includes all keypoints within the subset Set(10). Equation 6.3, formalises the definition of the 3D local interest point sets, and shows the inclusion of keypoints, $key_{idx}$, in a set, given that the dot product between the keypoint surface normal, $norm_{key_{idx}}$, and the viewing axis, $view$, is less than or equal the set threshold.

$$view.norm_{key_{idx}} \begin{cases} \leq 0^\circ & set_0 = \{set_o, key_{idx}\} \\ \leq 10^\circ & set_{10} = \{set_{10,}key_{idx}\} \\ \leq 20^\circ & set_{20} = \{set_{20}, key_{idx}\} \\ \vdots & \vdots \\ \leq 90^\circ & set_{90} = \{set_{90}, key_{idx}\} \end{cases} \tag{6.3}$$

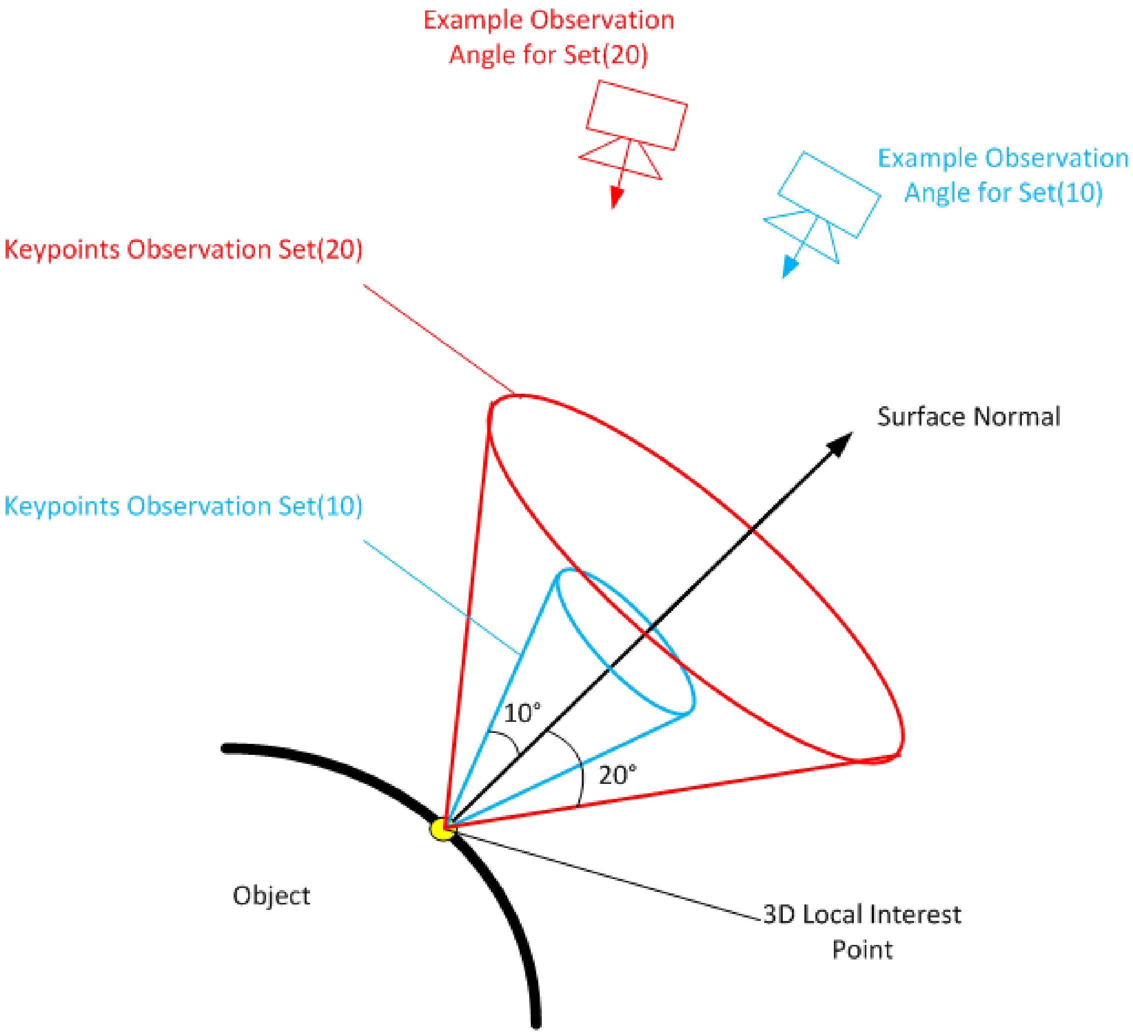

Example Observation Angle for Set(20)
Example Observation Angle for Set(10)
Keypoints Observation Set(20)
Surface Normal
Keypoints Observation Set(10)
10°
20°
Object
3D Local Interest Point

Figure 6.10: Keypoints from Multiple Observations

From the collection of keypoints in a set of angles, a characterisation of feature descriptors is performed to form the descriptor for the 3D local interest point. This descriptor comprises a mean feature descriptor for the set, and a measure of the variation within the set, $\mu_{set_i}$. To establish the optimum approach for characterising the variation within the set of feature descriptors, a number of approaches are investigated in this chapter. The investigated approaches are: Principal Component Analysis (Jolliffe, 2005), the Mahanalobis distance (Bishop, 2006), and Fishers Linear Discriminant (McLachlan and Wiley, 1992). Support Vector Machines (Cristianini and Shawe-Taylor, 2000), and the Earth Movers Distance (Zhang et al., 2007) were also considered as an approach for characterising the feature descriptor variation. However, due to the dimensionality of the feature descriptors and potential overlap in feature descriptor space between the area covered by differing features, the complexity for an SVM or EMD solution would be prohibitively expensive.

The remainder of this section covers the implementation of each of the proposed feature descriptor association approaches for creating a characterisation of 3D local interest points.

#### 6.1.3.1 Principal Component Analysis

Principal Component Analysis, PCA, is a method of reducing feature descriptor dimensionality without affecting match performance. This can be useful for saving storage space and increasing match efficiency (Bishop, 2006; Jolliffe, 2005). Typically PCA is used with SIFT to reduce the entire set of feature descriptors to a common principal component space in which the modes of variance of all feature descriptors is maximised (Ke and Sukthankar, 2004b). In this experiment a principal component space is created for each 3D interest point where the in-cluster descriptor variation is minimised using a PCA space trained on the in-custer descriptors. When matching, all feature descriptors are projected into this space and the Euclidean distance to the cluster feature descriptor mean is taken as the similarity measure. The projection of the feature descriptors on to the first 10 principal components of the in-cluster features can be seen in Figure 6.11. The in-cluster feature descriptors are shown in magenta, and the out-of-cluster are shown in green. These feature descriptors are taken from the cluster of keypoints shown in Figure 6.12, also with the in-class keypoints shown in magenta, and out-of-class in green. It can be seen by inspection that in the low dimensions of the PCA space, Figures 6.11a to 6.11h, the in-class cluster shown in magenta forms a point around the cluster mean, whereas the out-of-class cluster is not tightly formed and covers a larger area. Subfigure 6.11g, however shows the final dimension in the PCA space, where the within class features exhibit the most variation; here the magenta points can be seen to spread out through the same area covered by the out-of-class features. This result shows that feature descriptors projected into a PCA space based on their variance are well clustered when compared with all other feature descriptors projected into the same space; in this instance it is possible to form a characterisation for the 3D local interest points which fits well to all member keypoints using PCA.

PCA is used to create a set of projection vectors defining the reduced feature space for the principal components of the cluster feature descriptors, see Equation 6.4. The mean of the feature descriptors belonging to the 3D interest point from this space are also stored, see Equation 6.6. The PCA eigen vectors are used to project query descriptors into the reduced feature descriptor space of the in-class descriptors, Equation 6.5, and compare the distance between the query descriptor and the mean of the model descriptors using a weighted Euclidean distance based on the eigen values, Equation 6.9.

$$[w_{in}, \lambda_{in}] = \text{Eig}(\text{Cov}\,(desc_{in})) \tag{6.4}$$

$$de\hat{s}c_{in} = desc_{in}.w_{in}\,(1:n) \tag{6.5}$$

$$\mu_{in} = \text{mean}\left(de\hat{s}c_{in}\right) \tag{6.6}$$

$$de\hat{s}c = desc.w_{in}\,(1:n) \tag{6.7}$$

$$dist = \sqrt{\text{sum}\left(\left(de\hat{s}c - \mu_{in}\right)^2\right)} \tag{6.8}$$

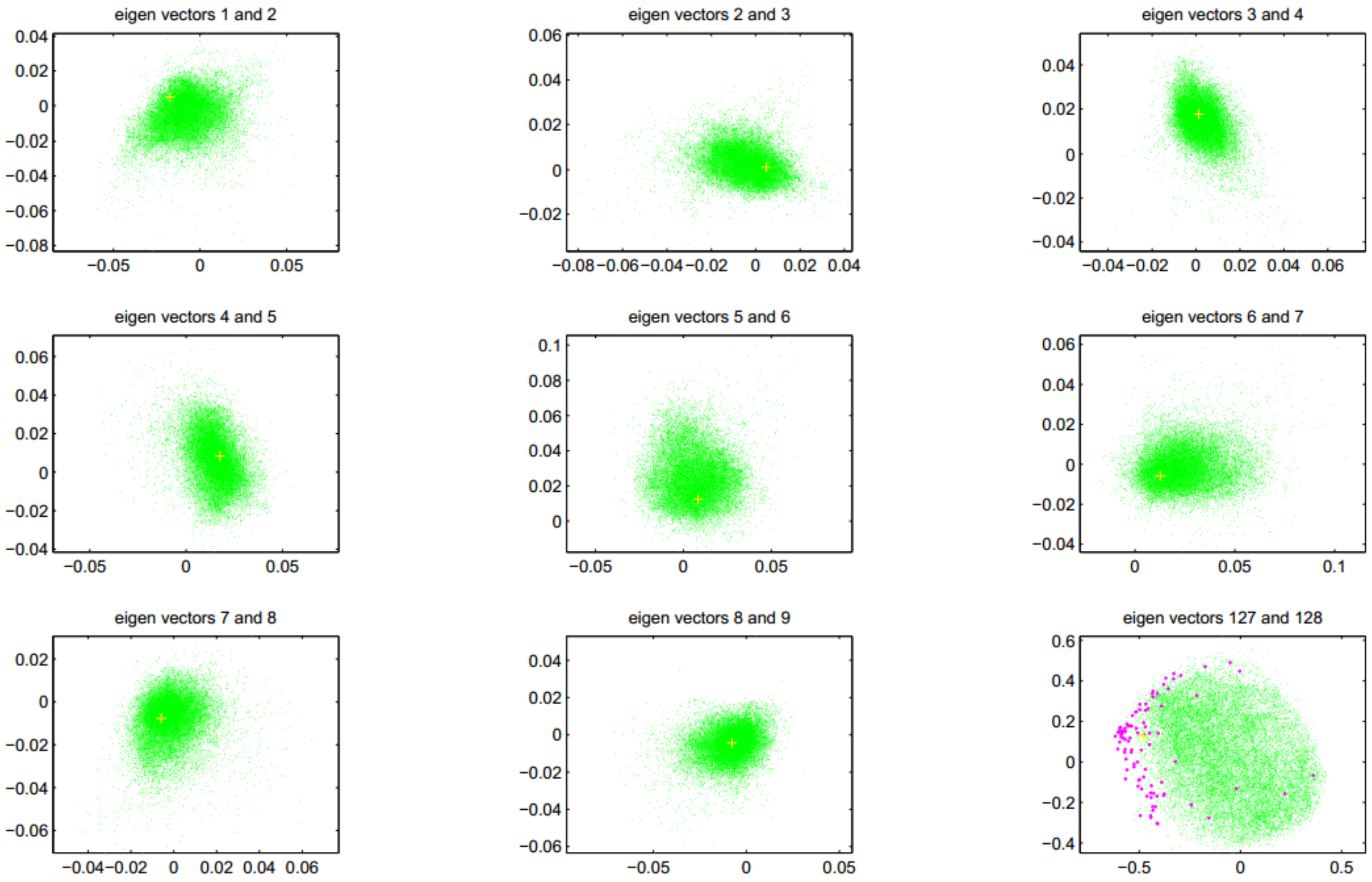


Figure 6.11: Principal components of feature descriptors for different dimensions in Eigenvector space.

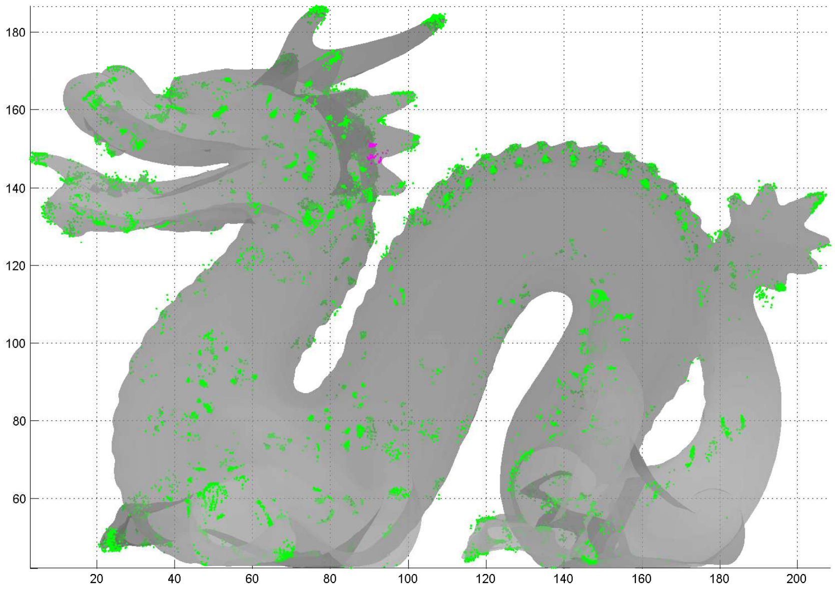


Figure 6.12: Single feature descriptor cluster, shown in magenta and located on the back of the dragon head

$$wghtdist = \text{diag}\left(\sqrt{\left(\hat{desc} - \mu_{in}\right)^T \lambda(1:n)^{-1}\left(\hat{desc} - \mu_{in}\right)}\right) \tag{6.9}$$

#### 6.1.3.2 Mahalanobis Distance

The covariance matrix created for each cluster of feature descriptors can be used directly as a measure of descriptor similarity. The Mahalanobis distance is a multi-variant analysis approach which decreases the weighting for the distance in the dimensions of the feature descriptor which have high in-class variation (McLachlan and Wiley, 1992). It is hypothesised that in the Mahalanobis space the distance within cluster features descriptors will show a smaller variation when compared with non-member descriptors.

$$\mu = \text{median}\left(desc_{in}\right) \tag{6.10}$$

$$\sum = \text{cov}\left(desc_{in}\right) \tag{6.11}$$

$$dist = \text{diag}\left(\sqrt{\left(desc - \mu\right)^T \Sigma^{-1} \left(desc - \mu\right)}\right) \tag{6.12}$$

#### 6.1.3.3 Fishers Linear Discriminant

Fishers Linear Discriminant, FLD, finds a vector on which to project the feature descriptors such that the variance between non-members is maximised and the variance between members is minimise. The FLD distance can be found from the ratio of the covariance of non-member feature descriptors, to the covariance of member feature descriptors, see Equation 6.16. Projecting the feature descriptors onto the FLD space shows a good separation of the within cluster feature descriptors against out-of-cluster feature descriptors. Figure 6.13a shows the distribution of the within cluster and out-of-cluster feature descriptors for a 3D interest point localised on the Dragon model, where the green curve shows the distribution of out-of-class interest points, and the magenta curve shows the distribution of in-class interest points. The in-cluster keypoints for the 3D interest point used is shown in magenta on dragon model, see Figure 6.13b.

$$\mu_i = \text{mean}\left(desc_{in(i)}\right) \tag{6.13}$$

$$\Sigma_b = \text{cov}(\mu) \tag{6.14}$$

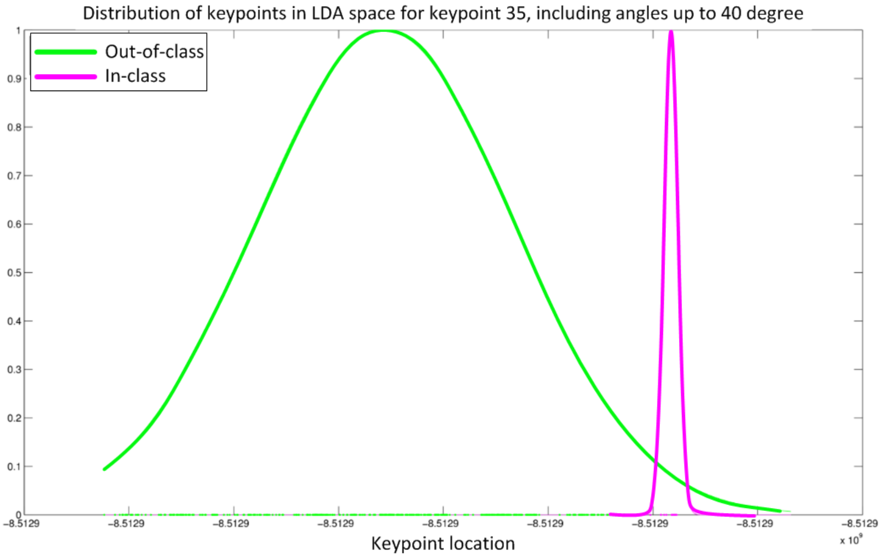
$$\sum = \text{cov}\left(desc_{in}\right) \tag{6.15}$$

$$dist = \frac{desc^T \Sigma_b desc}{desc^T \Sigma desc} \tag{6.16}$$

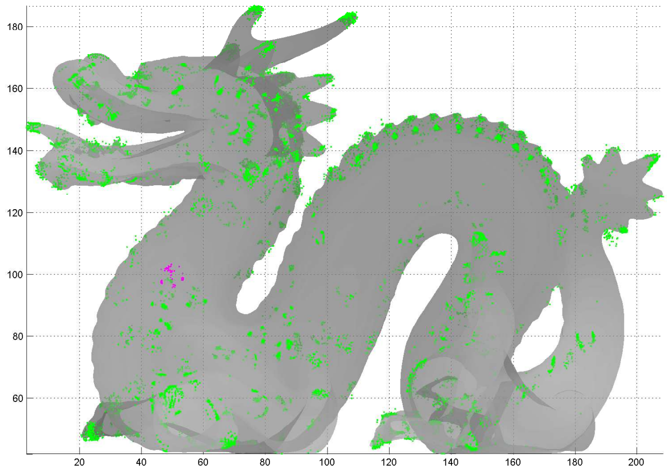
### 6.1.4 Methodology Summary

This section summarises the processing approach proposed in this chapter for creating a keypoint characterisation across a range of observations.

1. **Keypoint extraction**, keypoints are extracted from individual range images using 2.5D SIFT.



(a) Distribution of keypoints in FLD space



(b) Keypoint 35 on Dragon model

Figure 6.13: Keypoint and descriptors in FLD space

Keypoints are characterised as position in x, y and range, scale and SIFT feature descriptor. These are single observation keypoints and can be extracted using any approach from Chapter 5, in this chapter the extracted keypoints are standard 2.5D SIFT applied to the range image.

2. **Pose estimation**, extracted keypoints have their local pose estimated from an 10x10 pixel patch taken from the range image surrounding the keypoint location. The pose of the local keypoints is stored as the keypoint observation angle..

3. **Projection of keypoints into the canonical space**, all keypoints collected from all range images are projected in to a common, 'canonical space'. In this chapter, the projection of all keypoints into the canonical space is achieved using the initial transformation applied to the 3D object for creating the synthetic range image, Section 6.2.1.

4. **3D interest point filtering and clustering**, the 3D locations of all keypoints collected form all range images represented in the canonical space form clusters around repeatable 3D locations on the object surface. However, many 3D localised keypoints do not associate with any given cluster, these keypoints are removed through 3D density filtering, the remaining keypoints are associated together into clusters.

5. **Keypoint association**, K-means clustering is applied to the filtered keypoint locations to form groups of keypoints surrounding 3D interest points.

6. **3D interest point characterisation**, The groups of keypoints are used to form examples of different expressions of feature descriptor possible for a given 3D interest point on the object. The mean value of the feature descriptors and the variation present in the group is recorded.

7. **Matching**, Query keypoints are matched to the database of collected 3D interest points using either PCA, FLD or Mahanalobis distance, the effectiveness of each of the proposed matching approaches in investigated in Section 6.3.

## 6.2 Experimental Setup

The aim of this experiment is to establish which methodology proposed in Section 6.1.3 is most appropriate for creating a characterisation of a 3D local interest point, and using each approach what range of keypoint observations may be matched. In order to evaluate the performance of the proposed characterisations, outlined in Section 6.1.3, each approach must be used to form a characterisation for every 3D local interest point and evaluate which keypoints in the database belong to this 3D local interest point. As multiple keypoints in the database should all match to the same 3D local interest point, the experimental configuration outlined in Chapter 4 cannot be directly applied. Furthermore, the models from Chapter 4 demonstrate a out-of-plane orientation change in only the yaw axis. This limited range of object observations will result in a limited range of keypoints which are viewed across a significant portion of the view sphere. To address these issues this section decomposes the problem into two parts: the formation of a database with examples from a more complete view sphere, and an approach for evaluating the performance of matching descriptors extracted from example range images to their corresponding 3D local interest points.

### 6.2.1 Synthetic Range Image Database

The test data used in Chapter 4, contains range images which exhibit out-of-plane observations changes in the yaw axis only. This limitation in range image observations of an object reduces the range of angles at which a keypoint may be observed, therefore reducing the number of keypoints which may exist in each set of keypoints for a 3D local interest point, Section 6.1.3. To create a complete set of observations for an object, a database of synthetic range image examples with out-of-plane observation changes in both the yaw and pitch axes were created from models in the Stanford 3D Scanning Repository. The orientation of the roll axis was fixed for all observations. The synthetic range images are created at 10 degree intervals in azimuth and elevation surrounding the object. Figure 6.14 shows the Dragon model with camera locations surrounding the object on a sampling sphere. To reduce the level of variation in the extracted feature descriptors no orientations change in the roll axis was applied and feature extraction was applied with no in-plane orientation invariance. This approach has been shown by Agrawal et al to increase the performance of SIFT descriptors in situations similar to those created here where objects are consistently viewed with the same in-plane orientation, or roll axis (Agrawal et al., 2008).

To create the synthetic range images a rotation is applied to the 3D model to align it with the axis of viewing observation. The viewing observation axis for each observation instance is the vector which joins an observation point, shown in magenta in Figure 6.14, to the centre of the object.

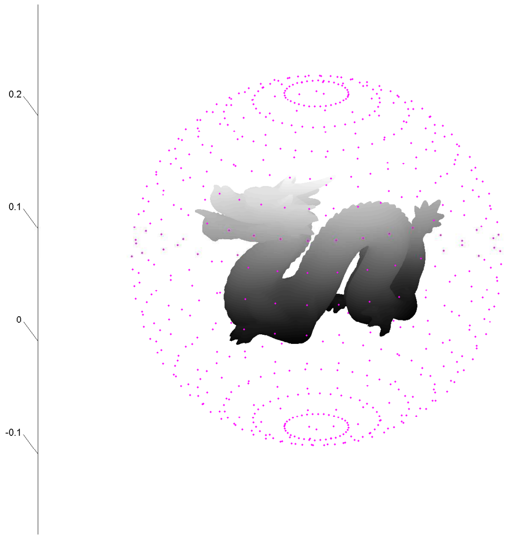


Figure 6.14: Dragon model with the locations of each view point shown as a magenta point

Hidden point removal is then applied to remove obscured vertices (Katz et al., 2007). The resulting point cloud is then resampled to create the synthetic range image at a resolution of 2000x2000 pixels. The resultant range images are then resized to 10% of their original range image size before extracting keypoints. All models in the Stanford 3D Scanning Repository have upwards of 5 million triangles.

### 6.2.2 Performance Evaluation

To evaluate the ability of the approaches outlined in Section 6.1.3 to characterise the variation within collections of descriptors in a 3D local interest point, *precision recall curves* are used. Precision recall curves are a parametric plot of the *precision* of a proposed approach against the approaches *recall*. In this experiment the precision is defined as the ratio of number of keypoints which are successfully matched to the given 3D local interest point, to the number of keypoints found. The recall is defined as the ratio of number of keypoints which are successfully matched to the given 3D local interest point, to the total number of keypoints which should match to the given 3D local interest point. In the evaluation methodology outlined here, all keypoints are ranked based on the distance to the cluster mean in the learned descriptor space.

Precision recall curves are formed by progressing through the ranked list of keypoints, $key(idx)$, and evaluating whether each entry increases the precision and recall values. Each entry in the ranked list is evaluated using the equations for precision and recall outlined in Equations 6.19 and 6.20, where $i$ is the index for each entry in the ranked list $idx$, $key$ is the database of keypoints and $set$ is the subset of keypoints that belong to the 3D local interest point. The precision recall curves in this chapter are plotted as 1-precision against recall, thereby giving a similar appearance to ROC curves used in previous chapters. Ideal performance is demonstrated at the point (0,1), where 1-precision is 0, indicating that all keypoints found belong to the set, and recall is 1, indicating that all keypoints which belong to the set were found. In this section the Euclidean distance measure is used as the base-line ranking for characterising the 3D interest points, Equation 6.17. All results are shown as a comparison between the proposed characterisation approach and the Euclidean distance.

$$dist = \sqrt{\mathrm{sum}\left((desc - \mu_{set_i})^2\right)} \tag{6.17}$$

$$idx = \mathrm{sort}\,(dist) \tag{6.18}$$

$$precision_i = \frac{\sum_{n=0}^{i} key(idx(n)) \cap set}{i} \tag{6.19}$$

$$recall_i = \frac{\sum_{n=0}^{i} key(idx(n)) \cap set}{\sum set} \tag{6.20}$$

Figure 6.16 shows different precision recall curves for keypoint sets formed using increasing viewing angles from the surface normal, and the Euclidean distance. Each curve is formed by varying the value of $i$ for a given inclusion angle set, $set$. The precision recall curves for sets of inclusion angles remove all keypoints in the database whose angle is greater than the set maximum angle. It can be seen that there is no curve for feature descriptors collected at a viewing angle closer than 0, 10, or 20 degrees from the surface normal as features detected at these angles are not present in the majority of 3D interest points, Figure 6.15 shows the distribution of keypoints against observation angle. The distribution of keypoint poses indicate that the majority of the keypoints are detected around occlusion boundaries. In Figure 6.16, the performance of including keypoints with surface normals greater than the viewing angle drops off steadily, with around 60 degrees inclusions representing the boundary of indecision. In the next section the performance of various distance metrics learned from cluster membership are compared for a range of observation angle

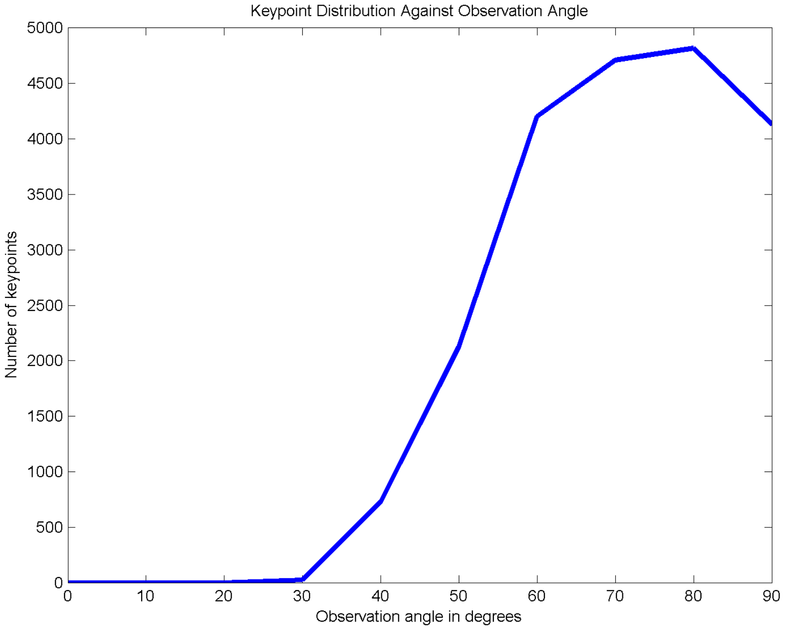


Figure 6.15: Distribution of keypoints against observation angle

inclusions.

## 6.3 Results and Analysis

This section presents the results of 3D interest point characterisation using the approach outlined in Section, under the performance criteria described in Section 6.1.3. . All precision recall curves for learned 3D interest point characterisation approaches are compared against the performance of the Euclidean distance in the subfigures below the performance for all angles.

### 6.3.1 PCA

Figure 6.17 shows the precision recall curve for PCA characterised 3D local interest points. Grouping descriptors based on the PCA space for each 3D interest point out performs the grouping based on the Euclidean distance. Descriptors maintain a stronger similarity at 60°, with better performance for all investigated angles.

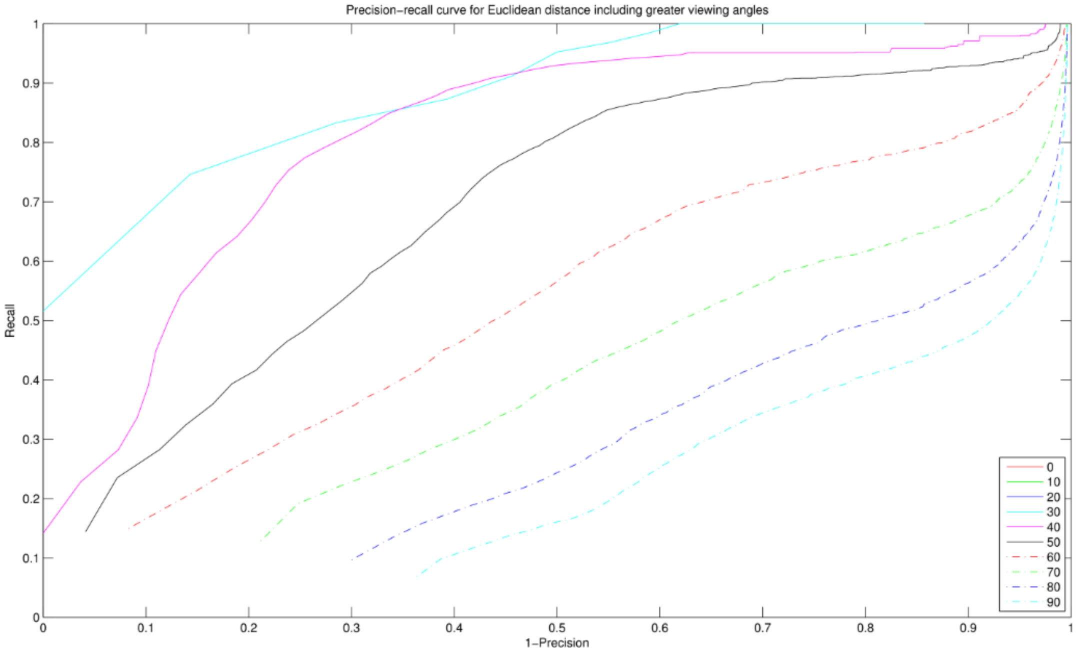


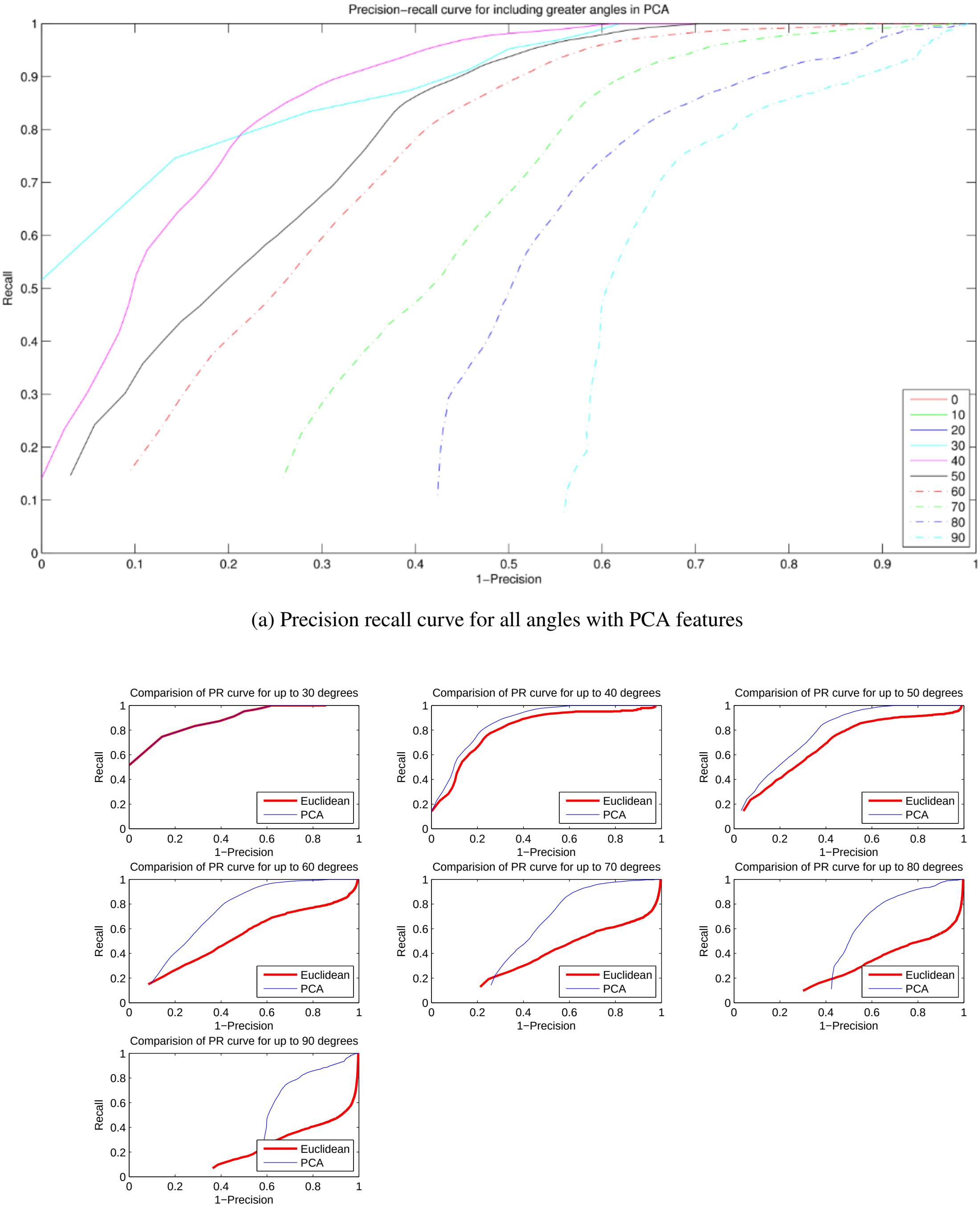
Figure 6.16: Precision recall curve for features clustered with the Euclidean distance

### 6.3.2 Mahalanobis distance

The Mahalanobis distance uses the same covariance matrix as the PCA approach for each view angle set. The distance of feature descriptors in the covariance space is compared. This approach shows a dramatic decrease in performance for angles up to 50°; however it maintains a higher performance when compared to the Euclidean distance in sets containing keypoints observed with up to 90° of deviation in out-of-plane rotation from the surface normal.

### 6.3.3 FLD

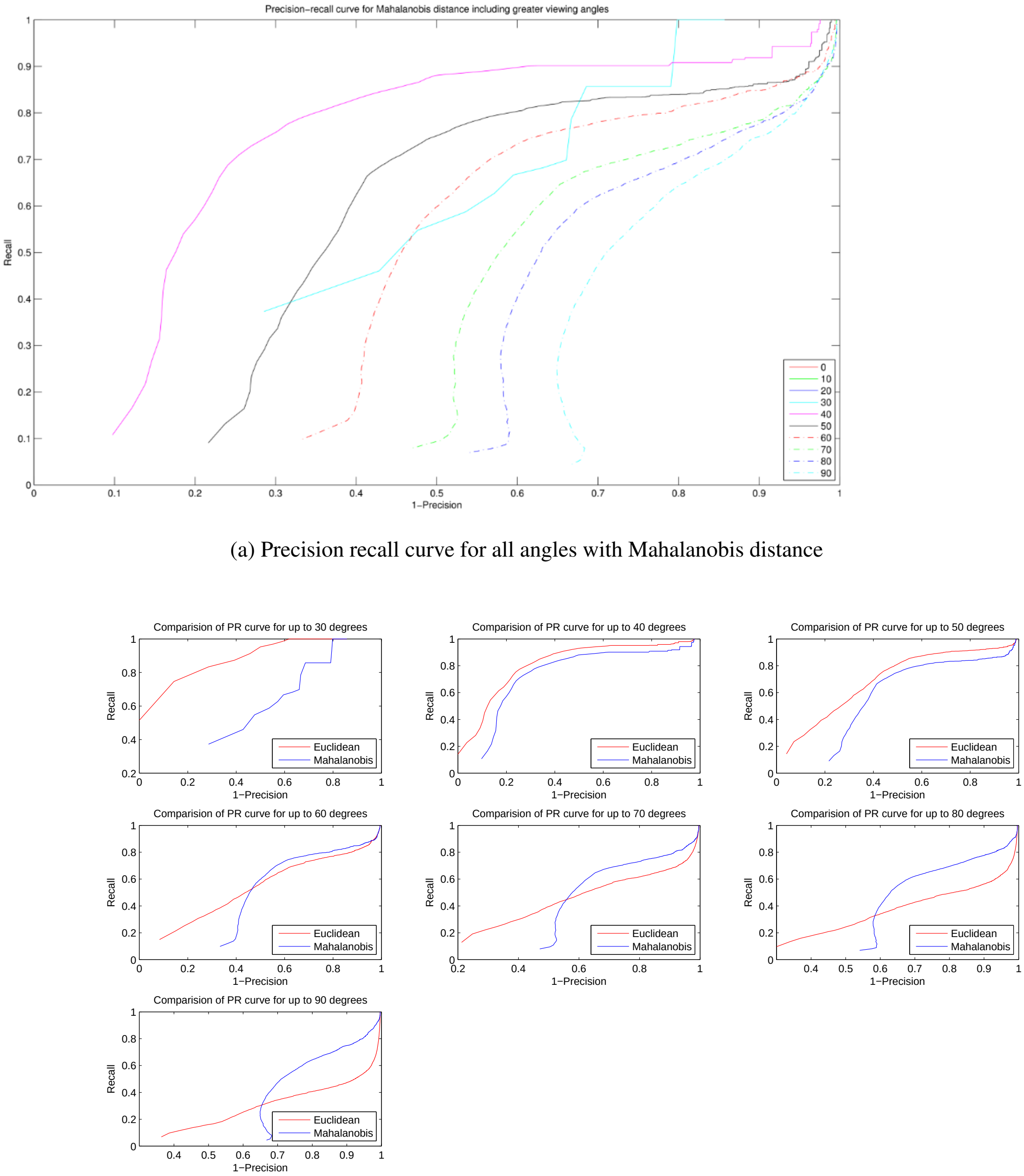
Figure 6.19, shows the precision recall curves for distances measured along the Fisher's Linear Discriminant. This measure performances the poorest out of all investigated characterisation approaches and worse than grouping based on the Euclidean distance, with precision for angles greater than 30° dropping to 0. The FLD distance finds a single vector along which the distances between model and query can be measured. The space occupied by within cluster descriptors is multidimen-



(a) Precision recall curve for all angles with PCA features

(b) Precision recall curve for all angles with PCA

Figure 6.17: Precision recall curves for PCA learned features



(a) Precision recall curve for all angles with Mahalanobis distance

(b) Comparison with Euclidean distance

Figure 6.18: Precision recall curve for Mahalanobis distance

sional and therefore this approach is inappropriate for characterising the 3D interest points.

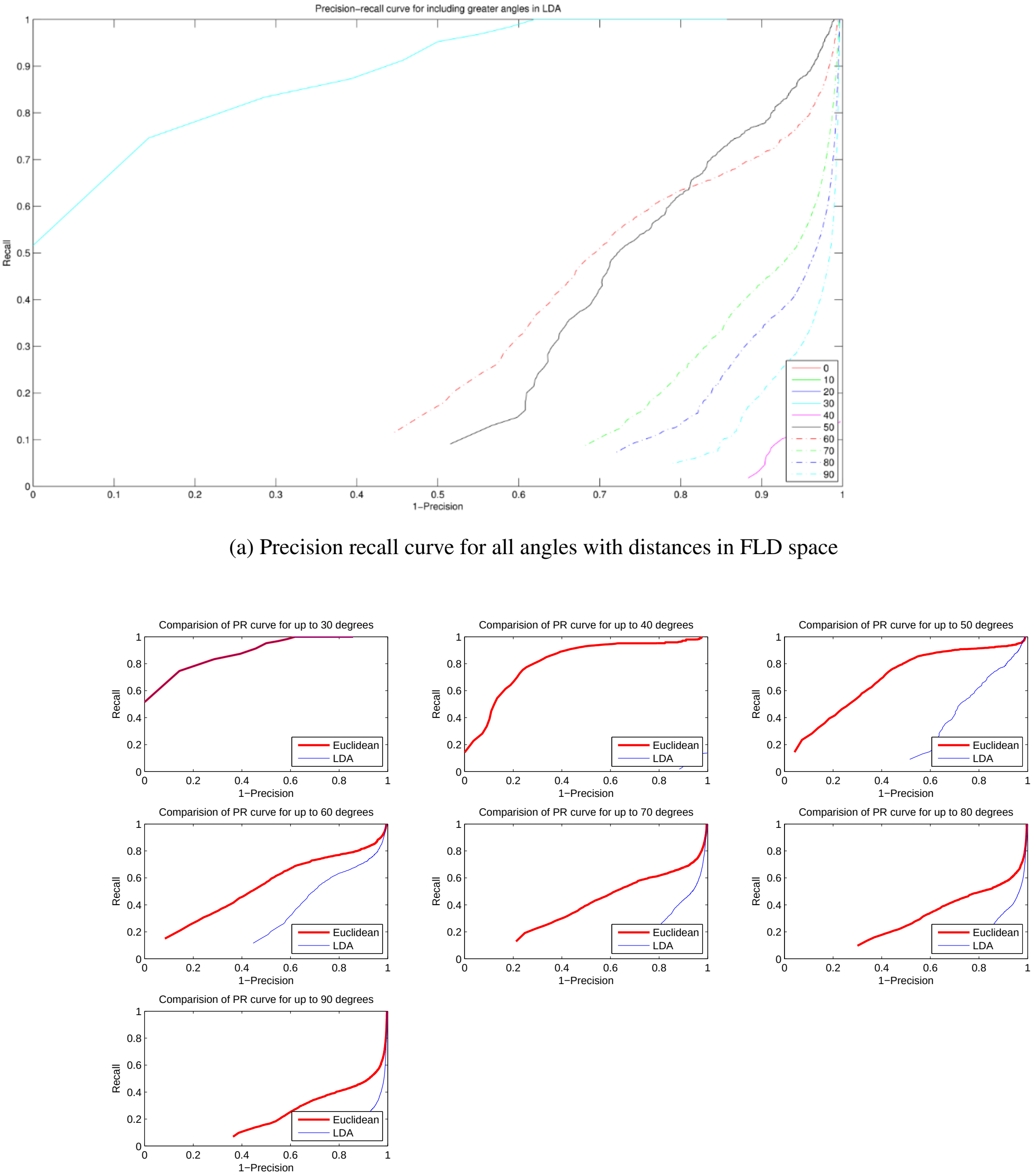
## 6.4 Summary and Conclusions

This chapter has presented a methodology for creating and characterising view independent 3D local interest points on a 3D free form object. The collection of view independent 3D local interest points are created from a collection of view dependant SIFT keypoints extracted from range images. An attempt to statistical characterise the underlying local structure of the object surface at these 3D local interest points has been made from the collection of SIFT descriptors taken from all views, such that the likelihood of matching a local feature descriptor from any given view to the 3D local interest point to which it belongs is increased. The characterisation of keypoints in this manner allows for the integration of multiple keypoint descriptors to form a 'typical' descriptor for a 3D local interest point and to give an indication of the variation in feature descriptor space which the descriptor may express.

From inspection of the 3D clusters formed by the 3D local interest point locator, keypoints characterising protrusions can be seen to be well represented. However, keypoints characterising surface textures or macro surface features such as large scale inflections of the object surface are not well represented. This chapter also found that using the current keypoint localisation approach in range images there are few keypoints which are detected in observations where the local pose is close to the surface normal.

Another limitation of the approach presented in this chapter is that the number of 3D local interest points must be specified in advance, this number is expected to vary between objects depending on the characteristics of the surface. As the number of 3D local interest points must be specified in advance, online learning of objects through a continuous exploration of the observation space is not possible. This requirement limits the applications of this approach to instances where the full observation space of an object is available for creating the initial keypoint database. Furthermore, the approaches presented for learning of variation in the feature descriptor space requires that all feature descriptors are present for a learning iteration; therefore to perform online learning and satisfy this requirement, all keypoints for a 3D local interest point must be stored.

The next chapter addresses the main issues identified with the approach detailed in this chapter.



(a) Precision recall curve for all angles with distances in FLD space

(b) Comparison with Euclidean distance

Figure 6.19: Precision recall curve for distance in FLD space

Keypoint localisation is addressed by keeping all keypoints from all views, such that the 3D local interest points are not required to be calculated. Furthermore, this chapter found that keypoints localised in the range image tend to be localised around regions where the local pose is greater than $40°$; keypoints in the next chapter may be localised in intensity images, therefore increasing the range of poses over which keypoints may be observed. The increase of variation of feature descriptors within sets with the increase of the surface pose inclusion criteria, is addressed by compartmentalising the feature view space into a series of canonical views. In this chapter PCA was found to give the highest performance of all characterisation approaches investigated. However, the performance compared to the Euclidean was only marginal. Therefore, the feature learning approach adopted in the next chapter will be performed by updating the view compartments with keypoints which best relate to the compartment criteria.

# Chapter 7

# View Compartmentalised Keypoints

Previous chapters have attempted to improve the performance of feature descriptors when observed under changes in view point using either an invariant representations for single observation of keypoints or by combining multiple observations to form a single characterisation of a keypoint location. This chapter joins these two concepts by creating a keypoint composed of multiple observations arranged in a compartmentalised keypoint. View compartmentalised keypoints have an additional weighting component which allows a confidence measure to be assigned to a compartment, based on the similarity of the viewing angle to the ideal observation angle for the compartment. A novel keypoint matching stage is introduced to incorporate the keypoint weighting component.

## 7.1 Objectives

The surface description methods used to formulate the feature descriptors in Chapter 5 have relied on a single observation of an object with which to create a view invariant descriptor. These approaches were found to only be capable of characterising a keypoint location on the object surface within a tolerance of ±30 degrees change in viewing angle from the original observation. Therefore, using the standard local feature extraction approach, without knowledge of the complete 3D structure of the object and only a single observation, it is not possible to achieve a fully pose invariant local description which will hold for all viewing angles. Chapter 6 approached the problem using a characterisation, modelling the descriptor space which a specific surface location can express within a keypoint. The characterisation of the descriptor space for each keypoint was based on observations from multiple viewing angles. Chapter 6 found that feature descriptors for the view sphere surrounding a keypoint demonstrated a large degree of variation depending on the viewing angle of the observer. This result shows that given different observations through the view sphere, keypoints may display a range of differing descriptor expressions; for example, the orthonormal view of a keypoint localised on a surface protrusion will contain many occlusion boundaries and other diagnostic information which will be characteristic of the keypoint location. Although appearance information from the orthonormal view of a keypoint is diagnostic for characterising the keypoint location, a keypoint may be equally well characterised by any one of its many silhouette images, no single *canonical view* for a keypoint exists. Additionally large differences between these multiple *canonical views* will cause each to fail to when matching to the others. Merging the descriptors from each of the possible canonical views was found in, Chapter 6, to reduce the distinctiveness of the resultant keypoint. Clearly a pose invariant feature descriptor must be capable of characterising all these expressions of the same keypoint independently in addition to structuring the observed data for matching purposes.

The approach taken in this chapter is to use the structure of the object surface and local pose information from the range image to associate keypoints together in a training phase. The keypoint training and association phases create descriptors which characterise the different view-based modes which a local feature can express: the orthonormal view, and a set of silhouette views covering a viewing hemisphere. The feature descriptors for each canonical view of a keypoint are then concatenated together and stored in a *view compartmentalised descriptor* with an associated *compartment weighting value* for each keypoint compartment. The view compartmentalised descriptor characterises the surface as measured from the given observation and the compartment weighting value describes the confidence which can be placed in the associated feature descriptor match for a compartment. Given a single keypoint observation, the extracted feature descriptor is initially assigned to every compartment and a weighting is calculated for the assignment to each compartment based on the 3D pose of the keypoint.

The 3D pose of a keypoint is calculated using the 3D pose estimation approach outlined in Chapter 6. This approach considers the range image as a 3D point cloud and calculates the principal components of the X, Y and Z points in a local support region surrounding the keypoint. In Chapter

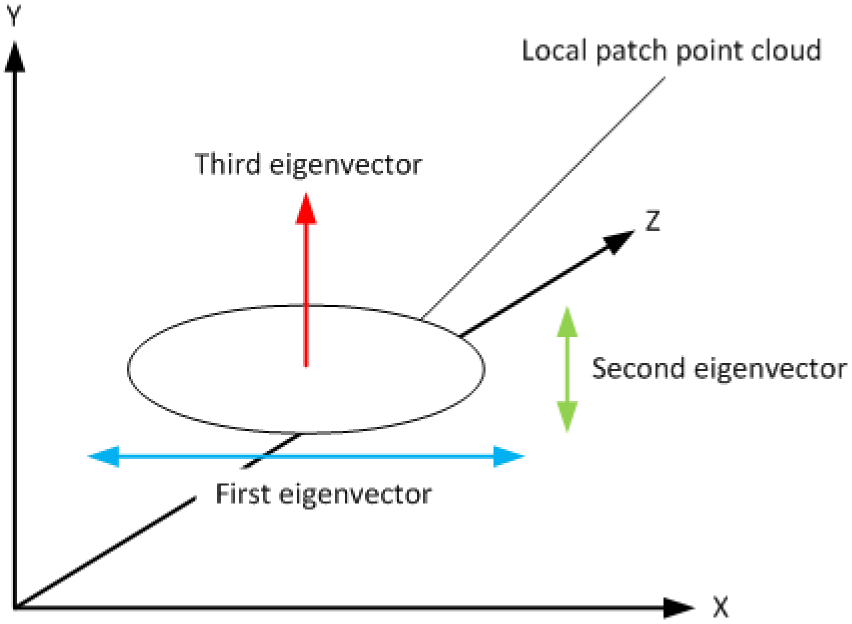


Figure 7.1: Local patch 3D orientation established through PCA, where the first eigenvector corresponds to the eigenvector with the largest associated eigenvalue

6, the Eigenvector corresponding to the lowest eigenvalue of the point cloud covariance matrix is used as an estimation of the surface normal of the keypoint. However, this chapter extends this approach by additionally using the second eigenvector as an estimate of 3D in-plane orientation of the local patch, see Figure 7.1.

As the compartment weighting value is to be included in the keypoint matching process the standard Euclidean distance metric cannot be directly applied for matching keypoints. To incorporate the compartment weighting value in the keypoint matching process, a weighted Euclidean distance approach is outlined and validated. Additionally, keypoints can be trained across a range of observations, to achieve this, an approach for updating existing keypoints in a database as new observations of the same keypoints from different poses become available is presented.

In this chapter a methodology is outlined for the association of the observed feature descriptors with their respective compartments in the view compartmentalised keypoint, and the implementation of the weighting function used for matching. This chapter contributes an approach for structuring keypoints based on 3D pose and for matching keypoints based on a confidence score for target and query view compartments. This chapter presents both the methodology and the experimental results, using the experimental design detailed in Chapter 4. The remainder of this chapter is organised as follows: Section 7.2.1 outlines the approach for determining the invariant local 3D pose space of the keypoint; Section 7.2.2 details the approach for establishing the compartment weighting value; Section 7.3 details the weighted Euclidean distance function for matching view compartmentalised descriptors; Section 7.5 details the results of the view compartmentalised descriptor in comparison with other local feature descriptors; Section 7.6 concludes the chapter with a summary and discussion of the approach and results.

## 7.2 View Compartmentalised Keypoint

The view compartmentalised keypoint is an approach for dividing a viewing hemisphere of an individual keypoint into a series of views, such that information regarding the appearance of the keypoint from each view can be encoded. The approach to formulating view compartmentalised keypoints was inspired by view bubbles, introduced by Peters and Zitova (Peters et al., 2002), where views of an object are characterised based on their similarity across the viewing space. However, in this application not all views are immediately available, therefore a characterisation of the difference between these is not possible; as a solution to this, the view compartmentalised keypoints store information from the views which have been observed in *view compartments* within the keypoint. In this approach each keypoint comprises a location, scale, 3D pose represented as a directional cosine matrix, and a series of compartments each with a SIFT feature descriptor and a weighting value, describing the confidence in each keypoint compartment for the viewing angles where the keypoint has been observed. The arrangement of compartments is aligned with the directional cosine matrix, defining the 3D orientation of the surface at the keypoint location. This alignment step aligns the *keypoint viewing hemisphere* with the keypoint surface normal and a calculated 3D in-plane orientation of the keypoint.

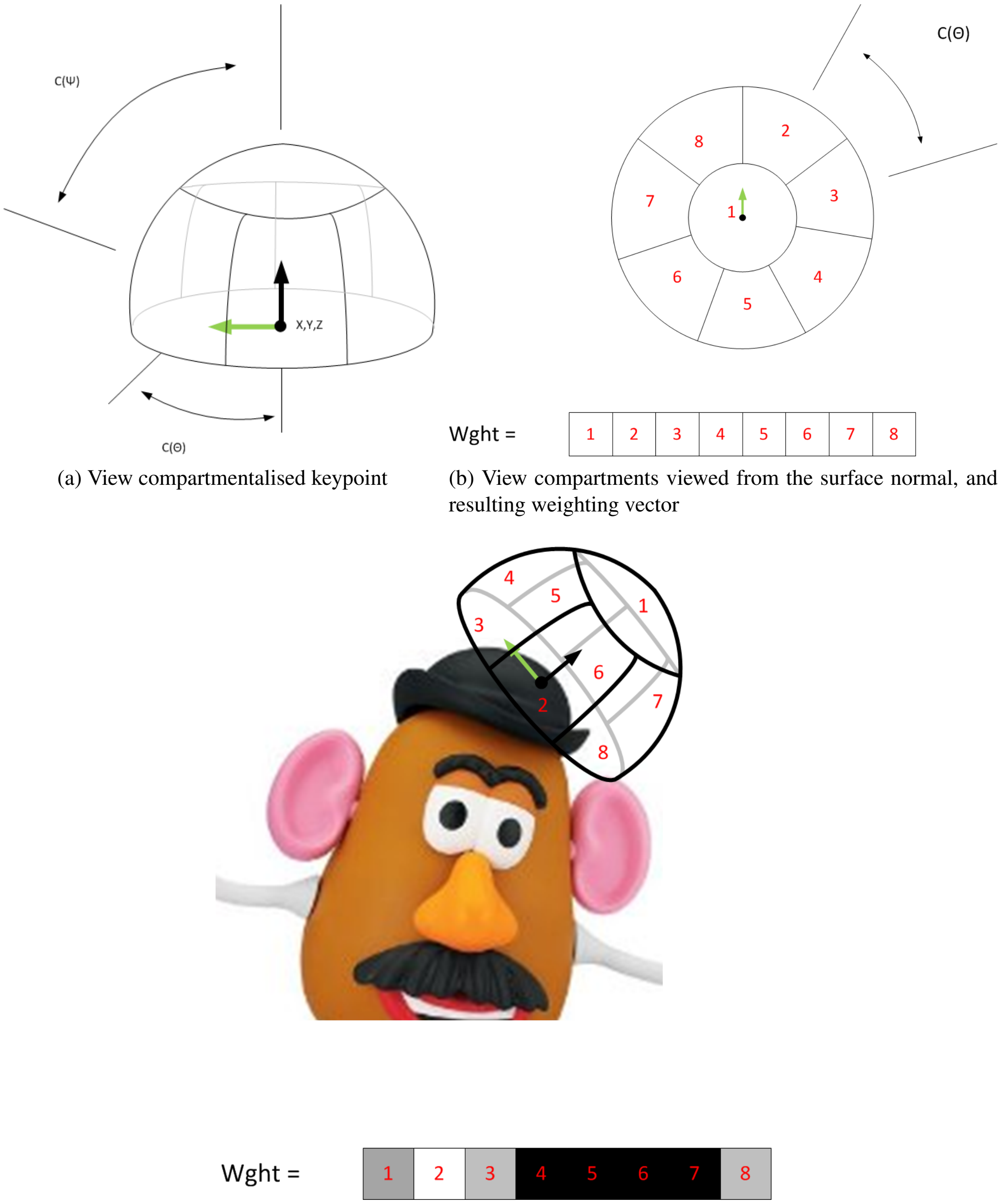
Figure 7.2a shows the arrangement of compartments covering the *keypoint viewing hemisphere* surrounding the keypoint location, the black arrow represents the surface normal at the keypoint location, the green arrow represents the in-plane surface orientation. Figure 7.2b shows the keypoint viewed from the surface normal with the *keypoint viewing hemisphere* decimated into compartments, numbered 1 to 8; the resulting descriptor weighting vector is shown below, the green arrow represents the in-plane keypoint orientation. Figure 7.2c shows an example of the keypoint weighting vector for a keypoint extracted from a 3D object. In the example shown in Figure 7.2c, compartment 2 best aligns with the viewing angle and therefore has the highest weighting value, visualised as a light coloured weighting element, represent a high weighting value. Equation 7.1 gives an outline of the extracted keypoint measurements; a SIFT descriptor for each view compartment, a corresponding weighting value for each view compartment, a keypoint pose component defining the 3D orientation of the keypoint, a keypoint scale and a 3D X, Y, Z keypoint position.

This keypoint extraction process allows for a keypoint to be detected, its descriptor computed and a match to be made based on a single observation. However, in addition to single observations keypoints the view compartmentalised keypoint extraction approach can allow keypoints locations which overlap between successive views to be integrated together, forming a more complete description of the 3D appearance of the local area surrounding a keypoint. The integration of keypoints from differing viewing angles can be achieved by updating the keypoint descriptor and weighting compartments when a new keypoint with a higher compartment weighting value for the same location on an object becomes available. The modular keypoint structure, see Equation 7.1, allows extracted descriptors to be of any type (investigated in Chapter 5), and extracted from either the intensity or depth modalities or combinations of both.

$$Key = \left\{ \begin{array}{c} \{Desc_{1,1:128}, Desc_{2,1:128} \ldots Desc_{n,1:128}\}, \\ \{Wght_1, Wght_2 \ldots Wght_n\}, \\ Pose_{3,3}, \\ \sigma, \\ x, y, z \end{array} \right\} \tag{7.1}$$

### 7.2.1 3D Keypoint Pose

The 3D pose of extracted keypoints is calculated using the local pose estimation outlined in Chapter 6. This chapter, however, extends the local pose estimation to use the second eigen vector as an estimation of in-plane 3D pose of the surface patch. The use of the second eigen vector to define an orthogonal principal axes is used in similar approaches for aligning 3D CAD models so as to extract descriptors (Furuya and Ohbuchi, 2009). The validation data from Chapter 6 was used to track a keypoint on the test object surface through examples of synthetic range images. A comparison of the estimated in-plane 3D pose angle against the ground truth is shown for a range of views of the

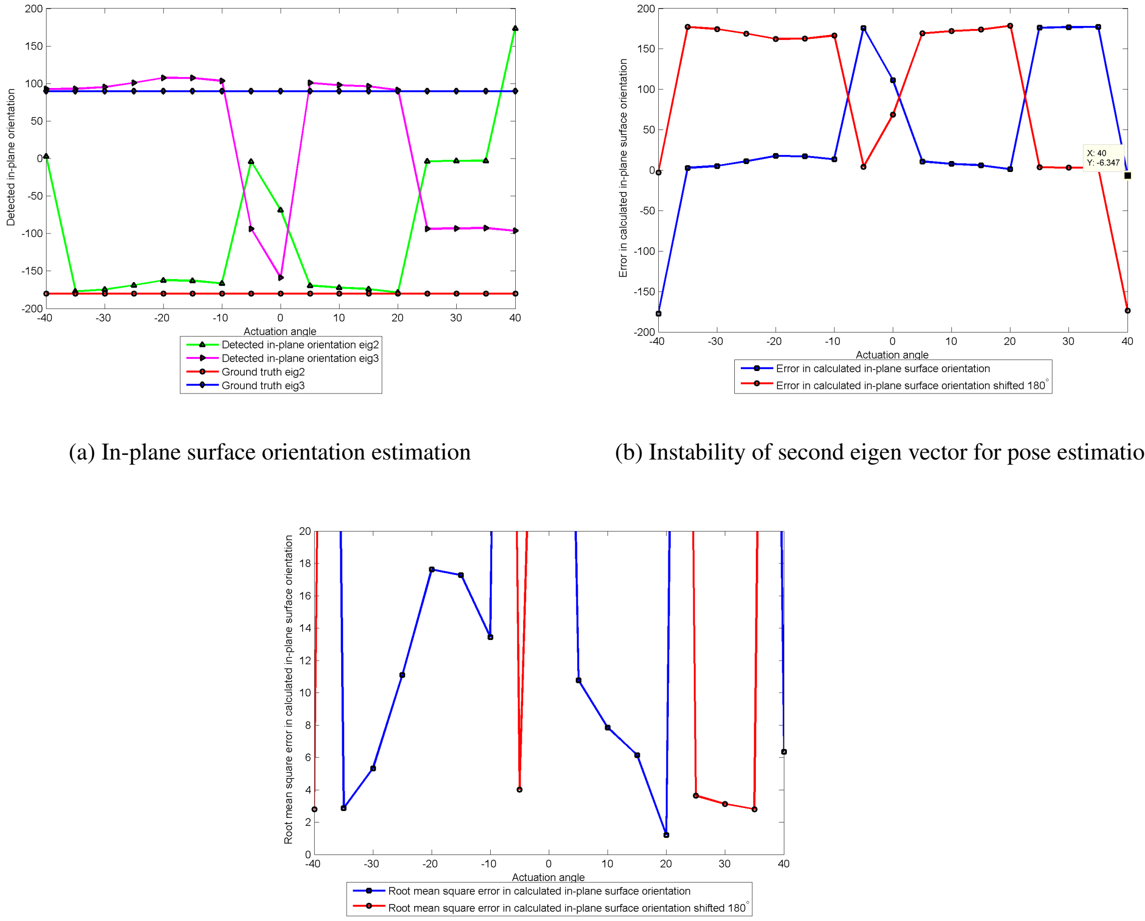


(a) View compartmentalised keypoint

(b) View compartments viewed from the surface normal, and resulting weighting vector

(c) Example of view compartmentalised keypoint sampling a 3D object

Figure 7.2: View compartmentalised descriptor



(a) In-plane surface orientation estimation

(b) Instability of second eigen vector for pose estimation

(c) RMS error for second eigen vector in pose estimation

Figure 7.3: In-plane surface orientation validation

keypoint in Figure 7.3a.

It can be seen that the calculated 3D in-plane pose orientation gives a close estimate of the actual in-plane orientation for 65% of the test examples. However, many of the incorrectly estimated examples show an in-plane ambiguity of 180°. Figure 7.3b shows the second eigen vector component of the 3D pose estimation duplicated with a 180°phase shift. Including this effect the in-plane pose estimation gives a close estimation for 95% of the example cases, only incorrectly labelling the 0° case. The root mean square error for the pose estimation is shown in Figure 7.3c, the maximum in-plane orientation inaccuracy is $\sim 18°$.

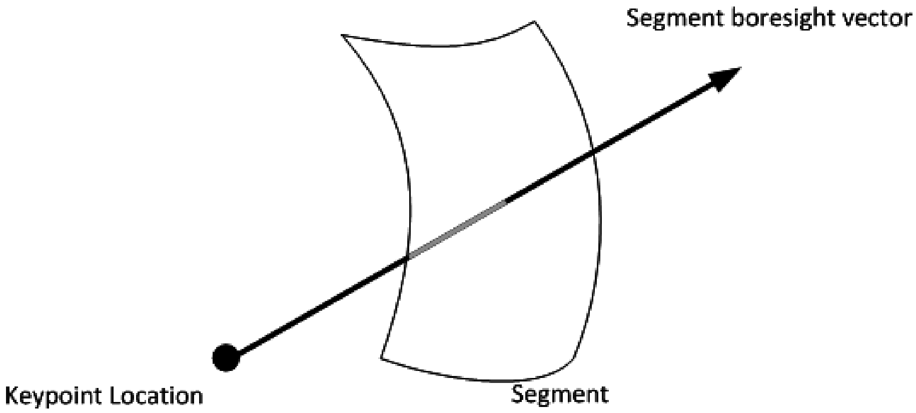


Figure 7.4: Segment Boresight

### 7.2.2 Pose Weighting

The pose weighting for a compartment is calculated as a linear function of the angular distance between the aligned compartment boresights and the viewing angle, $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$, see Equation 7.3. A compartment boresight is defined as the optimum viewing angle of the canonical view for the given compartment, see Figure 7.4. The compartment boresights are formed from the angle $C(\Psi)$, which defines the angular distance from the surface normal to the compartment boresight, and the angle $C(\Theta)$ which defines the angular division of the hemisphere surrounding the keypoint. These measures define a set of boresight vectors covering a viewing hemisphere, see Figure 7.2a. Figure 7.5, shows the compartment boresights on a range image before alignment, $\theta_{compartment}$, in blue. These are then rotated to align with the 3D local pose of the keypoint, $M_{pose}$, to give the boresights in the keypoint space, $\theta_{boresights}$, shown in red, see Equation 7.2.

The weighting function is calculated from the comparison of the observation vector, $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$, to each of the compartment boresights vectors, $\theta_{boresight_i}$. The dot product between these vectors gives the cosine distance which is linearised with the arc cosine. The result varies between 0 and$\pi/2$, such that dividing by $\pi/2$ allows the pose weighting to vary between 0 and 1, see Equation 7.3.

$$\theta_{boresight_i} = M_{pose}\theta_{compartments} \tag{7.2}$$

$$Wght_{pose_i} = \frac{\operatorname{acos}(\operatorname{dot}(\theta_{boresight_i}, viewangle))}{\pi/2} \tag{7.3}$$

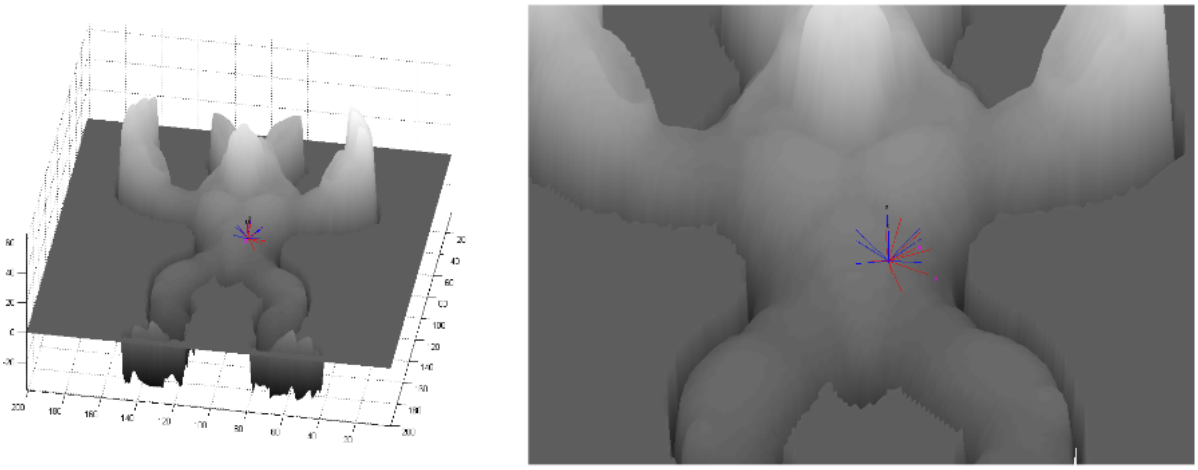


Figure 7.5: View Compartment Boresights

### 7.2.3 View sphere decimation

The view sphere is partitioned by the choice of boresight locations. These are chosen such that each compartment covers an angular range greater than twice the uncertainty in the keypoint pose estimation. The keypoint pose uncertainty was found in Section 7.2.1 to be $\sim 18^\circ$. Therefore, view compartments should have a minimum width of $36^\circ$. Additionally, the maximum separation should be set to reflect the range of views over which a local feature descriptor remains stable, this sets an ideal maximum separation of $30^\circ$. However, an increase in the number of view compartments creates an increase in the length of keypoint feature descriptor, and therefore an increase the processing power required to match descriptors. Accounting for these factors the view sphere decimation is initially set as a compromise between descriptor length and feature descriptor stability, with boresight separations of $C(\Theta), C(\Psi) = 45^\circ$, allowing $180^\circ$ of view hemisphere to be partitioned into 3 compartments. This arrangement gives 9 compartments, 1 central, and 8 surrounding the surface normal, and a feature descriptor of length 9x128 = 1152 elements.

### 7.2.4 Compartment Assignment

During a training phase the view sphere of an object may be explored. Locations on the object surface may be tracked using motion ground truth, ICP, optical flow or keypoint consensus approaches such as the GHT; using these approaches it is possible to establish the association between individual keypoints from multiple observations. Individual keypoints of the same scale from differing observations which demonstrate large overlaps in sample area can be considered to belong to the same master keypoint, with the individual keypoints describing differing expressions of the keypoint for differing viewing angles. The view compartmentalised keypoint approach integrates individual keypoints, describing the same location on the object, together to form a master keypoint by substituting compartments from keypoints in the database with corresponding keypoint compartments when a higher confidence value becomes available.

Figure 7.6, shows an example of this substitution approach where 3 observations of a keypoint are collected from differing viewing angles. The keypoint observations have different compartment weightings based on the pose of the keypoint at the point of observation, these are shown as shading of the keypoint compartments. In the association stage the view compartments with the higher weighting values replace those view compartments with lower weightings to form the single master keypoint which then describes all the original keypoint locations.

## 7.3 Matching

The matching function for the view compartmentalised keypoint extraction algorithm presented in this chapter must account for both descriptor matches between the compartments of the model and query keypoints and the compartment weighting values. The confidence values for each compartment are based on the pose of a query keypoint and the learned appearances of the target keypoints. In order to utilise the additional confidence information in the matching approach a weighted Euclidean distance is proposed. The weighted Euclidean distance weights the dimensions of a Euclidean space by the product of confidence measures for corresponding compartments in the model and query keypoints. The keypoint compartment weightings in the model and query keypoints are termed: $modelWght_{comp_i}$ and $queryWght_{comp_i}$. The product of these compartment weighting values are used to create a weighted space which is unique to each model-query keypoint pair. To compare the distances between all query and model keypoints, the weighted spaces of each model-query keypoint pair must be normalised to unit length, see Equations 7.6 and 7.7. Matches between model and query keypoints are then established using the Log-likelihood of the model-query pair distances as used in Chapters 3 and 5, and in SIFT (Lowe, 2004, 1999), see Equation 7.8.

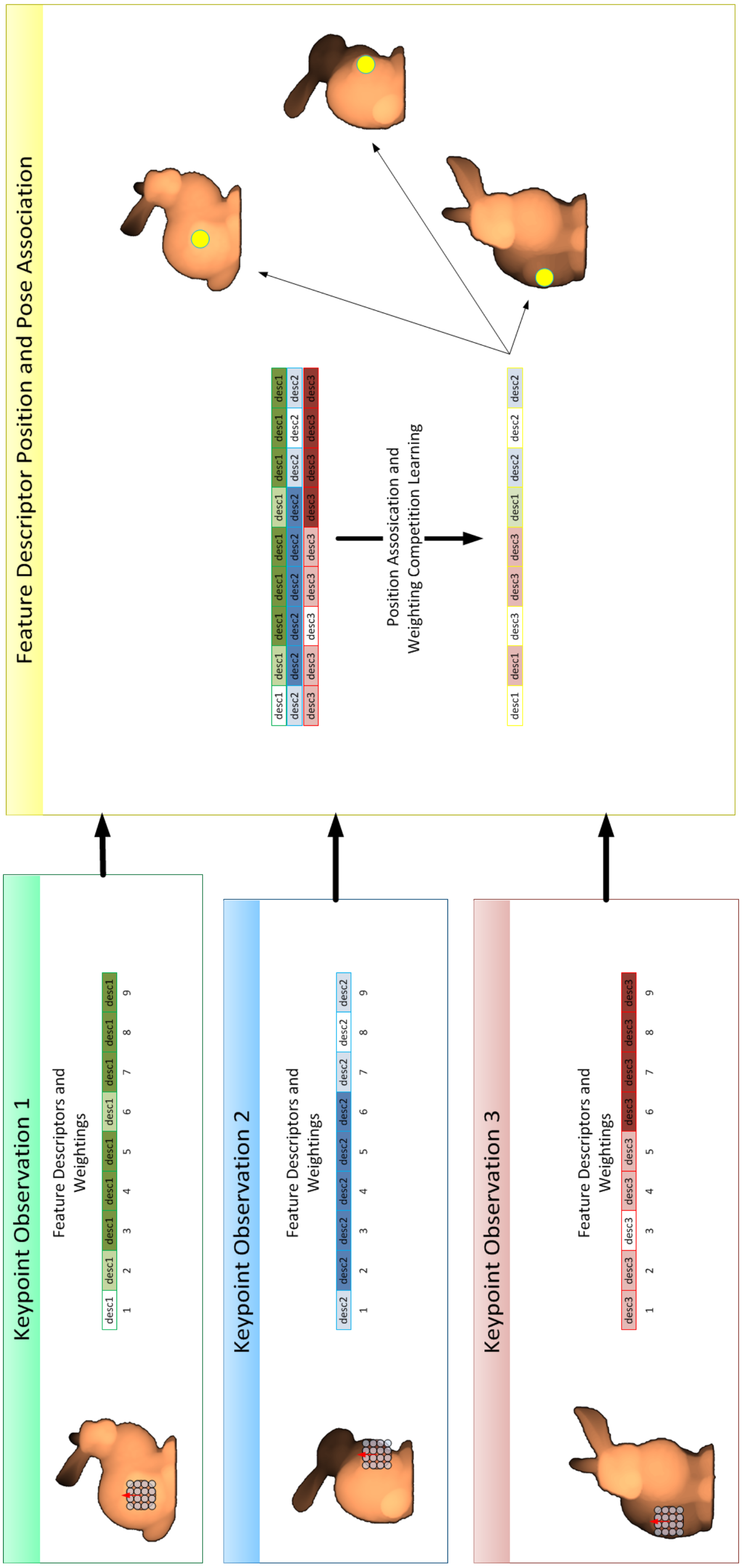

Keypoint Observation 1
Feature Descriptors and Weightings
desc1 desc1 desc1 desc1 desc1 desc1 desc1 desc1 desc1
1 2 3 4 5 6 7 8 9
Keypoint Observation 2
Feature Descriptors and Weightings
desc2 desc2 desc2 desc2 desc2 desc2 desc2 desc2 desc2
1 2 3 4 5 6 7 8 9
Keypoint Observation 3
Feature Descriptors and Weightings
desc3 desc3 desc3 desc3 desc3 desc3 desc3 desc3 desc3
1 2 3 4 5 6 7 8 9
Feature Descriptor Position and Pose Association
Position Assosication and Weighting Competition Learning
desc1 desc1 desc3 desc3 desc3 desc1 desc2 desc2 desc2

Figure 7.6: Combining View Compartmentalised Keypoints

$$Val_i = queryWght_{comp_i} modelWght_{comp_i} \left(queryDesc_{comp_i} - modeDesc_{comp_i}\right)^2 \tag{7.4}$$

$$\hat{Val} = \sqrt{\sum Val_i} \tag{7.5}$$

$$normWght = \sqrt{\sum \left(queryWght_{comp_i} modelWght_{comp_i}\right)^2} \tag{7.6}$$

$$dist = \frac{\hat{Val}}{normWght} \tag{7.7}$$

$$Match = \begin{cases} 1, & \text{if} \frac{dist_1}{dist_2} < distRatio \\ 0, & otherwise \end{cases} \tag{7.8}$$

### 7.3.1 Matching Validation

A system for matching view compartmentalised descriptors should have the following desirable qualities:

- A match should be established where a descriptor match in a compartment with a high weighting value exist.
- Keypoints where a descriptor match does not exist but a view compartment has a high confidence should be rejected.
- Descriptor matches which exist in incorrect view compartments should be rejected.
- Keypoint matching should be robust to noise in either the feature descriptor, the compartment weighting or both.

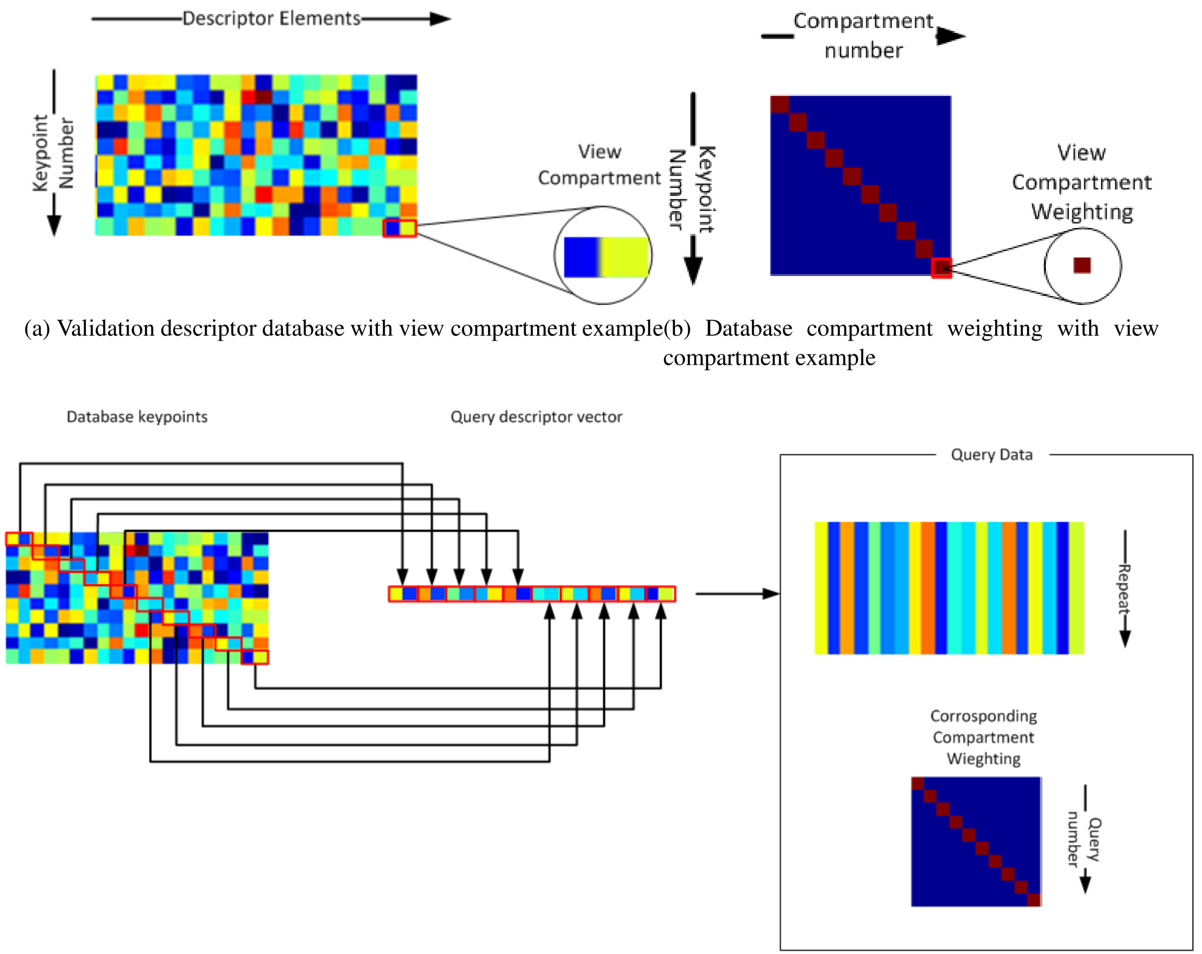
To investigate the effectiveness of the proposed weighted Euclidean distance with regards to these criteria, the system was tested on synthetic data. The synthetic example data was created with characteristics similar to the intended application. However, in order to visualise the data a simplified validation was conducted with reduced complexity descriptors.

A database of 10 keypoints each with descriptors of length 20 elements was created using a random number generator. The descriptors are divided into 10 compartments each with descriptors of length 2 elements, each descriptor compartment vector is normalised to unit length, see Figure 7.7a. Note that this normalisation stage reflects the normalisation stage of the view compartmentalised keypoint descriptors proposed in Section 7.2; where the feature descriptor for each compartment is normalised to unit length. However, the final view compartmentalised feature descriptor vector is not of unit length. The baseline query view compartmentalised feature descriptor vector is created from the concatenation of one descriptor compartment from each of the keypoints in the database, see Figure 7.7c. The query compartmentalised feature descriptor vector is repeated to create a database of query view compartmentalised descriptors. The database of keypoint weighting vectors is formed by assigning each of the query descriptor vectors a different compartment weighting vector. In the initial experiments in this validation the query weighting value is either set to 1 to represent the presence of a compartment or 0 to denote the absence, see Figure 7.7b. Using this synthetic data 4 experiments were conducted. The remainder of this section details these validation experiments.

#### 7.3.1.1 Descriptor matches in compartments with high compartment weightings

The weighted Euclidean matching function was tested for its ability to minimise a distance between query keypoints and a known database keypoint match. The results for the distances are displayed as a match matrix showing the distance between every query keypoint and database keypoint pair, dark blue indicates a low distance between the keypoints in match space, which corresponds to a high match score. Figure 7.9b, shows the results of comparing keypoints with identical feature descriptors, although differing compartment weighting values for each keypoint. In this experiment there is a strong keypoint association along the diagonal of the match matrix where the correct matches between query and database keypoints exist. This result shows that correct matches exist where descriptor matches are in the correct compartments with a high compartment weighting value. Matches exist for all other compartments, however these compartments have low weighting values and as a result are suppressed by the matching function.

Shuffling the query weighting function shows the match matrix following the new weighting function, maintaining correct matching correspondence between the query and database keypoints, see Figure 7.8b. This result shows that the order of the keypoints in the query database does not



(a) Validation descriptor database with view compartment example

(b) Database compartment weighting with view compartment example

(c) Query keypoint data creation

Figure 7.7: View compartment matching function validation data

affect the match score.

#### 7.3.1.2 Descriptor matches with low compartment weightings

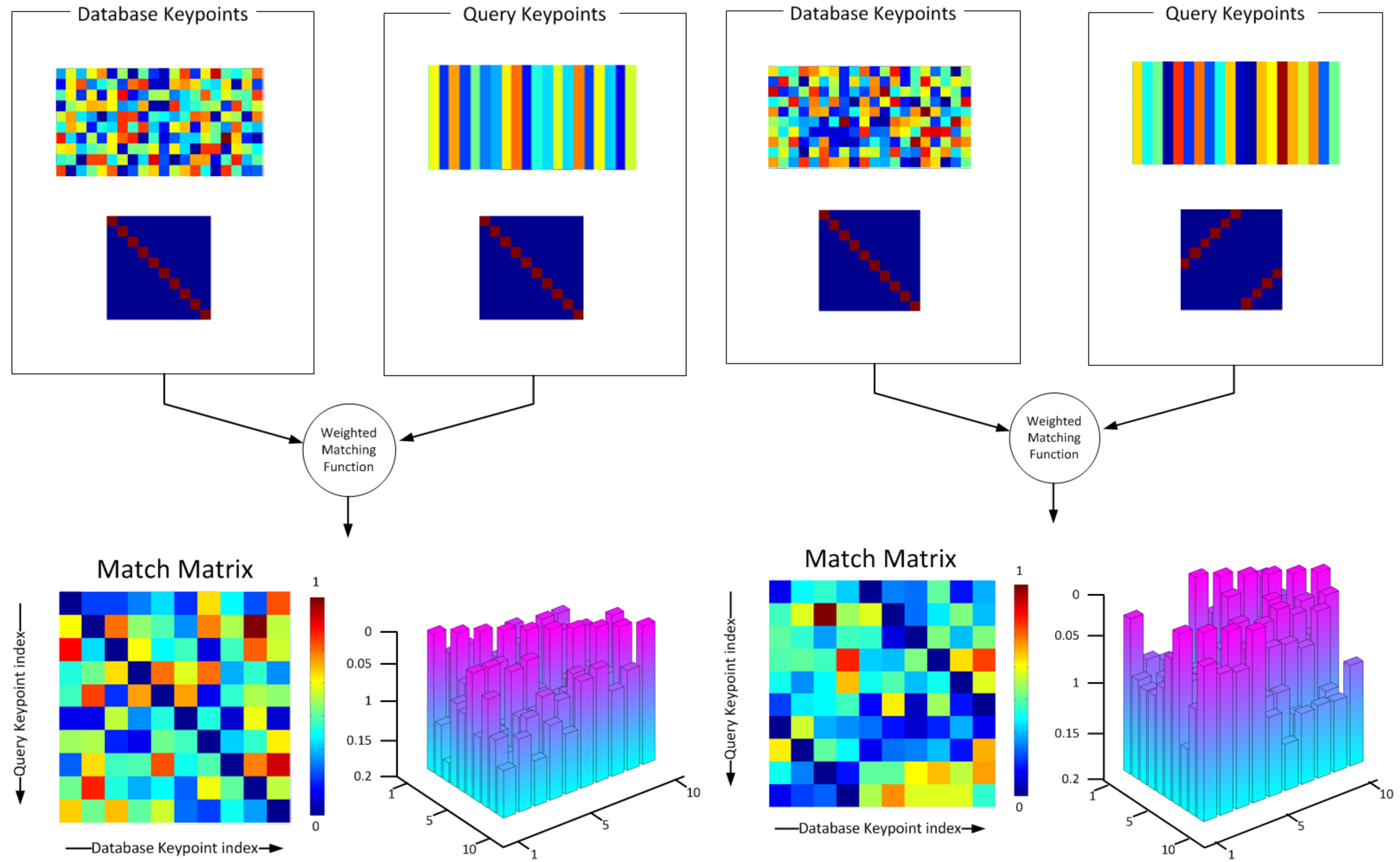
To investigate the case where descriptor matches exist with low compartment weightings, the database compartment weightings were shuffled. In this example, descriptor matches exist along the diagonal of the database keypoints; however these compartments in the database keypoints have now been assigned a low weighing value. As a result the match matrix for this experiments shows that no distinct keypoint matches, see Figure 7.9a. This result demonstrates the ability of the weighted Euclidean distance to suppress keypoints which have a descriptor match in the correct compartment, although have a low confidence measure for the matching compartment.

#### 7.3.1.3 Descriptor matches in incorrect compartments

This section tests the ability of the weighted Euclidean distance to discern between similar appearances existing in different canonical views of an example of keypoints. In this circumstance individual feature descriptors will form matches however the compartment ordering will be incorrect. The compartment ordering in a view compartmentalised descriptor is a result of the structuring of the keypoint based on 3D pose. An example of this is created in the synthetic keypoints by moving the query keypoint compartments round by one compartment, see Figure 7.9b. In this example the keypoint match scores indicate that there are no distinct matches. This validation establishes the ability of the matching function to reject correct descriptor matches where high compartment weightings exist in both model and the query keypoints, although the order of the compartments is incorrect.

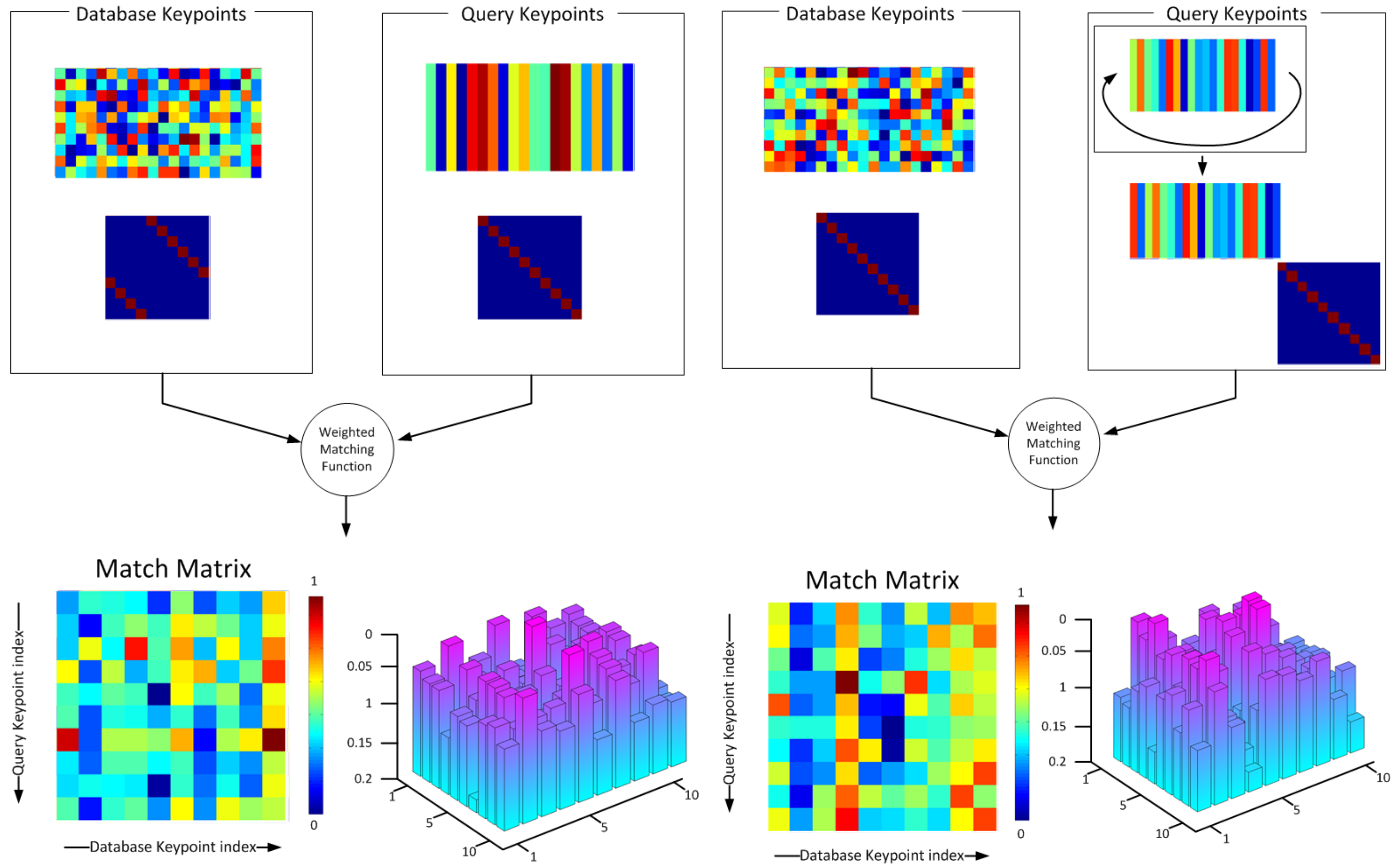
#### 7.3.1.4 Descriptor matching in the presence of noise

Robust keypoint matching where noise is present is another desirable property of any matching system. To investigate this 10% descriptor noise and 10% weighting noise are applied to the query keypoints, the match matrix for this result shows high match scores across the diagonal. However, in this experiment high match scores can be seen elsewhere in the match matrix, indicting false positive matches. This effect is potentially the result of low descriptor dimensionality in the test data. To investigate further a set of database view compartmentalised keypoints were created with



(a) Matching function validation results with no noise and direct correlation, match matrix(b) Matching function validation results with no noise and query keypoints weightings shuffled, match matrix

Figure 7.8: View Compartmentalised Matching Validation 1



(a) Matching function validation results with no noise and database weightings shuffled, match matrix

(b) Matching function validation for keypoints which have descriptor matches in incorrect compartments

Figure 7.9: View Compartmentalised Matching Validation 2

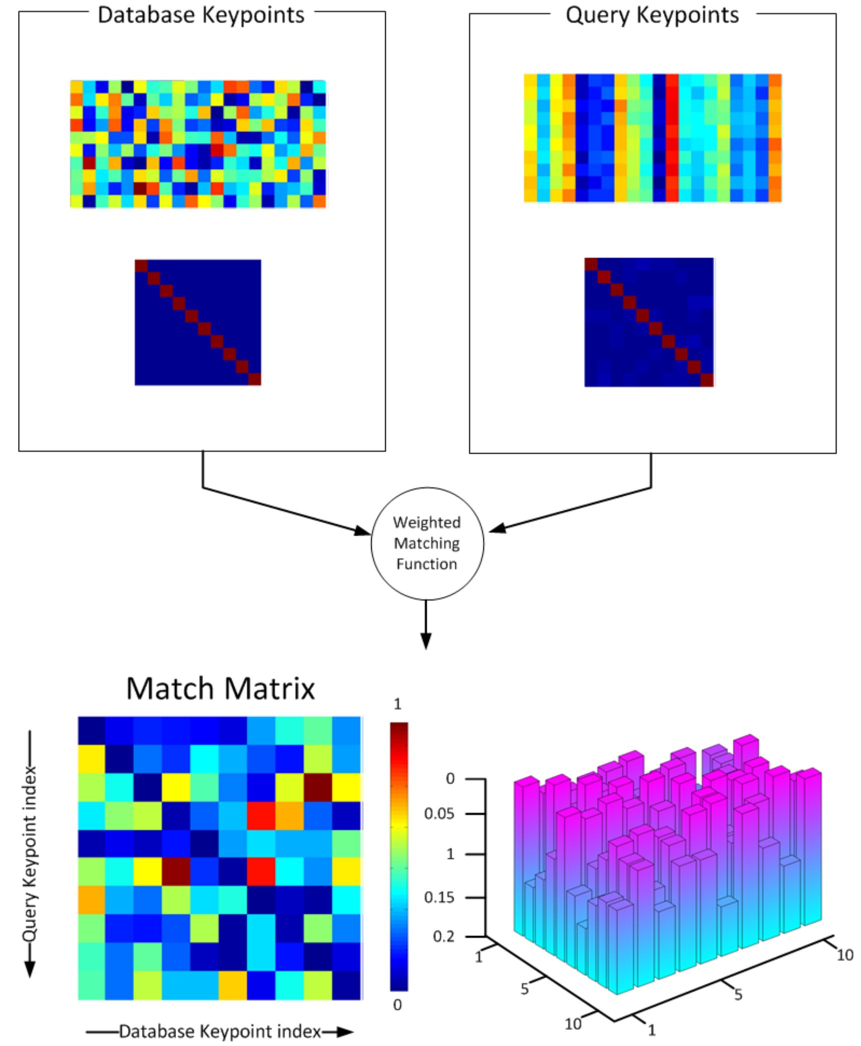


Figure 7.10: Matching function validation results with 10% descriptor noise and 10% weighting compartment noise, match matrix

compartment descriptor lengths of 30 elements. The experiment using the 30 element compartment descriptors uses a lower descriptor noise of 1% was and a higher weighting noise of 30% to emphasis the effect of noise in the compartment weighting. Figure 7.11, shows the query view compartment weightings and the corresponding match matrix. In the match matrix it can be seen that the highest match scores are those for the correct matches. However, good match scores can also be seen to be highly correlated with query compartment weightings. This artefact is explain by the fact that all query descriptors have a corresponding database match in every view compartment. However, this result does indicate that the proposed weighted Euclidean distance has a high sensitivity to the view compartment weightings.

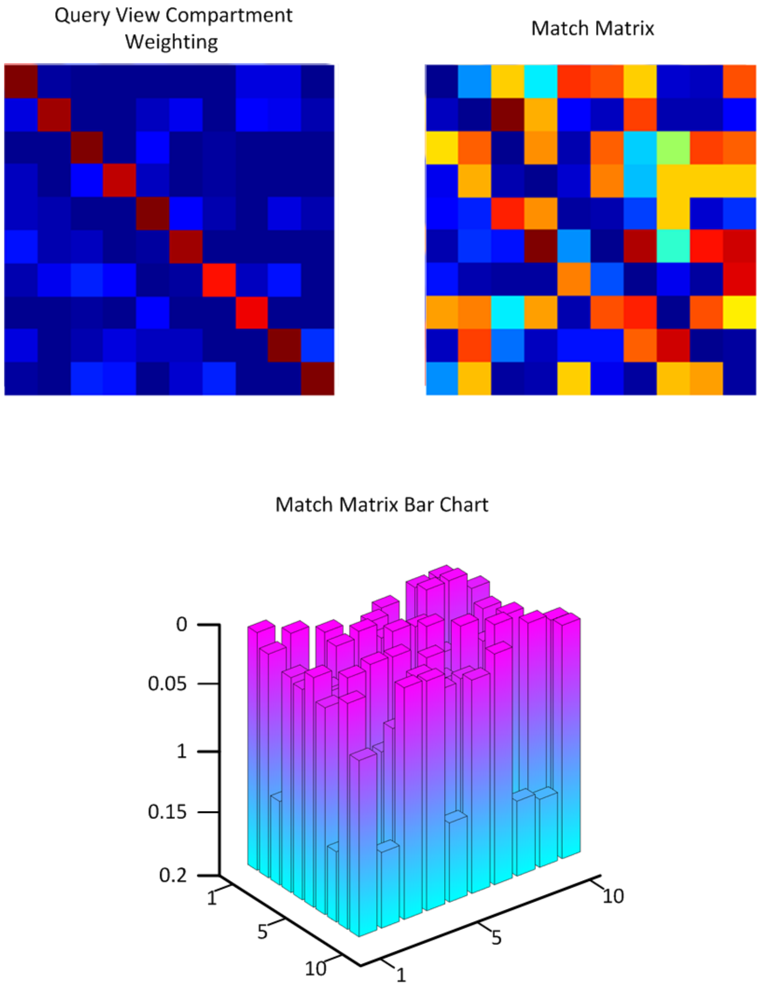


Figure 7.11: Compartment weighting and match matrix for view compartmentalised keypoints, with descriptor length of 30, 10 compartments, 1% descriptor noise, 30% weighting noise

## 7.4 Approach Overview

This section summarises the steps taken to create a view compartmentalised keypoint.

1. **Keypoint localisation,** Keypoints are localised in either range or intensity images or both using the SIFT keypoint localising approach. The structure of keypoint localisation used by the view compartmentalised keypoints allows the extraction process to take advantage of multimodal features, investigated in Chapter 5.

2. **Keypoint extraction,** keypoint are extracted using the processes found in Chapter 5 to give the most repeatable keypoint descriptors, these were: Affine corrected SIFT keypoints and element corrected 2.5D range SIFT keypoints. In addition to these standard SIFT is included as a baseline comparison.

3. **Keypoint pose estimation,** Following feature localisation and extraction, each of the keypoint locations has an associated local 3D pose. This pose is based on the principal axes of sample points on the range surface surrounding the keypoint. The principal axes are defined as the surface normal and a 3D in-plane orientation component.

4. **Compartment weighting function,** A weighting value for each compartment is established through the comparison of the viewing angle with a set of compartment boresights aligned to the keypoint pose. The weighting function is determined as a linear function of the angular difference between the current view and the canonical view for any given compartment.

5. **Update database keypoints,** the database of keypoints is created during an exploration of the view space. While exploring the view space, points on the surface are tracked and keypoints extracted from the same scale and are of a distances of less than 50% of the sample patch size are associated together. The association of keypoints from the same location is achieved by updating existing keypoint compartments in the database with new keypoint compartments which have a higher compartment weighting value.

6. **Keypoint matching,** A weighted Euclidean distance has been introduced which allows keypoints to be compared based on compartments with a shared level of confidence. To match view compartmentalised keypoints, descriptor matches must exist in compartments which

have a significant weighting value. The compartment weighting value serves to reject keypoints where the level of confidence for a given observation is low, while the compartment assignment function serves to compare keypoint appearances from similar views.

## 7.5 Results and Analysis

This section covers the direct comparison of keypoint pairs from single observations using the experimental methodology outlined in Chapter 4 to form ROC curves, Section 7.5.1. An example of the exploration of the view sphere for view compartmentalised keypoints with 9 compartments is presented in Section 7.5.2; where the range images which formed the best match to each compartment for a single keypoint through multiple observations are displayed. Additionally a view compartmentalised descriptor with 3 compartments is also investigated in Section 7.5.3.

### 7.5.1 ROC curve performance

The ROC performance of the top feature extraction approaches found in Chapter 5 were compared with their equivalent view compartmentalised keypoint extraction approaches. This experiment investigates the effects of view compartmentalisation on a pair of keypoints each generated from a single observation. The view compartmentalised descriptors have a repetition of the feature descriptor component in every compartment with only the compartment weighting value changing between compartments in a keypoint. Therefore, this experiment investigates the combination of keypoint pose estimation and the resulting weighting and matching function. Figure 7.12 shows the performance for keypoint matches formed from a single observation.

The results show that for all cases the resultant view compartmentalised keypoints exhibit a decreased performance when compared with the non-view compartmentalised keypoints, with the exception of the unmodified SIFT case, where the view compartmentalised keypoints exhibit a marginal improvement in performance. As all keypoint compartments contain the same feature descriptor, which forms matches between instances in the non-view compartmentalised keypoints, the decrease in performance is expected to be attributed either to the instability in pose estimation observed in Section 7.2.1, or the weighted Euclidean distance function.

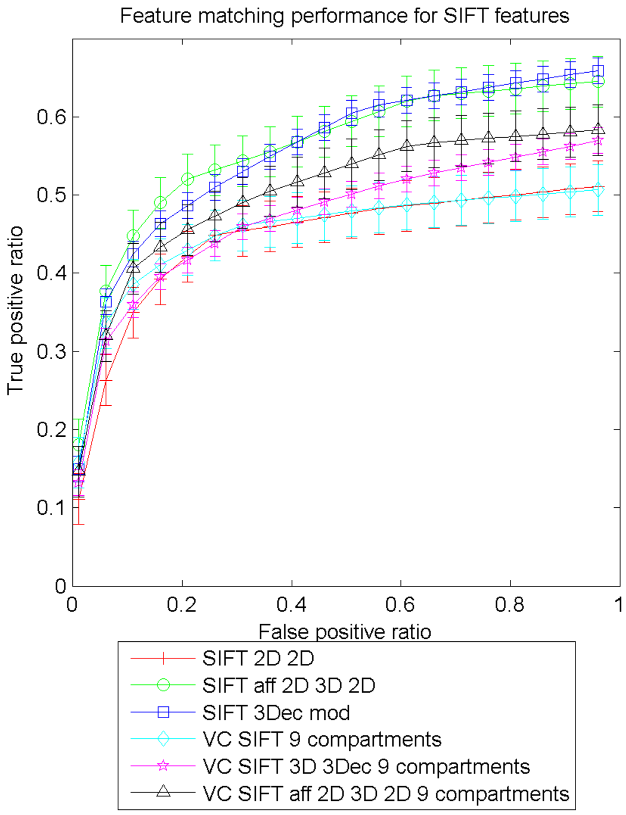


Figure 7.12: View Compartmentalised Keypoints ROC curve

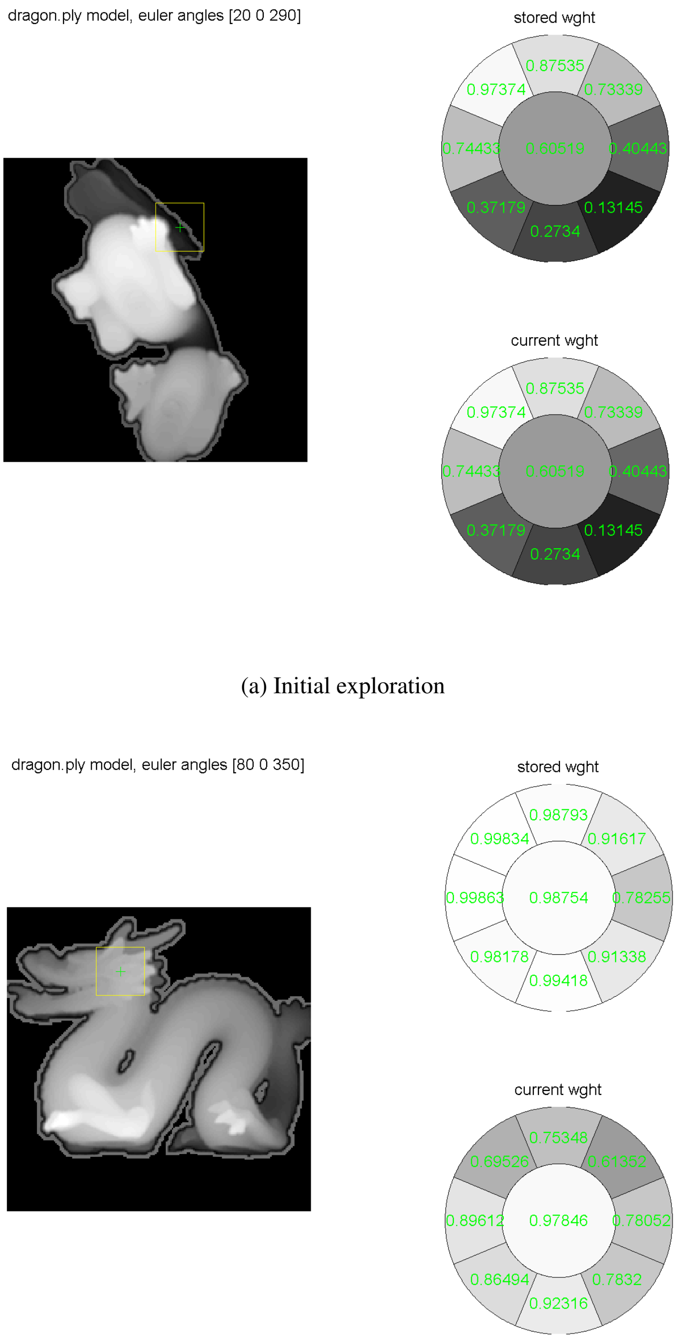
### 7.5.2 Pose based compartment assignment

Keypoints compartments in the database may be updated during an online learning phase when keypoint observations with higher compartments weighting values become available. The compartments with higher weighting values then replace the original database keypoint compartment characterising the same keypoint view. To investigate the views which are stored as the canonical views of the object for each compartment the view space of the Dragon model, used in Chapter 6, was explored for a manually labelled location on the object surface. The manually labelled location was then transformed between observations instances using the inverse of the transformation applied to create each synthetic range image. Figure 7.13a, shows the first view of the object in the exploration; here every compartment of the the database keypoint is updated with the those from the current observation. In the initial view the upper left most compartment, compartment 9, shows the highest confidence, this indicates that the in-plane orientation has assigned this compartment to the lower left most corner of the object. Figure 7.13b, shows the state of the view sphere exploration after the lower half the view sphere has been covered, half the compartments in the database keypoint have high confidence values. Additionally, the weighting values indicate that the current view of the keypoint best aligns with the central view compartment.

By recording the range image from each of the views which best aligns with the compartment boresights a visual representation of the optimum viewing angle for each compartment can be created, see Figure 7.14. This figure shows the compartments in the same arrangement as shown in Figure 7.13a, where compartment 1 is in the centre, with compartments 2 to 9 surrounding starting from the top and increasing in number clockwise. From this figure it can be seen that the highest candidate matches for each of the keypoint compartments appear to be incorrectly identified, indicating that the pose estimation for the keypoint at these locations has been incorrectly estimated. From the examples presented in Figure 7.14, the incorrect pose estimation appears to result from the object viewed at extreme observation angles resulting in the keypoint pose estimation sample patch sampling the surface as a point cloud with high variability.

### 7.5.3 Reduced complexity compartmentalised keypoint

In Section 7.2.1 the in-plane orientation estimation was shown to be unstable and liable to 180° shifts in pose estimation, whereas the surface normal estimation outlined and investigated in Chapter



(a) Initial exploration

(b) Half view sphere explored, observation taken at surface normal

Figure 7.13: View space exploration and database keypoint compartment weighting value update

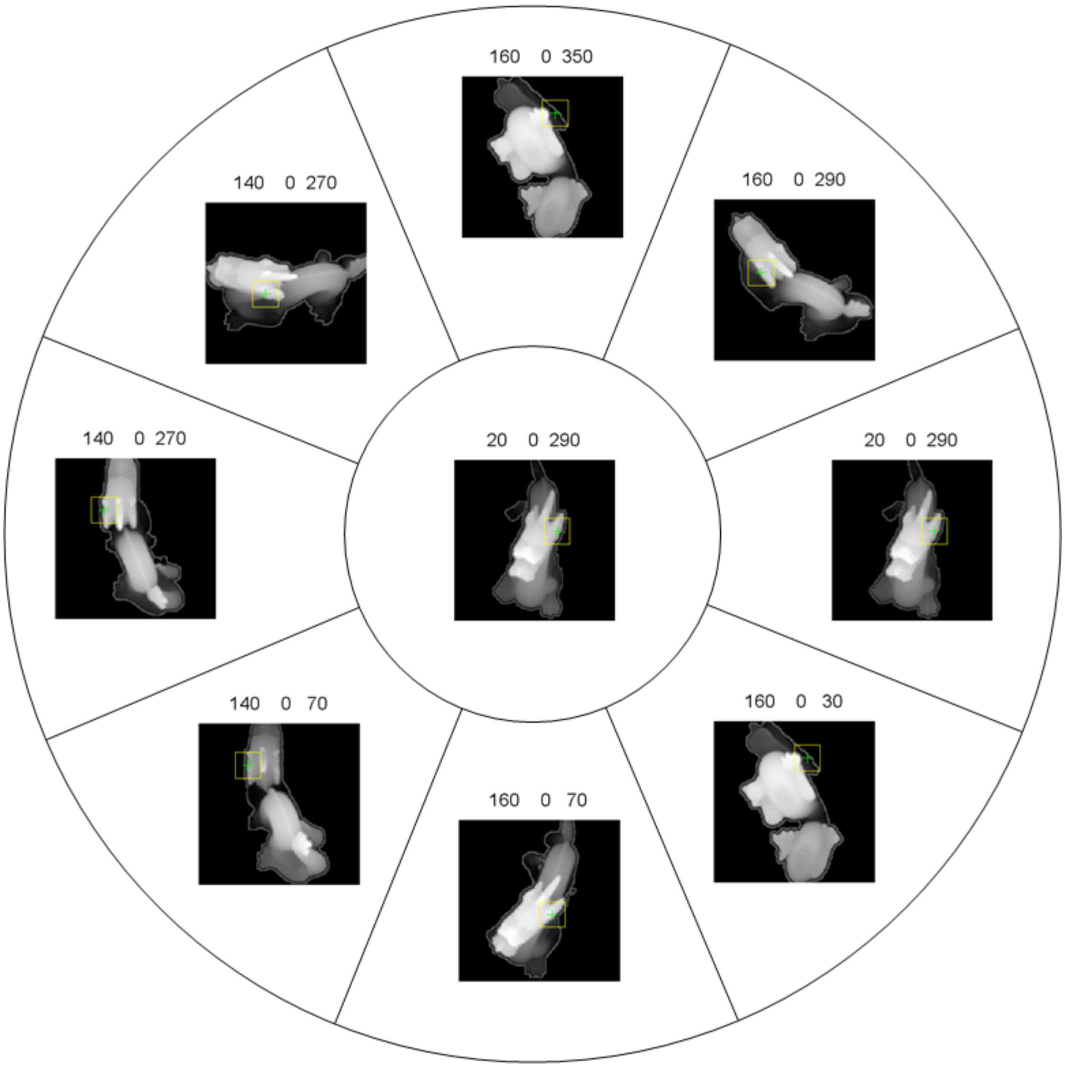


Figure 7.14: Top range image matches for each compartment

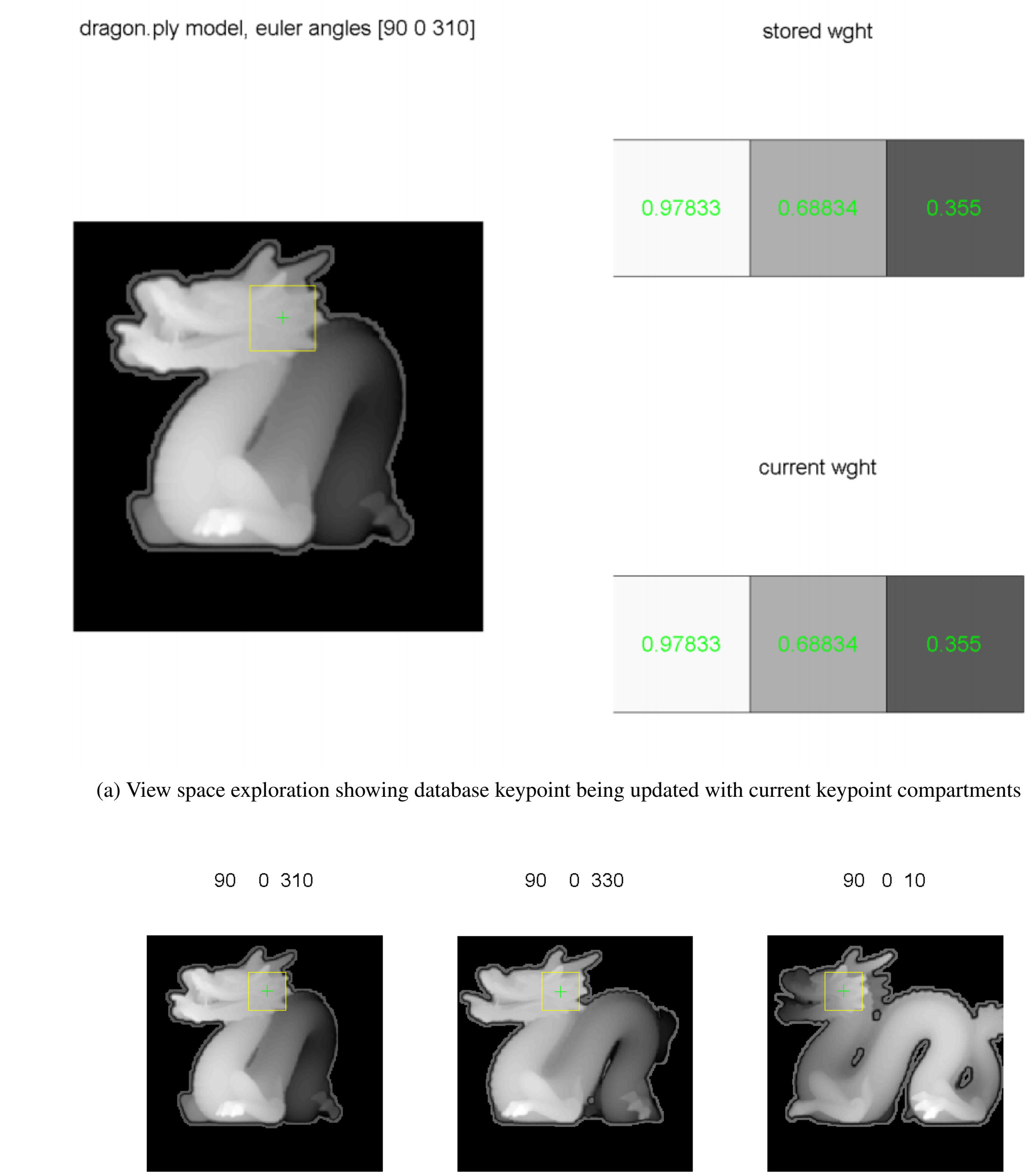
6 was found to be more stable, due to the ±90° limitation in surface normals offered by the range image. In addition to the limitation in views offered by the nature of the range image, the test data used for the experiments is limited to applying an out-of-plane rotation only in the yaw axis. These limitations in data allows a reduced complexity compartmentalised keypoint to be formulated which investigates the performance of the view compartmentalised keypoint methodology in an example of reduced complexity.

The reduced complexity example has compartment boresights along the yaw plane at ±45° and 0°, see Figure 7.15a shows an example of the weighting component of the keypoint. The view space was then explored along the yaw plane only and the range images with the highest confidence in each compartment is shown in Figure 7.15b. The highest candidate matches in this example appear, by inspection, to be more intuitively correct. Figure 7.16, show the ROC curve for this arrangement of view compartmentalised keypoint. It can be seen that view compartmentalised SIFT with 2D intensity gradients shows a significant improvement in ROC performance when compared with the 9 compartment view compartmentalised keypoints. View compartmentalised Element Corrected SIFT shows a marginal improvement in ROC performance, however view compartmentalised Affine corrected SIFT shows a decrease in ROC performance. Further improvements are expected with more stable estimations of keypoint 3D pose.

## 7.6 Summary and Conclusions

This chapter has introduced view compartmentalised keypoints, which utilise range data to create a 3D structured keypoint description from standard SIFT keypoints. Standard SIFT keypoints have stages in processing pipe-line for correcting for in-plane orientation invariance and increasing feature descriptor distinctiveness through a measure of the composition of the surface statistics. This is achieved through aligning the measurement aperture with a calculated canonical orientation and spatially dividing the area surrounding the keypoint into a series of sub patches. The aim of the approach presented in this chapter has been to apply an analogous methodology, inspired by the concept of canonical views from view bubbles (Peters et al., 2002), for aligning keypoints with a canonical 3D pose and compartmentalising an appearance of the keypoint at each observation angle.

In Chapter 5 it was noted that range images alone do not form many keypoints and typically have low surface variation. Whereas, intensity images form a greater number of keypoints and encode a



(a) View space exploration showing database keypoint being updated with current keypoint compartments

(b) Top range image matches for each compartment

Figure 7.15: View space exploration for view compartmentalised descriptor with 3 compartments

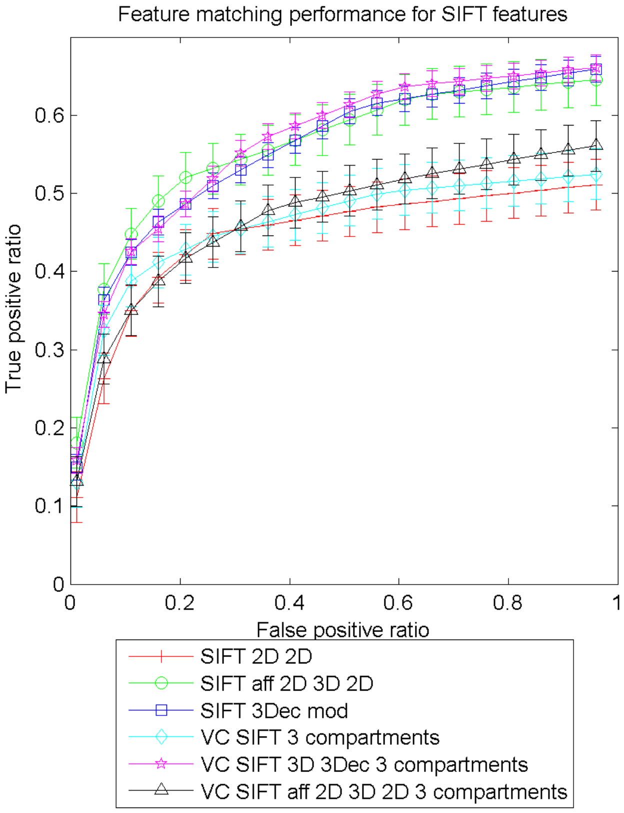


Figure 7.16: View compartmentalised keypoint with 3 compartments

richer level of texture information. Both modalities, however suffer from appearance changes due to pose, such that single observation keypoints are restricted to $\pm 30^\circ$ of out-of-plane view point changes. Therefore, the approach outlined in this chapter was aimed at realising the potential of co-aligned range and intensity images to create a pose invariant object description, by utilising range images to establish a set of canonical views, and both range and intensity features to characterise the keypoint appearance in these views.

This was achieved through a view compartmentalised keypoint arrangement comprising a keypoint feature descriptor and weighting value for each view compartment, and a novel weighted Euclidean distance. The compartment weighting values characterise how well a given observation of a keypoint aligns with the optimum view for each view compartment. The weightings are used in a training phase, where the view sphere of the object is explored, and subsequent views of a keypoint are used to update a master keypoint. During the matching phase, a weighted Euclidean distance is used to ensure that only descriptors as observed from similar viewing angles are compared. A set of 4 desirable properties for the proposed weighted Euclidean distance were identified and the system was validated with tests conducted on synthetic data.

The performance of point-to-point correspondences from single observations of view compartmentalised keypoints was investigated using the experimental configuration from Chapter 4. The results of the ROC curves for view compartmentalised keypoints with 9 compartments exhibited a significantly decreased performance when comparing view compartmentalised keypoints to their non-view compartmentalised counterparts. An exploration of the view space of the object for a single keypoint instance showed that this result was due to an instability in the calculation of the compartment weighting resulting from an instability in pose estimation. In order to investigate view compartmentalised keypoints with a reduced dependency on the 3D in-plane pose estimation, a view compartmentalised keypoint with 3 compartments was proposed and investigated. Through the exploration of the view space of an object for single keypoint, the compartment assignment for these keypoints demonstrated a greater level of stability than the case of 9 view compartments. Using the 3 compartment view compartmentalised keypoints the ROC curves, in most cases, showed a marginal, however statistically insignificant, increase in feature matching performance when compared with non-view compartmentalised keypoints, with the exception being for Range-Affine intensity features.

# Chapter 8

# Conclusions and Future Work

This chapter summaries the research conducted in this thesis and the placement of this work in the current literature. The achievements and limitations of the work conducted are outlined. The chapter concludes with potential directions for future work, where the work conducted in this thesis may be furthered or applied.

## 8.1 Thesis Objectives

The objective of this work was to investigate the use of range and intensity imaging modalities in local features. Range images have been frequently cited in the literature as having desirable properties for many applications. The recent availability of co-aligned range and intensity domain information for applications, such as robotics, has prompted the question of how to combine these imaging modalities to form a single robust local feature representation for a location on an object surface. The SIFT architecture was chosen to achieve this goal as this have been frequently cited, and the structure of the algorithm pipeline is readily reconfigurable, allowing the application of the range imaging modality in a local feature context to be investigated.

This objective was decomposed into the following goals:

- Create a validation approach and co-aligned range and intensity image dataset to gauge the performance of all modifications, and ensure that the work consistently progressed with measurable improvements.
- Examine the applications of range domain images and co-aligned range and intensity images

in the existing SIFT structure, as an approach for improving the invariance of keypoints to common transformations.

- Extend the current SIFT structure through processing of extracted feature descriptors to account for variation resulting from pose changes.

In order to address these objectives this thesis has followed a structured approach for evaluating the available design choices when formulating SIFT-like features, and analysing where range and intensity information may be applied to best improve the resultant description of a location on an object between observations.

At the beginning of this thesis the following hypothesis was made:

> "This thesis argues that it is possible to achieve a performance improvement over existing local feature matching approaches by exploiting information from co-aligned range and intensity domain images. Range images offer a partial representation of the 3D object surface which can allow keypoint feature descriptors to encode further information regarding the 3D appearance of the keypoint, which is unavailable when using only the intensity imaging modality. This additional information regarding the object structure can be encoded together with the intensity domain information to form a robust local feature descriptor for an image location."

These objectives have been achieved and the hypothesis has been validated through the work conducted and detailed in this thesis. The conclusions of this work are summarised in the next section. Furthermore, the future work section details areas of research which may advance the work in this thesis to improve range and intensity local features further.

## 8.2 Contributions

The key intellectual contribution of this thesis is to advance the understanding of how to best exploit range images and show how to best combine the information from range and intensity modalities for the purposes of feature extraction. The following contributions resulting from this work address the initial objectives and hypothesis:

The Major Contributions of this work are:

- Analysis and formulation of multimodal combinations of features, and feature structures.
- Element Corrected SIFT features and Range-Affine SIFT features, both of which advance the state of the art in local feature repeatability.
- A study of local feature integration in range images captured from multiple observation angles.
- View compartmentalised features, which allow the structure of keypoints to be view optimised.

The Minor Contributions of this work are:

- Benchmark experimental methodology for examining the effect of local feature matching between observations differing in pose by an out-of-plane rotation.

### 8.2.1 Multimodal SIFT features

Prior to this study SIFT features have been typically localised and described in a single imaging domain for creating repeatable local features. This thesis extends this approach by including information from both the range and intensity domains as a means to localise and formulate keypoint descriptions. The modularity of the SIFT processing pipeline allows the extraction process to be decomposed into the stages of scale space feature localisation and characterisation, which comprises canonical orientation sample patch correction, and feature descriptor extraction. The use of information from the range or intensity domain, or the use of information from both domains to perform the actions of each of these two stages, allows for the creation of 9 differing types of features extraction processes.

Cross modal features, combining range localisation and intensity descriptors, and intensity localisation and range descriptors, were formed by applying the extraction process in one domain and the characterisation process to the other. However, the combination of features which were localised in both domains or characterised in both domains presented two distinct questions of how to combine keypoint locations and how to combine keypoint characterisations. As a solution to these challenges a methodology was proposed, whereby the extraction process remains as two distinct stages and the combination of multimodal information is executed in each stage independently. For multimodal keypoint localisation, keypoints are localised in each domain and their resultant keypoint locations and scales are concatenated. For multimodal keypoint descriptors, keypoint locations had a canonical orientation estimated independently for each domain; a descriptor sample patch was aligned to the calculated canonical orientation in each domain; and a surface gradient measure is extracted to form the feature descriptor for each domain. The two feature descriptors both describing the same image location, however extracted from differing domains are then concatenated and renormalised to form a 256 element multimodal feature descriptor. This feature descriptor comprises a 128 element SIFT descriptor from each domain. The equal number of elements in the descriptors resulting from range and intensity, attributes equal weighting to information extracted from each domain.

The performance of the proposed cross modal and multimodal features for matching under out-of-plane pose variation was evaluated using the turn-table and stereo camera experimental configuration. From the analysis of the combinations of cross modal and multimodal features it was found that the optimum performance is achieved using a keypoint extraction process localising keypoints in both range and intensity, and applying a keypoint characterisation stage using information from the intensity domain images only. By the examination of other combinations of cross modal and multimodal keypoints it was found that range localised features typically gave better ROC performance for high sensitivity levels, although fewer keypoints are produced. Feature descriptors formed from range image surface gradients do not form as robust feature descriptors as expected.

### 8.2.2 Element Correction and Range Estimated Affine

The analysis of features extracted from co-aligned range and intensity images was extended to investigate an increased complexity in sample patch pose correction and surface measurements. This resulted in two sets of experiments, one to investigate the effect of sample patch pose correction approaches and another to investigate feature descriptor formation from a range of different surface measures.

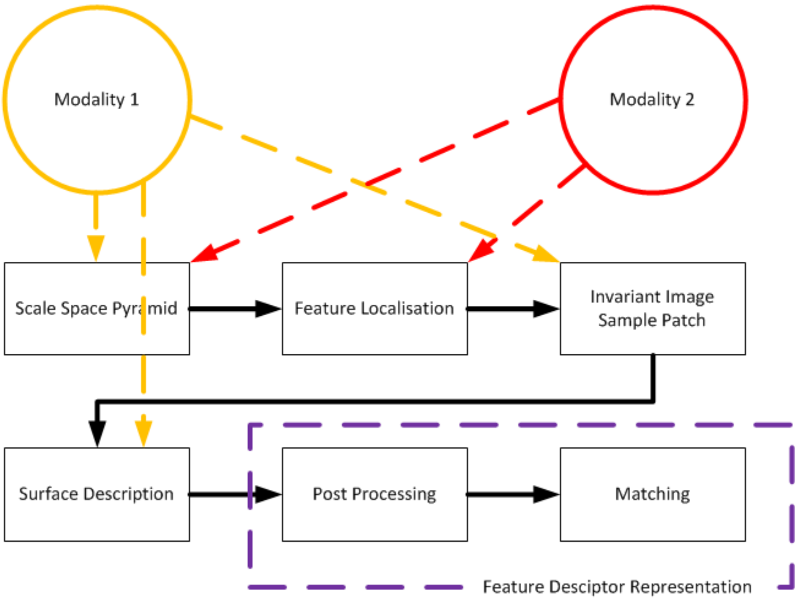


Figure 8.1: Multi and Cross Modal Feature Extraction Pipe-line

#### 8.2.2.1 Sample Patch Pose Correction

The sample patch pose correction approaches presented in this thesis investigates in-plane orientation invariance, affine correction, and full projective pose corrected sample patches. In addition to the level of sample patch pose correction applied, the modality from which the pose estimation of the sample patch was formed was also investigated.

The out-of-plane pose estimation of the sample patches was determined by means of a calculation of the local slant and tilt measures from either the range or intensity image values of a limited region surrounding the keypoint. Using these measures the sample patch correction approaches were applied. For the in-plane orientation invariance and affine pose correction, the pose corrected sample patches were established deterministically by applying the corrective transformation to the sample point locations. However, full projective pose correction required that the viewing axis of range surface was aligned with the surface normal and a rectilinear sample grid of sample points placed. The sample points were then projected back into the space of the original range and intensity images and then used as sample point locations for creating a feature descriptor.

The performance of the formulated pose corrected features was established using the calibrated turn-table and stereo camera configuration. The performance of feature extraction approaches were expected to increase as the level of out-of-plane correction was increased; affine features were expected to out perform in-plane corrected, and full pose corrected features were expected to increase

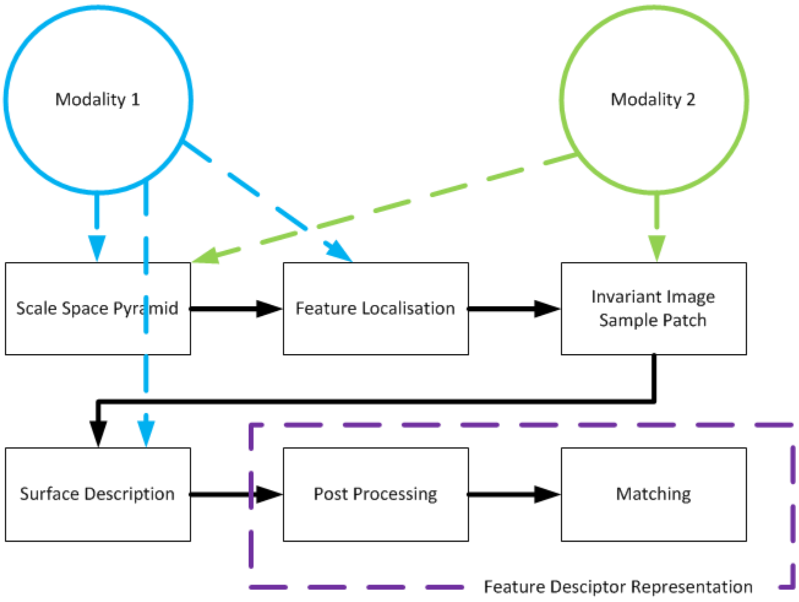


Figure 8.2: Sample Patch Corrected Feature Extraction Pipe-line

performance further. However, it was found that the optimum performance resulted from a local feature extraction approach comprising detection with descriptor extraction from the intensity images with an affine pose estimation and sample patch correction established from the range image. Conversely, range localised and range pose corrected features showed a decrease in performance. These results can be explained by the poor performance of the pose estimation approach in regions with high range variability, in these the pose estimation remains unstable, no matter how many range features are localised. The performance of intensity features may be explained by a more stable pose estimation from the surrounding range surface.

Full pose corrected features exhibited a further decrease in performance, this result may be explained by the non-linear placement of sample points on the image, such that the level in scale space where descriptors are extracted displays an inappropriate set of spatial frequencies. Additionally an instability in pose estimation applies a greater degree of patch shape for full pose corrected features; such that an error in pose estimation causes two patches describing a similar location to have a greater difference in shape, using full pose correction, than patches formed with an equivalent affine pose correction.

#### 8.2.2.2 Element Corrected Range SIFT

A range of surface descriptor extraction approaches were investigated, including the proposal of a novel range surface descriptor termed Element Corrected SIFT, that consists of a feature descriptor comprising a measure of surface gradients as for standard SIFT. However, the mean value across the whole sample patch is calculated and subtracted from the surface gradient calculated for each pixel in the patch. This has the effect of correcting the pose of the surface gradient measure for each pixel in the sample patch, without applying sample patch pose correction. This approach allows robust local features to be extracted from range images under changes in out-of-plane view point, for features localised around regions of high range variability. Initial experiments showed another measure, of Shape Index and Curvature, out performed Element Corrected SIFT features. However when the pixel-wise sample patch for both approaches was made equivalent, 9 pixels for each, Element Corrected SIFT features were found to exceed prior state of the art performance.

### 8.2.3 Multiview Integration

The investigations in Chapter 5 focus on the use of range and intensity information as a means to improve out-of-plane descriptor performance based on single observation of a keypoint. However, range information gives a partial representation of the 3D structure of an object, whereby features from multiple observations may be associated together to create a collection of possible descriptor expressions. To investigate the potential use of multiple observations, as a means to account for changes in keypoint feature descriptors as a result of out-of-plane pose change, a collection of keypoint observations from a range of views was used to form a model of the feature descriptor space of variation.

In order to achieve this, a set of 3D local interest points were selected around regions of high keypoint density in multiple object views. The descriptors from the keypoints associated with a 3D local interest point were used to form a measure of the mean descriptor and the variance within the group of descriptors. These measurements form the characterisation a 3D local interest point. To match any query keypoints to a model 3D local interest point, query descriptors are projected into the space of the model descriptors for each 3D local interest point where the match distance is then calculated. A range of statistical measures of variance were investigated to establish the most appropriate for characterising feature descriptor variations. Additionally, as descriptor variance increases with out-of-plane orientation changes the effect of increasing the range of keypoint poses was investigated.

The performance of 3D local interest point characterisations was investigated as keypoint precision versus recall for synthetic range images with observation changes along the pitch and yaw axes. In the experiments conducted it was found that the variation in the feature descriptor corpus for a 3D local interest point was best characterised using a PCA space for each 3D local interest point; the optimum performance for the range of keypoint observations was found when the inclusion of keypoints was limited to those with a surface normal less than $40°$off the viewing axis. However, the performance of PCA features showed only a marginal improvement on the standard Euclidean distance. Additionally, the structure of the 3D local interest point localiser and descriptor characterisation approaches prohibit the use of the multiview integration approach presented in this thesis from applications which require on-line learning, such as robotics.

### 8.2.4 View Compartmentalised Keypoints

View Compartmentalised keypoints were proposed in Chapter 7, these features comprise an additional processing step in the SIFT processing pipe-line which allows the resulting feature descriptor to encapsulate a measure of the 3D appearance of the keypoint, and the observation angles at which the keypoint has been observed. Extracted SIFT features are stored in a series of keypoint view compartments, the 3D pose of the keypoint is used to assign a weighting value to each compartment, representing the confidence which may be invested in a descriptor match in any given compartment. During a learning phase the system explores the view sphere of an object, subsequent observations of a keypoint may be used to update compartments in a master keypoint when a compartment with a higher confidence becomes available. In order to match any pair of keypoints, a weighted Euclidean distance was formulated and validated. The weighted Euclidean distance finds Euclidean distances between the descriptors from all compartments and weights the combination of these by the product of the confidence scores for the corresponding compartments. This allows the confidence in the descriptor match between two compartments to be factored into the keypoint matching approach.

The system introduces three novel stages to the extraction and matching of SIFT features: the weighted Euclidean distance, the compartment weighting and the compartment assignment. The weighted Euclidean distance was validated through matching examples of synthetic descriptors where weighting values were set, this allowed the resultant matches to be compared to the expected matches. The weighted Euclidean distance performed well on the synthetic keypoints and gave an indication that matches obtained using this approach would be sufficiently robust. The compartment assignment was investigated through an example of exploring a view hemisphere of a keypoint in synthetic range images. In this experiment, incorrect example range images were stored as the closest match for each compartment, these images tended to be at the extremes of the view hemisphere, indicating that pose instability around occlusion boundaries significantly affects the choice of compartment assignment. Additionally the keypoint pose experiments showed a significant instability in 3D in-plane pose estimation from the range image, where ambiguities of $\pm 90^\circ$ and $180^\circ$ were observed. The observed pose instability affects the compartment weight, having the effect of incorrectly labelling the structure of the compartments with erroneous weighting scores, and thereby giving erroneous descriptor matching. The ROC curves for view compartmentalised features were compared with their non-view compartmentalised counterparts, for the cases of 9 view compartments and 3 view compartments, in both cases of view compartmentalisation an improvement was observed using standard SIFT descriptors. However, the lack of sufficiently robust 3D pose estimation from range images limits the effectiveness of the view compartmentalised keypoints.

## 8.3 Future Work

The work presented in this thesis furthers the state of the art in local feature matching performance using combinations of range and intensity imaging modalities. This section suggests research routes whereby further developments, which may improve the repeatability of individual range and intensity local features are identified. In addition to generalising the approaches outlined in this thesis to other applications of local features.

### 8.3.1 Evaluation

The evaluation approach outlined in this thesis is based upon matching a query image to the nearest target image. This approach limits the number of target keypoints in the database to which a query keypoint must match. An alternative approach might extend the search space by matching the query keypoints to all target keypoints in the database. However, it is expected that this more expensive approach would produce fewer correct keypoint matches per query image, requiring an increase in the number of iterations of the experiment in order to gain a sufficient sample of keypoint matches to allow a ROC curve to be constructed for each keypoint extraction process. The resultant ROC curves are predicted to show decreased ROC performance for all proposed approaches when compared with the current experimental design, in accordance with those results observed by Moreels and Perona (Moreels and Perona, 2007).

#### 8.3.1.1 Illumination and Clutter

The experiments outlined in Chapter 4 use a lighting source either side of the stereo camera configuration, this subsequently limits the effects of illumination to be explored only in conjunction with a 3D out-of-plane rotation. Other experimental designs such as the Coil-100 database have used a series of lamps which may be selected to create a range of illuminations patterns for an object observed at a single pose (Nene et al., 1996). To implement similar experimental design using the approach adopted in this thesis would require a stereo matching approach robust to illumination changes, in addition to a formalisation of the range of illumination changes which will be investigated. A more robust stereo matching approach may be achieved using range image processing approaches such as Adaptive Surface Smoothing (Sun et al., 2002), Anisotropic diffusion (Weickert, 1998) or energy minimisation constraints (Li, 1992).

### 8.3.2 Range SIFT

A number of SIFT variations which utilising range domain information were proposed and investigated in this thesis. This section details potential future work which could be conducted to improve the work outlined in this thesis.

#### 8.3.2.1 Pose Estimation in Range Images

Large errors were observed when tracking the pose of local features between range images between observations. Many of the range based approaches presented in this thesis rely on accurate measurements of keypoint surface normals based on range images. Therefore, improvements in the performance of all of these approaches can be achieved through greater accuracy in pose estimation

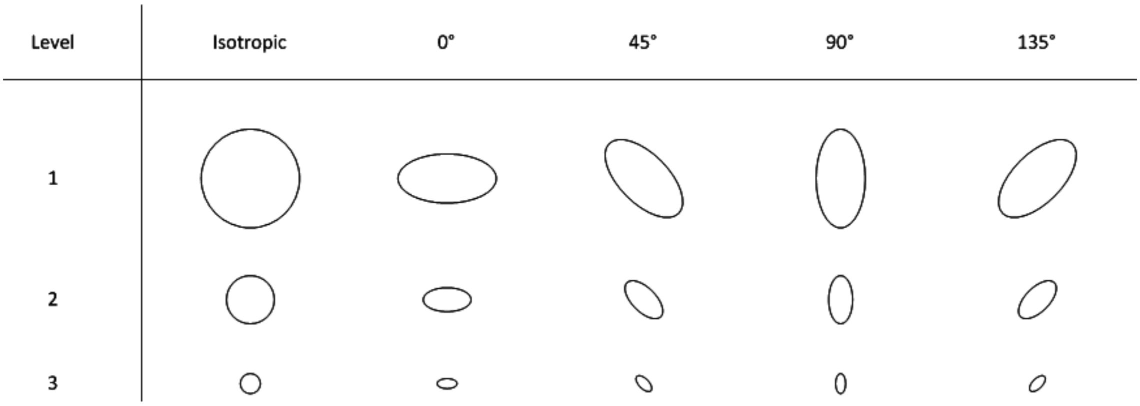


Figure 8.3: Gaussian Blur Functions for Affine Scale Space Pyramid

from the range surface.

#### 8.3.2.2 Normalisation

The normalisation of 2.5D range values in order to create a descriptor of unit length removes diagnostic information regarding the local surface surrounding interest points in range images. To account for the loss of information resulting from the normalisation stage, a further investigation into how to encode the local range surface variation into local features while allowing feature descriptors to be normalised to unit length is required.

#### 8.3.2.3 Affine Scale Space Pyramid

The sample patch warping approaches outlined in Chapter 5.5.2 were shown to compress the axes of the sampling patch, thereby changing the spatial sampling frequency in the direction of compression. This effect may be accounted for by introducing an *affine scale space pyramid*. The affine scale space pyramid should comprise a scale space pyramid in which each level includes additional images formed with elliptical Gaussian blur functions, see Figure 8.3. When applying an affine warp to a sample patch, an appropriate affine scale space image can therefore be selected in which the effect of the axis compression of the sample patch is minimised.

### 8.3.3 View Compartmentalised SIFT

The performance of view compartmentalised SIFT keypoints relies on correctly establishing the pose of the keypoint in the given observation for compartment assignment and the associated pose weighting. Therefore, as an approach to increasing the performance of view compartmentalised SIFT keypoints a more accurate pose estimation stage is required, as recommended in the previous section. However, this section details a number of applications for view compartmentalised SIFT keypoints.

#### 8.3.3.1 Multiview View Compartmentalised Keypoints

Local features are frequently used as an approach for building multiview representations from sparse collections of points. However, the source images frequently include large baseline separations, resulting in large change in the feature appearance between observations. As an approach to reducing the feature descriptor distance between keypoint correspondences a view compartmentalised feature descriptor methodology could be adopted. This approach may use affine pose estimations for structuring the compartments and the level of warping applied to create the compartment weighting function for the feature. Matching between keypoint instances can be established using the weighted Euclidean matching scheme outlined in Chapter 7.

#### 8.3.3.2 BOF compartmentalised SIFT

Bag-Of-Features approaches to image analysis for object instance recognition and scene interpretation have recently become popular among the vision community (Sivic et al., 2005; Fei-Fei and Perona, 2005). Ohbuchi et al have proposed a method by which range features can be extracted to from a Bag-Of-Features representation utilising the range domain representations of an object captured from multiple observations (Ohbuchi et al., 2008). However, this approach may be extended using view compartmentalised SIFT descriptors which more explicitly encode the local range structure in the weighting vector. The compartment weighting vector can be used to create a weighted vote in a bag-of-features approach, which reflects the encoding of the keypoint appearance. Additionally, where both range and intensity are available, such as in many robotics applications, view compartmentalisation based on range measurements may be performed on extracted intensity features. Lai et al recently present a range and intensity image repository of 300 instances of objects in varying poses captured with a Microsoft Kintect (Lai et al., 2011). This repository may be used to investigate the potential application of view compartmentalised features in a bag-of-words model.

## 8.4 Concluding Remarks

This thesis investigates the combination of range and intensity domain information to formulate robust and highly distinctive local features. The work has shown the potential for multimodal RGB-D representations to improve local feature matching under changes in observer location. Additionally, work on View Compartmentalised features has introduced an approach for combining a variety of keypoint observations together to form a single keypoint with differing confidences associated with each of its constituting descriptions. The above future work section highlights existing computer vision algorithms which may be used in conjunction with the techniques developed in this dissertation to further advance the state-of-the-art. Vision systems based on these techniques may find applications in clinical landmarking, robotic navigation, manipulation, and inspection.

# Bibliography

References are included with the referencing section numbers appended following the entry details

Dimensional Imaging Ltd @http://www.di3d.com/index.php, Dec 2008. 2.2.1.4

Andrea F Abate, Michele Nappi, Daniel Riccio, and Gabriele Sabatino. 2D and 3D face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885–1906, 2007. 2.3.1.1, 5.2.1

Y. I. Abdel-Aziz and H. M. Karara. Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. In *Symposium on Close-range Photogrammetry*, pages 1–18, Falls Church, VA, 1971. American Society of Photogrammetry. 3

Ankur Agarwal and Bill Triggs. Hyperfeatures–multilevel local coding for visual recognition. In *European Conference on Computer Vision*, pages 30–43. Springer, 2006. 2.3.2.2, 2.3.2.5

Motilal Agrawal, Kurt Konolige, and Morten Blas. Censure: Center surround extremas for realtime feature detection and matching. In *European Conference on Computer Vision*, pages 102–115. Springer, 2008. 6.2.1

Erdem Akagunduz and Ilkay Ulusoy. Extraction of 3D transform and scale invariant patches from range scans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 5.2.1

Alberto Albiol, David Monzo, Antoine Martin, Jorge Sastre, and Antonio Albiol. Face recognition using HOG–EBGM. *Pattern Recognition Letters*, 29(10):1537–1543, 2008. 2.3.1.3

Gerardo Aragon-Camarasa and J. Paul Siebert. Unsupervised clustering in Hough space for recognition of multiple instances of the same object in a cluttered scene. *Pattern Recognition Letters*, 31(11):1274–1284, August 2010. 4.3.2.2

Jürgen Assfalg, Marco Bertini, A Del Bimbo, and Pietro Pala. Content-based retrieval of 3-D objects using SPIN image signatures. *IEEE Transactions on Multimedia*, 9(3):589–599, 2007. 5.4.4

Indriyati Atmosukarto, Katarzyna Wilamowska, Carrie Heike, and Linda G Shapiro. 3D object classification using salient point patterns with application to craniofacial research. *Pattern Recognition*, 43(4):1502–1517, 2010. 2.3.2.4, 3.3.4, 5.4.2

D.H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981. 2.3.2.5, 4.3.2

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. 2.3.1, 2.3.1.2, 2.3.2.1, 2.3.2.3, 2.3.2.4

Neslihan Bayramoglu and A Aydin Alatan. Shape index sift: Range image recognition using local features. In *IEEE International Conference on Pattern Recognition*, pages 352–355, 2010. 3.3.4

Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 2.3.1.3

Serge Belongie and Jitendra Malik. Finding boundaries in natural images: A new method using point descriptors and area completion. In *European Conference on Computer Vision*, pages 751–766. Springer, 1998. 2.3.1

Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002. 2.3.1.3, 2.3.2.4

Paul J Besl and Ramesh C Jain. Invariant surface characteristics for 3D object recognition in range images. *Computer Vision, Graphics, and Image Processing*, 33(1):33–80, 1986. 1.1, 2.2.4, 2.3.1.1

P.J. Besl and N.D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. 2.3.1.1

Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115, 1987. 2.3.1

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 6.1.1.3, 6.1.3, 6.1.3.1

Matthew Blaschko and Christoph Lampert. Learning to localize objects with structured output regression. In *European Conference on Computer Vision*, pages 2–15. Springer, 2008. 2.3.1

Chris Boehnen and Patrick Flynn. Impact of involuntary subject movement on 3d face scans. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2009. 2.2.1.1

Kevin W Bowyer, Kyong Chang, and Patrick Flynn. A survey of approaches and challenges in 3D and multi-modal 3D+ 2D face recognition. *Computer Vision and Image Understanding*, 101(1): 1–15, 2006. 1.1.2, 2.3.1.1, 2.3.2.4, 2.4

M. Brown and D.G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007. 2.3.2.5

Matthew Brown and David G Lowe. Invariant features from interest point groups. In *British Machine Vision Conference*, volume 21, pages 656–665, 2002. 3.3.1

Matthew Brown, Richard Szeliski, and Simon Winder. Multi-image matching using multi-scale oriented patches. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 510–517, 2005. (document), 2.10

Matthew Brown, Gang Hua, and Simon Winder. Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):43–57, 2011. 1.2.1, 2.3.1

Gertjan J Burghouts and Jan-Mark Geusebroek. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 113(1):48–62, 2009. 1

GJ Burton, ND Haig, and IR Moorhead. A self-similar stack model for human and machine vision. *Biological Cybernetics*, 53(6):397–403, 1986. 2.3.1

Benjamin Bustos, Daniel A Keim, Dietmar Saupe, Tobias Schreck, and Dejan V Vranić. Feature-based similarity search in 3D object databases. *ACM Computing Surveys (CSUR)*, 37(4):345–387, 2005. 2.2

Michael Calonder, Vincent Lepetit, and Pascal Fua. Keypoint signatures for fast learning and recognition. In *European Conference on Computer Vision*, pages 58–71. Springer, 2008. 1.1.2, 2.2, 2.2.1.1

Richard J Campbell and Patrick J Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, 81(2):166–210, 2001. 2.3.1.1, 2.3.2.4

John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):679–698, 1986. 2.3.1

N. Canterakis. 3D Zernike moments and Zernike Affine invariants for 3D image analysis and recognition. In *Scandinavian Conference on Image Analysis*, pages 85–93, 1999. 2.3.1.1

Chin-Seng Chua, Feng Han, and Yeong-Khing Ho. 3D human face recognition using point signature. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 233–238, 2000. 2.3.1.1, 2.3.1.2

Jinyun Chung, Taemin Kim, Yeong Nam Chae, and Hyun S Yang. Unsupervised constellation model learning algorithm based on voting weight control for accurate face localization. *Pattern Recognition*, 42(3):322–333, 2009. 2.3.1.3

Ken Conley. Robotic Operating System Wiki @http://www.ros.org/wiki/, Jul 2012. 2.2.1.3

Timothy F Cootes, Christopher J Taylor, David H Cooper, Jim Graham, et al. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. 2.3.1.3, 2.11b

Timothy F Cootes, G Edwards, and Christopher J. Taylor. Active appearance models. In *European Conference on Computer Vision*, pages 484–498. Springer, 1998. 2.11a

Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001. 2.3.1.3

Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000. 6.1.3

Yan Cui, Sebastian Schuon, Derek Chan, Sebastian Thrun, and Christian Theobalt. 3D shape scanning with a time-of-flight camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1173–1180, 2010. 2.2.1.2

B. Cyganek and J.P. Siebert. *An introduction to 3D computer vision techniques and algorithms*. Wiley, 2011. 2.2.1, 2.2.2, 4.3.1.2, 4.4.1, 5.2.1

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005. 2.3.2.2, 2.3.2.5

M Dekiff, P Berssenbrügge, B Kemper, C Denz, and D Dirksen. Three-dimensional data acquisition by digital correlation of projected speckle patterns. *Applied Physics B: Lasers and Optics*, 99(3): 449–456, 2010. 4.3.1.2

Chitra Dorai and Anil K. Jain. COSMOS-A representation scheme for 3D free-form objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(10):1115–1130, 1997. 2.3.1.1, 2.3.2.4, 5.4.2, 6.1.1

Itiel E Dror, Mark Zagaeski, and Cynthia F Moss. Three-dimensional target recognition via sonar: a neural network model. *Neural Networks*, 8(1):149–160, 1995. 2.2

Richard O Duda and Peter E Hart. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972. 2.3.1

R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification and Scene Analysis 2nd ed.* 1995. 4.3.2

Thomas F. El-Maraghi. Matlab SIFT Code @http://www.cs.toronto.edu/ tem/research.html, Dec 2008. 3.1, 3.3

Timothy C Faltemier, Kevin W Bowyer, and Patrick J Flynn. Using a multi-instance enrollment representation to improve 3D face recognition. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6, 2007. 1.1, 2.2.1.1

Timothy C Faltemier, Kevin W Bowyer, and Patrick J Flynn. Using multi-instance enrollment to improve performance of 3D face recognition. *Computer Vision and Image Understanding*, 112 (2):114–125, 2008. 2.3.1.1

Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-D audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010. 2.2

Olivier Faugeras. *Three-dimensional computer vision: a geometric viewpoint*. MIT Press, Cambridge, MA, USA, 1993. 2.3.1

Olivier D Faugeras and Martial Hebert. The representation, recognition, and locating of 3-D objects. *International Journal of Robotics Research*, 5(3):27–52, 1986. 2.3.1

T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. (document), 3.2.4, 3.2.4.1, 3.13

L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 524–531, 2005. 2.3.2.5, 2.4, 8.3.3.2

Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. 2.3.1.3

Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. 2.3.1.3

Rob Fergus, Pietro Perona, and Andrew Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–275, 2003. 2.3.1.3

Virgilio F Ferrario, Chiarella Sforza, Carlo E Poggio, Massimiliano Cova, and Gianluca Tartaglia. Preliminary evaluation of an electromagnetic three-dimensional digitizer in facial anthropometry. *The Cleft Palate-craniofacial Journal*, 35(1):9–15, 1998. 5.2.2

W. Freeman, P. Perona, and B. Scholkopf. Guest Editorial: Machine Learning in Computer Vision. *International Journal of Computer Vision*, 2008. 2.1, 2.3.1

Andrea Frome, Daniel Huber, Ravi Kolluri, Thomas Bülow, and Jitendra Malik. Recognizing objects in range data using regional point descriptors. In *European Conference on Computer Vision*, pages 224–237. Springer, 2004. 2.2.4, 2.3.1.1, 5.4.4

Takahiko Furuya and Ryutarou Ohbuchi. Dense sampling and fast encoding for 3D model retrieval using bag-of-visual features. In *ACM International Conference on Image and Video Retrieval*, page 26, 2009. 7.2.1

Bernat Gabor. Camera calibration With OpenCV @http://opencv.itseez.com/trunk/doc/ tutorials/calib3d/camera_calibration/camera_calibration.html, Dec 2011. 4.3.1.1

Davi Geiger, Bruce Ladendorf, and Alan Yuille. Occlusions and binocular stereo. *International Journal of Computer Vision*, 14(3):211–226, 1995. 2.2.4

Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd edition, 1992. 2.1, 2.3.1.2

Raghuraman Gopalan, Pavan Turaga, and Rama Chellappa. Articulation-invariant representation of non-planar shapes. In *European Conference on Computer Vision*, pages 286–299. Springer, 2010. 2.3.1.3

Gaile G Gordon. Face recognition based on depth and curvature features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 808–810, 1992. 1.1, 1.1.2, 2.3.1.1

He Guo, Kai Zhang, and Qi Jia. 2.5 D SIFT Descriptor for Facial Feature Extraction. In *IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 723–726, 2010. 3.3.4

Raj Gupta, Harshal Patil, and Anurag Mittal. Robust order-based methods for feature description. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 334–341, 2010a. 2.3.2.4

S. Gupta, M.P. Sampat, M.K. Markey, A.C. Bovik, and Z. Wang. Facial range image matching using the complexwavelet structural similarity metric. In *Workshop on Applications of Computer Vision*, pages 4–4. IEEE, 2007. 2.3.2.2, 2.3.2.4

Shalini Gupta, Kenneth R Castleman, Mia K Markey, and Alan C Bovik. Texas 3D face recognition database. In *Southwest Symposium on Image Analysis & Interpretation*, pages 97–100. IEEE, 2010b. 2.3b, 2.2.1.4, 4.4.1

R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*, volume 2. Cambridge Univ Press, 2005. (document), 1.1.2, 2.4, 2.2.2, 3, 4.3.1.2

James Hays and Alexei A Efros. IM2GPS: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2.2

Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid. Description of interest regions with local binary patterns. *Pattern Recognition*, 42(3):425–436, 2009. 2.3.2.4

Gunter Hetzel, Bastian Leibe, Paul Levi, and Bernt Schiele. 3D object recognition from range images using local feature histograms. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–394, 2001. 2.3.1.1, 2.3.2.4, 3.3.4, 5.4.2

Adam Hoover, Gillian Jean-Baptiste, Xiaoyi Jiang, Patrick J. Flynn, Horst Bunke, Dmitry B. Goldgof, Kevin Bowyer, David W. Eggert, Andrew Fitzgibbon, and Robert B. Fisher. An experimental comparison of range image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):673–689, 1996. 2.3.1

Berthold KP Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, 4(4):629–642, 1987. 4.3.2.2

J Illingworth and J Kittler. A survey of efficient Hough Transform methods. In *Third Alvey Vision Conference*, pages 319–326, 1987. 2.3.1

S Islam, Mohammed Bennamoun, Ajmal Mian, and R Davies. Score level fusion of ear and face local 3D features for fast and expression-invariant human recognition. In *International Conference on Image Analysis and Recognition*, pages 387–396. Springer, 2009. 2.3.1.1

Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. A category-level 3-d object dataset: Putting the kinect to work. In *IEEE International Conference on Computer Vision Workshops*, pages 1168–1174, 2011. 1.1.2, 1.2.1

Andrew E. Johnson and Martial Hebert. Using SPIN images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5): 433–449, 1999. 2.3.1.2, 2.3.2.3, 5.4.4

Andrew Edie Johnson. *SPIN-images: A representation for 3-d surface matching*. PhD thesis, Carnegie Mellon University, 1997. 2.3.1.1, 2.3.1.2, 5.4.4

Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2005. 6.1.3, 6.1.3.1

Judson P Jones and Larry A Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987. 2.3.1

X. Ju, T. Boyling, and P. Siebert. A high resolution stereo imaging system. *3D Modelling*, 2003. 1.3, 2.2.2, 1, 2, 4.3.1.1, 4.4, 4.4.1

Takeo Kanade. *Three-dimensional machine vision*. Springer, 1987. 2.3.1

S. Katz, A. Tal, and R. Basri. Direct visibility of point sets. In *ACM Transactions on Graphics (TOG)*, volume 26, page 24, 2007. 5.3.3, 6.2.1

Yan Ke and Rahul Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 506–513, 2004a. 2.3.1.2

Yan Ke and Rahul Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 506–513, 2004b. 2.3.2.5, 6.1.3.1

K. Khoshelham. Extending generalized hough transform to detect 3d objects in laser range data. In *ISPRS Workshop on Laser Scanning, Proceedings, LS 2007*, pages 206–210, 2007. 4.3.2, 4.3.2.2

Gudrun J Klinker, Steven A Shafer, and Takeo Kanade. A physical approach to color image understanding. *International Journal of Computer Vision*, 4(1):7–38, 1990. 2.3.1

KH Ko, T Maekawa, NM Patrikalakis, H Masuda, and F-E Wolter. Shape intrinsic fingerprints for free-form object matching. In *ACM Symposium on Solid Modeling and Applications*, pages 196–207, 2003. (document), 2.7, 2.3.1.1

Jan J Koenderink and Andrea J Van Doorn. Receptive field assembly pattern specificity. *Journal of Visual Communication and Image Representation*, 3(1):1–12, 1992. 2.3.1

J.J. Koenderink and A.J. van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10(8):557–564, 1992. (document), 2.3.1.1, 2.8, 3.3.4, 5.4.2

Jonathan Kofman and George K Knopf. Point correspondences between successive range views using localized SPIN images. In *International Society Conference on Optical Engineering*, volume 3837, pages 289–298, 1999. 1.1.2

Iasonas Kokkinos, Michael M Bronstein, Roee Litman, and Alex M Bronstein. Intrinsic shape context descriptors for deformable shapes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 159–166, 2012. 2.3.1.3

Venkat Krishnamurthy and Marc Levoy. Fitting smooth surfaces to dense polygon meshes. In *ACM Conference on Computer Graphics and Interactive Techniques*, pages 313–324, 1996. 2.2, 2.3.1.1

K. Lai and D. Fox. 3D laser scan classification using web data and domain adaptation. In *RBO International Conference on Robotics: Science and Systems*, 2009. 5.4.4

Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A scalable tree-based approach for joint object and pose recognition. In *AAAI Conference on Artificial Intelligence*, 2011. (document), 1.2.1, 2.2.1.3, 2.3.2.5, 2.18, 8.3.3.2

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005. 2.3.1.2

Daniel D Lee, HSebastian Seung, et al. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 2.3.1.3

Sukhan Lee, Zhaojin Lu, and Hyunwoo Kim. Probabilistic 3D object recognition with both positive and negative evidences. In *IEEE International Conference on Computer Vision*, pages 2360–2367, 2011. 2.3.2.5

Stan Z Li. Toward 3D vision from range images: An optimization framework and parallel networks. *CVGIP: Image Understanding*, 55(3):231–260, 1992. 8.3.1.1

Z Lian, A Godil, B Bustos, M Daoudi, Jeroen Hermans, S Kawamura, Y Kurita, G Lavoue, H Nguyen, R Ohbuchi, et al. SHREC11 track: shape retrieval on non-rigid 3D watertight meshes. In *Eurographics Association Workshop on 3D Object Retrieval*, volume 11, pages 79–88, 2011. 6.1.1

T. Lindeberg. *Scale-space theory in computer vision.* Springer, 1993. 2.3.2.1

Haibin Ling, Xingwei Yang, and Longin Latecki. Balancing deformability and discriminability for shape matching. In *European Conference on Computer Vision*, pages 411–424. Springer, 2010. 2.3.1.1, 2.3.1.3

C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011. 2.3.2.5

Jianqing Liu and Yee-Hong Yang. Multiresolution color image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(7):689–700, 1994. 2.3.1

Rachel Lo. *Feature extraction for range image interpretation using local topology statistics.* PhD thesis, University of Glasgow, UK, 2009. 1.1.2, 3.3.2

T.W.R. Lo and J.P. Siebert. SIFT keypoint descriptors for range image analysis. *Annals of the BMVA*, pages 1–17, 2008. 2.3.2.4, 5.4.2

T.W.R. Lo and J.P. Siebert. Local feature extraction and matching on range images: 2.5 D SIFT. *Computer Vision and Image Understanding*, 113(12):1235–1250, 2009. 1.1.2, 2.3.1, 2.3.1.1, 2.3.2.3, 2.3.2.4, 2.4, 3.3.2, 2, 3.3.4, 5.3, 5.3.2, 5.4.2

David Lowe. Demo SIFT Code @http://www.cs.ubc.ca/ lowe/keypoints, Dec 2008. 1

David G Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1987. 2.3.1

David G Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE, 1999. 1.3, 2.3.1, 2.3.1.2, 2.3.2, 2a, 7.3

D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. (document), 1.3, 2.3.1, 2.3.1.2, 2.3.2, 2.3.2.1, 2.3.2.3, 2.3.2.5, 2.4, 3.1, 3.2.1, 3.2, 3.2.4, 3.3.1, 4.3.2.1, 2a, 5.4.1, 7.3

TC Lukins and RB Fisher. Qualitative characterization of deforming surfaces. In *International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 287–293. IEEE, 2006. 2.3.2.4, 5.4.2

S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. 2.3.2.1

Alan P. Mangan and Ross T. Whitaker. Partitioning 3D surface meshes using watershed segmentation. *Visualization and Computer Graphics, IEEE Transactions on*, 5(4):308–321, 1999. 6.1.1

Jiri Matas, Ondrej Chum, Martin Urban, and Tomáš Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004. (document), 2.3.2.2, 2.14

Bogdan J Matuszewski, Wei Quan, and L-K Shark. High-resolution comprehensive 3-D dynamic database for facial articulation analysis. In *IEEE International Conference on Computer Vision Workshops*, pages 2128–2135, 2011. 2.2.1.4, 2.3.1.3

G.J. McLachlan and J. Wiley. *Discriminant analysis and statistical pattern recognition*, volume 5. Wiley Online Library, 1992. 6.1.3, 6.1.3.2

David Meger, Per-Erik Forssén, Kevin Lai, Scott Helmer, Sancho McCann, Tristram Southey, Matthew Baumann, James J Little, and David G Lowe. Curious George: An attentive semantic robot. *Robotics and Autonomous Systems*, 56(6):503–511, 2008. 2.1

Xing Mei, Xun Sun, Mingcai Zhou, Shaohui Jiao, Haitao Wang, and Xiaopeng Zhang. On building an accurate stereo matching system on graphics hardware. In *IEEE International Conference on Computer Vision Workshops*, pages 467–474, 2011. 2.2.1.4, 2

Ajmal S Mian, Mohammed Bennamoun, and Robyn Owens. An efficient multimodal 2D-3D hybrid approach to automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1927–1943, 2007. 2.3.1.1

Ajmal S. Mian, Mohammed Bennamoun, and Robyn Owens. Keypoint Detection and Local Feature Matching for Textured 3D Face Recognition. *International Journal of Computer Vision*, 79(1): 1–12, 2008. 1.1.2

K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004. (document), 1.2.1, 2.3.2.3, 2.17, 5.3.2, 5.6.2

K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005. 1.2.1, 2.3.1, 2.3.1.2, 2.3.2.2, 2.3.2.3, 2a

K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L.V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1):43–72, 2005. 5.3.2, 5.6.2

Pierre Moreels and Pietro Perona. Common-frame model for object recognition. In *Neural Information Processing Systems*, pages 953–960, 2004. 1.2.1

Pierre Moreels and Pietro Perona. Evaluation of features detectors and descriptors based on 3D objects. *International Journal of Computer Vision*, 73(3):263–284, 2007. 1.2.1, 2.3.2.2, 8.3.1

Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *IEEE International Conference on Computer Vision Theory and Applications*, pages 331–340, 2009. 2.3.2.5

Hiroshi Murase and Shree K Nayar. Learning object models from appearance. In *National Conference on Artificial Intelligence*, pages 836–836. John Wiley & Sons Ltd, 1993. 2.3.1

Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). *Internal Report CUCS-005-96, Columbia University Computer Science*, 1996. 8.3.1.1

Marcin Novotni and Reinhard Klein. 3D Zernike descriptors for content based shape retrieval. In *ACM Symposium on Solid Modeling and Applications*, pages 216–225, 2003. 2.3.1.1

Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*, pages 490–503. Springer, 2006. 2.3.2.2

Ryutarou Ohbuchi and Takahiko Furuya. Scale-weighted dense bag of visual features for 3d model retrieval from a partial view 3d model. In *IEEE International Conference on Computer Vision Workshops*, pages 63–70, 2009. 1.1.2

Ryutarou Ohbuchi, Kunio Osada, Takahiko Furuya, and Tomohisa Banno. Salient local visual features for shape-based 3D model retrieval. In *IEEE International Conference on Shape Modeling and Applications*, pages 93–102, 2008. 1.1.2, 6.1.1, 8.3.3.2

Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996. 2.3.2.4

Nikhil R Pal and Sankar K Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993. 2.3.1

S.E. Palmer. *Vision science: Photons to phenomenology*, volume 1. MIT Press Cambridge, MA, 1999. 2.2, 6.1

Martin P Parsley and Simon J Julier. Avoiding negative depth in inverse depth bearing-only SLAM. In *IEEE International Conference on Intelligent Robots and Systems*, pages 2066–2071, 2008. 2.2

Nick Pears, Tom Heseltine, and Marcelo Romero. From 3d point clouds to pose-normalised depth maps. *International Journal of Computer Vision*, 89(2):152–176, 2010. 1.1, 1.1.2, 2.3.1.1

Gabriele Peters, Barbara Zitova, and Christoph Von der Malsburg. How to measure the pose robustness of object views. *Image and Vision Computing*, 20(4):249–256, 2002. 7.2, 7.6

Jean Ponce, Tamara Berg, Mark Everingham, David Forsyth, Martial Hebert, Svetlana Lazebnik, Marcin Marszalek, Cordelia Schmid, Bryan Russell, Antonio Torralba, et al. Dataset issues in object recognition. In *Toward category-level object recognition*, pages 29–48. Springer, 2006. 2.2, 2.3.1

Massimiliano Pontil and Alessandro Verri. Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):637–646, 1998. 2.3.1, 2.4

A.R. Pope and D.G. Lowe. Probabilistic models of appearance for 3-D object recognition. *International Journal of Computer Vision*, 40(2):149–167, 2000. 2.3.2.5, 4.3.2.1

Mark W Powell, Kevin W Bowyer, Xiaoyi Jiang, and Horst Bunke. Comparing curved-surface range image segmenters. In *IEEE International Conference on Computer Vision*, pages 286–291, 1998. 2.3.1

ID Reid and JM Brady. Recognition of object classes from range data. *Artificial Intelligence*, 78 (1):289–326, 1995. 2.3.1

Lawrence G Roberts. Machine perception of three-dimensional solids. Technical report, Massachusetts Institute of Technology Lexington Lincoln Lab, 1963. 2.3.1

Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). In *IEEE International Conference on Robotics and Automation*, pages 1–4, 2011. 2.2, 2.2.1.3

Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1):7–42, 2002. 2, 2.4

B. Schiele and J.L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000. 1.3

Cordelia Schmid and Roger Mohr. Combining greyvalue invariants with local constraints for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 872–877, 1996. 2.3.1, 2.3.1.2

Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000. 1.2.1

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 2.3.1

J Paul Siebert and Stephen J Marshall. Human body 3D imaging by speckle texture projection photogrammetry. *Sensor Review*, 20(3):218–226, 2000. 1.1.2, 2.2.2, 4.3.1.2, 4.4.1

J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, pages 1470–1477, 2003. 2.3.1, 2.3.2.5, 2.4

J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman. Discovering objects and their location in images. In *IEEE International Conference on Computer Vision*, volume 1, pages 370–377, 2005. 2.4, 8.3.3.2

Michael Stark, Michael Goesele, and Bernt Schiele. Back to the future: Learning shape models from 3d cad data. In *British Machine Vision Conference*, pages 1–11, 2010. 2.3.1.3

F. Stein and G. Medioni. Structural indexing: efficient 3-D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):125 –145, 1992. 2.3.1.1

Y. Sun, D.L. Page, J.K. Paik, A. Koschan, and M.A. Abidi. Triangle mesh-based edge detection and its application to surface segmentation and adaptive surface smoothing. In *IEEE International Conference on Image Processing*, volume 3, pages 825–828, 2002. 8.3.1.1

Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010. (document), 2.1, 2.3.2.2, 2.13

Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):1068–1080, 2008. 2

Babak Taati and Michael Greenspan. Local shape descriptor selection for object recognition in range data. *Computer Vision and Image Understanding*, 115(5):681–694, 2011. 1.1.2

Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, 2010. 2.3.1.2, 2.9a, 2.3.2.3

Federico Tombari and Luigi Di Stefano. Object recognition in 3D scenes with occlusions and clutter by Hough voting. In *Image and Video Technology. Fourth Pacific-Rim Symposium on*, pages 349–355. IEEE, 2010. 4.3.2, 4.3.2.2, 4.3.2.2

Emanuele Trucco and Alessandro Verri. *Introductory techniques for 3-D computer vision*, volume 93. Prentice Hall Upper Saddle River, 1998. 2.2.2

Greg Turk and Marc Levoy. Zippered polygon meshes from range images. In *ACM Conference on Computer graphics and interactive techniques*, pages 311–318, 1994. 2.2

Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3 (1):71–86, 1991. 2.3.1.3

T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008. 2.3.1.2, 2.3.1.3, 2.4

Colin Urquart. *The active stereo probe: The design and implementation of an active videometrics system*. PhD thesis, University of Glasgow, UK, 1997. 2.2.2, 4.4.1

Koen EA Van De Sande, Theo Gevers, and Cees GM Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9): 1582–1596, 2010. 1.2.1

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–511, 2001. 2.3.1.2, 2.3.1.3, 2.3.2.1

Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 2.3.1.3

Christian Wallraven, Barbara Caputo, and Arnulf Graf. Recognition with local features: the kernel recipe. In *IEEE International Conference on Computer Vision*, pages 257–264, 2003. 2.3.1

Zhenhua Wang, Bin Fan, and Fuchao Wu. Local Intensity Order Pattern for Feature Description. In *IEEE International Conference on Computer Vision*, pages 603–610, 2011. (document), 2.3.2.3, 2.16

J. Weickert. *Anisotropic diffusion in image processing*, volume 1. Teubner Stuttgart, 1998. 8.3.1.1

Kilian Q Weinberger and Lawrence K Saul. Unsupervised learning of image manifolds by semidefinite programming. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–988, 2004. 2.3.1

Willow-Garage and Intel. Camera calibration With OpenCV @http://opencv.willowgarage.com/wiki/, Dec 2011. 2.2.1.3

Simon Winder, Gang Hua, and Matthew Brown. Picking the best daisy. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 178–185, 2009. 2.3.2.3

Simon AJ Winder and Matthew Brown. Learning local image descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. (document), 1.2.1, 2.3.1, 2.9b, 2.15, 2.3.2.3, 2.4, 5, 3.3.3.1

Laurenz Wiskott, J-M Fellous, N Kuiger, and Christopher von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997. 2.3.1.3, 2.11c

Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *International ACM Workshop on Multimedia Information Retrieval*, pages 197–206, 2007. 6.1

Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *European Conference on Computer Vision*, pages 151–158. Springer, 1994. 2.3.2.4

Xenophon Zabulis, Thomas Sarmis, and Antonis A. Argyros. 3D head pose estimation from multiple distant views. In *British Machine Vision Conference*, 2009. 2.3.1.3

Hui Zeng, Ji-Yuan Dong, Zhi-Chun Mu, and Yin Guo. Ear recognition based on 3D keypoint matching. In *IEEE International Conference on Signal Processing*, pages 1694–1697, 2010. 3.3.4

Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009. 2.2

Dongmei Zhang and Martial Hebert. Harmonic maps and their applications in surface matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 1999. 2.3.1.1

Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007. 6.1.3

Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, 13(2):119–152, 1994. 2.3.1.1

C Zitnick. Binary coherent edge descriptors. In *European Conference on Computer Vision*, pages 170–182. Springer, 2010. 2.3.2.4