Robust Image Registration using Improved Local Descriptors and Support Vector Machines

Yu-Chun Danny Cheng

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy in Engineering

June 2010

Mechatronics, Department of Mechanical Engineering
The University of Auckland, New Zealand

Abstract

This thesis presents a detailed study and improvement to local descriptor processes for registering images for the purpose of three-dimensional reconstruction, using four Māori artefacts as case studies. The motivation for the research came from the issues which still exist in image registration when dealing with large magnitudes of image transformations.

Four major pieces of work were carried out in the course of this research. First, an evaluation was carried out to study the performance of local descriptor processes and based on the results, the local descriptor process was divided into three stages, of which two were closely analysed. Second, the local descriptor formation stage was studied, and two methods, colour and hybrid local descriptor methods, were developed using colour images instead of greyscale images to improve the uniqueness of local descriptors. Third, the local descriptor matching stage was studied, and a new method based on support vector machines was developed. Fourth, an assisted image registration programme was developed and is a semi-automatic approach for registering images.

Extensive amount of experiments were carried out to validate these work. It was found that the colour and hybrid local descriptor methods had gains in matching accuracy of up to 10% over existing methods, and the support vector machine maching method had increased matching performance of up to 20%. When the two methods were combined, it was found that performance gains of up to 25% could be achieved. For the assisted image registration programme, up to 50% improvement was achieved, and the advantage was more significant as the magnitude of image transformation increased, highlighting the need for such programme.

These results show that the proposed work in this research are significant contributions to literature. In addition, these results show that the proposed methods can be used successfully for registering images for three-dimensional reconstruction, where the image transformation between images are often large. As there is currently a need to reconstruct Māori artefacts, this research has provided a new approach for registering images of these artefacts, which could then be used to construct three-dimensional models of the artefacts.

Acknowledgements

Throughout the course of this research, help were received from many people for both PhD-related and non-related matters. There are in fact far too many to acknowledge all of them in this section, but there is no doubt they have contributed without expecting anything in return. Here is a short list, in no particular order, of the people who deserve a special thanks for their assistance and being part of my PhD.

- My supervisors, Dr Shane Xie and Dr Enrico Hämmerle, for their continuous support and feedback on the direction of my research. Without their support, especially during the hard times when the research was going nowhere, I would not be completing this research.

- Assoc. Prof. Xun Xu for arranging meetings with the Engineering Project in Community Service team and the Auckland War Memorial Museum, where the need for 3D reconstruction of Māori artefacts was identified.

- Dr Oliver Stead (formerly from the Auckland War Memorial Museum) who provided access to a wide range of artefacts in the museum as well as providing invaluable knowledge on the history of Māori artefacts.

- Assoc. Prof. Kecman (now part of Virginia Commonwealth University) for his input and guidance in support vector machines and machine learning algorithms.

- Dr Kepa Morgan (Associate Dean Māori) for kindly lending me some of his Māori artefacts to complete the series of experimental work conducted.

- Mr Kenneth Snow from the Manufacturing Systems laboratory who have provided technical support whenever a test rig needed to be built.

- All the colleagues in the office who has provided suggestions and feedbacks throughout my research, in particular, Dr Thomas Scelo, Dr Avinda Weerakoon, Dr Ariful Islam and Mr Gareth Ferrari.

- Lucy Chen who is the single person apart from my supervisors that had the honour of proofreading this thesis, not to mention all her mental support throughout the course of this research.

- Last but by no means least, my family who has provided me with this opportunity to carry out research at a world-leading university and provided both mental and financial support. My father whom I have inherited my curiosity for the unknown, and my mother who, above all things, has taken care of almost all of the housework during the

crucial times of my PhD which I would otherwise have spent a large amount of my precious time on. Without them I would not be here today.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
ALH	Adaptive Local Hyperplane
CLD	Colour Local Descriptors
CSIFT	Colour Scale Invariant Feature Transform
CT	Computed Tomography
DoG	Difference of Gaussian
DoH	Difference of Hessian
EPICS	Engineering Projects In Community Services
GLOH	Gradient Location Orientation Histogram
HLD	Hybrid Local Descriptors
ISDA	Iterative Single Data Algorithm
KHNN	*k*-local Hyperplane Nearest Neighbour
LESH	Local Energy-Based Shape Histogram
LoG	Laplacian of Gaussian
LOOCV	Leave-One-Out Cross-Validation
MATLAB	MATrix LABoratory
MRI	Magnetic Resonance Imaging
NN	Nearest Neighbour
NNR	Nearest Neighbour Ratio
PC	Principal Component
PCA	Principal Component Analysis
RANSAC	RANdom SAmple Consensus
RFE-SVMs	Recursive Feature Elimination with Support Vector Machines
SIFT	Scale Invariant Feature Transform
SVM	Support Vector Machines

List of Symbols

Due to the vast number of symbols present in this thesis, the symbols are listed in the chapter they first appear.

List of symbols from Chapter 2: Literature Review.

c_x, c_y	Principal point
f_x, f_y	Focal length
$O(n)$	Big O notation
$\mathbf{H}^{12}$	Homography matrix
$\mathbf{K}$	Camera/intrinsic matrix
$(\mathbf{R}^1, \mathbf{T}^1)$, $(\mathbf{R}^2, \mathbf{T}^2)$	Extrinsic parameters consisting of the rotation and translation matrices for the reference and sensed images, respectively
s	Distortion factor
T_n	Triangular number
(x^1, y^1)	Location of a single pixel in the reference image
$\hat{x^1}$, $\hat{y^1}$, $\hat{l^1}$	Components of $\mathbf{x}^1$
$\mathbf{x}^1$, $\mathbf{x}^2$	Location of pixels in the reference and sensed images, respectively
$\mathbf{X}$	Location of 3D points which correspond to $\mathbf{x}$

List of symbols from Chapter 3: Performance Evaluation of Local Descriptor Methods.

C_o^1, C_o^2	Camera centre of the reference and sensed images, respectively
d^1, d^2	Distance from the reference and sensed image planes to the object, respectively
h_O^1,h_O^2	Height of the object in the reference image and sensed images, respectively
I^1, I^2	Reference and sensed image planes, respectively
O	Object
S	Scale
x	Location of a pixel along the x-axis
x_c	Radial distortion centre along the x-axis
$\hat{x}$	Corrected pixel location along the x-axis taking into account radial distortion
$\hat{y}$	Corrected pixel location along the y-axis taking into account radial distortion
ΔT	Translation change from C_o^1 to C_o^2
θ	Angle of change when the camera is moved from C_o^1 to C_o^2

List of symbols from Chapter 4: Colour and Hybrid Local Descriptors Methods.

a, b, c, d	Real numbers of quaternions
ang^{O_1}, ang^{O_2}	Opponent angles
ang^{S_1}, ang^{S_2}	Spherical angles
c_1, c_2, c_3	Channels of the $c_1c_2c_3$ colour model
$\mathbf{C}$	Covariance matrix
d_{m_1}	Difference of two neighbouring pixels in a given orientation for the m_1 colour channel
d_x, d_y	Wavelet responses along the horizontal and vertical axes
$\mathbf{D}$	Diagonal matrix of eigenvalues
$H(R,G,B)$	Hue image
i, j, k	Imaginary components of quaternions
$I(R,G,B)$	Intensity image
l_1, l_2, l_3	Channels of the $l_1l_2l_3$ colour model
L^2	Euclidean distance measure
m	Number of local descriptor pairs
m_1, m_2, m_3	Channels of the $m_1m_2m_3$ colour model
M	Maximum value for R, G and B
MNCC	Normalised cross-correlation
p	Number of pixels along the horizontal or vertical axis in an interest region
p^1, p^2	Number of pixels along the horizontal or vertical axis in an interest region for the reference and sensed images, respectively
q	Quaternion
r	Number of interest regions
$r(R,G,B)$, $g(R,G,B)$, $b(R,G,B)$	Normalised R, G and B
$\mathbf{r}^1$, $\mathbf{r}^2$	Interest regions of the reference and sensed images, respectively
$\overline{\mathbf{r}^1}$, $\overline{\mathbf{r}^2}$	Mean of the interest regions of the reference and sensed images, respectively
$\widehat{\mathbf{r}^2}$	Mean subtracted region from the sensed image
R, G, B	Red, green and blue pixels
$S(R,G,B)$	Saturation image
v	Size of the local descriptor concerned
v_{m_1}, v_{m_2}, v_{m_3}	Number of vectors for each of the three colour channels of the $m_1m_2m_3$ colour model
$\mathbf{V}$	Eigenvectors

w_{L^2}, w_{MNCC}	Weights for the Euclidean distance measure and modified normalised cross-correlation values, respectively
w_{m_1}, w_{m_2}, w_{m_3}	Weights for the three colour channels of the $m_1m_2m_3$ colour model
(x,y), (u,v)	Indices for the pixels in an image
x_i^1, x_i^2	i^{th} vector of local descriptors from the reference image and sensed image, respectively
$\tilde{x}(m_1)$, $\tilde{x}(m_2)$, $\tilde{x}(m_3)$	Median for the three colour channels of the $m_1m_2m_3$ colour model
$\mathbf{x}_1$, $\mathbf{x}_2$	Image locations of two neighbouring pixels
$\mathbf{X}$	Input data matrix
Y	Greyscale pixel
δang^O	Opponent angle with error analysis
δang^S	Spherical angle with error analysis
λ_i	i^{th} eigenvalue
τ_{HLD}	Threshold for the hybrid local descriptor method
τ_{L^2}	Threshold for the threshold matching method using the Euclidean distance measure

List of symbols from Chapter 5: Local Descriptor Matching with Support Vector Machines.

b	Bias term for SVM
C, σ	Penalty and threshold for SVM with a Gaussian kernel
c_i	Square of the weights for RFE-SVMs
d	Order of polynomial for SVM with a polynomial kernel
$G(\mathbf{x}, c_i)$	Gaussian kernel
k	Number of iterations for cross-validation
L^1	Rectilinear distance measure
L^p	p-norm distance measure
n	Number of available classes for SVM
$n(\text{LD}^1)$, $n(\text{LD}^2)$	Number of local descriptors in the reference and sensed images, respectively
$n(\text{LD}_{\text{correct}})$, $n(\text{LD}_{\text{incorrect}})$	Number of correctly and incorrectly matched local descriptor pairs, respectively
$n(\text{LD}_{\text{total}})$	Total number of local descriptors
p	Number of features to be removed at each iteration by RFE-SVMs
$\mathbf{r}$	Ranking vector for RFE-SVMs
$\mathbf{s}$	Indices of vectors of local descriptors to be emptied and ranked in $\mathbf{r}$

w_i	Weight of the i^{th} value in the difference vector for SVM with a Gaussian kernel
w_i	Weights of each input vector for RFE-SVMs
$\mathbf{x}$, $\mathbf{x}_i$	Input matrix for SVM and the i^{th} input vector
$\mathbf{X^0}$	Input matrix for RFE-SVMs
$\mathbf{y}$ y_i	Output vector for SVM and the i^{th} output value
$\Re^n$	n-dimension real number
τ_T	Threshold for the threshold matching method
τ_{NN}	Threshold for the nearest neighbour method
τ_{NNR}	Threshold for the nearest neighbour ratio method
τ_{SVM}	Threshold for SVM output in the range of [-1, 1]

Chapter 1

Introduction

This chapter provides an insight into the field of image registration in computer vision, focusing on the registration of images in preparation for 3D reconstruction. The motivation behind the research is discussed leading to the definition of a list of objectives. The contributions of the research based on these objectives are also presented. The structure of the thesis is outlined, as well as an overview of the topics covered in this thesis.

Image registration is the process of aligning two or more images automatically by first identifying a set of corresponding features or regions between the images and then using this set of correspondences to compute the transformation matrix. Image registration has been used in various computer and machine vision applications over the years, and while methods have been developed for different applications, because of the vast number of research fields image registration can be applied to, it is almost impossible to develop a generic method that is optimised for the requirements of all the different applications.

One particular issue of concern is the registration of images when large magnitudes of image transformations exist [9]. An example is when only 16 images were used for constructing 3D models of objects [3]. This means that the methods would need to handle image transformations of approximately 22.5°. This research aimed to improve on existing image registration methods for dealing with this issue, as this is encountered in a variety of fields. Examples include medical applications [10, 11] and the construction of panoramic images from multiple images [12]. This research addressed the registration of images for the purpose of three-dimensional (3D) reconstruction. In order to reduce the number of images required for reconstruction, which is often desirable, the magnitude of image transformations involved is potentially large. Four Māori artefacts were used as case studies to validate the performance of the developed methods and algorithms. These methods focused on improving the robustness of image matches by two means, first by increasing the uniqueness of the local descriptors, and secondly by better using the data available from these local descriptors for matching. Because of the focus of development of the methods, they are not

limited to images of Māori artefacts, and can be applied to objects in other applications.

This chapter is structured as follows. First the motivation behind the research, as well as the current approach for the application concerned are presented in Section 1.1. The requirements of the application, based on the existing approach for the task, are outlined in Section 1.2. The objectives, scope and contributions are presented in Section 1.3 and lastly, this chapter is concluded with the outline of this thesis in Section 1.4.

1.1 Motivation and Case Study

A comprehensive review of image registration methods carried out during the course of this research showed that despite numerous improvements proposed by various researchers over the years, the task of registering images using image registration methods is still unsatisfactory. One of the areas which drew attention to this research was the registration of images when large magnitudes of image transformations exist. In many applications, there is often a need to register images with large viewpoint changes. One example is in medical applications, where multiple x-ray images of a patient might be required and these images may need to be joined together. In such circumstances, it is often desirable to reduce the number of images taken to minimise the amount of radiation the patient is exposed to [10, 11]. Another example is when panoramic images need to be constructed from multiple images. In such instances, the fewer images required, the easier it is for the end-user to take the necessary images [12]. In the case of 3D reconstruction, the goal is to construct 3D models of objects, therefore images encompassing the objects are required. Many published articles on 3D reconstruction have focused on using a minimal number of images to construct 3D models of objects. One example is the evaluation study in [3], where as few as 16 images have been used to successfully construct 3D models. In these conditions, the image transformation between each image is approximately 22.5° and this magnitude of image transformation is considered large for many methods [13]. While specialised equipment can be used for recording the necessary location and orientation information of the camera for each image taken for 3D reconstruction, this is not practical in many real-life applications and as such, further study was pursued to enhance image registration techniques for dealing with this issue.

Four Māori artefacts were used as case studies in this research to verify the performance of the methods, algorithms and tools developed. In recent years, various research institutes and museums worldwide have begun work in reconstructing artefacts and historical sites, with the aim of digitally preserving these valuable objects. Some examples include the Monticello [14], the Cuneiform tablet [15], the plastico di Roma antica [16] and the David statue [17]. Māori artefacts were used as case studies as these artefacts have a similar cultural importance to those digitised in existing studies. The Māori people are the indigenous

Polynesian people of New Zealand, arriving and settling in New Zealand some seven centuries ago. Over the years, a distinct culture has been developed due to the geographic isolation of the country. Due to this distinctiveness, there is a high cultural significance of the artefacts produced by the Māori people. The need to digitally reconstruct these artefacts in 3D has recently arisen [18], as there are many advantages in having 3D models of the artefacts, including digital archiving, analysing the geometrical structure of the artefacts without risking damaging the physical artefacts, and online display to showcase the artefacts. 3D models of the artefacts are also a better way of storing information compared to two-dimensional (2D) photographs, as 3D models contain significantly more information than 2D images and can be used for restoration of the artefacts in the future or for producing replicas.

Currently, the most commonly used methods for the reconstruction of objects from museums around the world involve different varieties of laser scanners, due to their ability to achieve a high resolution in the reconstruction of objects [18]. Laser scanners work on the principal of time of flight, and rely on the reflection of the laser beam. Thus, the surface of the object that needs to be reconstructed can be a limiting factor in determining whether or not the method can be used. In the case of the Māori artefacts, many artefacts possess reflective surfaces, a typical example of this being carvings made of pounamu or greenstone. In addition, laser scanners are also very time-consuming, labour-intensive and intrusive. This is not desirable due to fear of damaging high-valued artefacts.

Recently, a group of engineers from the University of Auckland have been working with the Auckland War Memorial Museum [19] on the Engineering Projects In Community Service (EPICS) project [18, 20, 21], focusing on the 3D reconstruction of artefacts by using laser scanners. While positive results and models have been obtained in the recent past, the drawbacks with the use of laser scanners discussed above meant that the selection of objects which can be reconstructed using laser scanners have been limited. Recent advances in 3D reconstruction using a computer vision approach have shown positive results [3], therefore further analysis into the feasibility of a computer vision approach to the problem was pursued.

1.2 Requirements of Application

As with all real-life applications, there are constraints and requirements that needed to be considered for the computer vision approach to image registration and 3D reconstruction. Māori artefacts are no exception. These artefacts need special care due to their historical values, as well as the fragility of many of the artefacts due to their age. Based on the issues encountered by the EPICS team, and taking into account practical limitations, the following requirements for the 3D reconstruction of these objects were defined [18, 20, 21]:

(a) robustness; (b) non-intrusive; and (c) not labour-intensive. The quality of the 3D model constructed is also an important factor. However, this research focused on the registering of the images in preparation for 3D reconstruction and not on the reconstruction itself, it therefore did not affect the design of the methods discussed in this thesis.

1.2.1 Robustness

To accurately reconstruct 3D models of objects using computer vision methods, a set of accurately aligned images is required to provide the intrinsic and extrinsic parameters required for 3D reconstruction methods. Image registration methods can be used to register the images and provide the transformation matrices of image pairs. These matrices can then be used to align the images in preparation for the 3D reconstruction stage. To accurately align these images, the method needs to be sufficiently robust against various condition changes such as image transformation and illumination changes. This requirement was considered the main focus of this thesis, and methods to deal with various image transformation and illumination condition changes needed to be studied and improved or developed.

1.2.2 Non-Intrusive

In many 3D reconstruction applications, holding the objects concerned by hand is not desirable. This is particularly true for the objects used as case studies due to the historical value of the artefacts concerned. In previous studies [18, 20, 21], this constraint meant that the objects that can be reconstructed using laser scanners were limited, and a non-intrusive approach would not only reduce the risk of damaging the objects, but also opens a new door to objects which are previously unavailable for 3D reconstruction.

1.2.3 Not Labour-Intensive

The use of laser scanners meant that the scanning process is labourious and time-consuming in both the scanning and post-processing stages. By reducing the time required for scanning and processing each artefact, it becomes possible to increase the efficiency and hence increase the number of artefacts that can be reconstructed in the same time frame. This is currently limited due to the amount of time and effort required in completing the digitisation and construction of the 3D model of each artefact using laser scanners. This requirement, along with the requirement on non-intrusiveness, were considered to be secondary since for a computer vision approach to 3D reconstruction using image registration algorithms, photographs are used as the data source. By its nature, this is relatively non-intrusive and not labour-intensive compared to laser scanners.

1.3 Scope of Research and Contributions

Based on the constraints of laser scanners in Section 1.1 and the requirements of the application concerned in Section 1.2, the aim was to develop new methods to provide a better solution to the problem faced. These methods needed to be robust against various imaging conditions and at the same time non-intrusive and not labour-intensive. The issues of non-intrusive and non labour-intensive posed by the case studies can be overcome with the use of digital cameras instead of laser scanners. As digital cameras can be considered an everyday item in current society, the use of this type of equipments for taking images of objects can be achieved in a non-intrusive and non labour-intensive manner. It then became clear that the focus of the research needed to be placed on the requirement of robustness of image matches. Based on this requirement and an in-depth study of existing image registration methods in literature, the following objectives were defined:

- Compare the performance of existing local descriptor methods.
- Develop new methods for two different stages of the local descriptor process: local descriptor formation and local descriptor matching.
- Analyse alternative approaches to a pure computer vision method to handle images where large magnitudes of image transformations exist.
- Implement and verify the performance of the proposed methods.

Based on these objectives, the main areas that have been studied and improved throughout the course of this research include: (a) two new local descriptor formation methods based on colour images instead of greyscale images; (b) a new local descriptor matching method based on Support Vector Machines (SVM); and (c) an assisted image registration programme which integrates local descriptor processes with the guidance of end-users. In addition, experiments were carried out to verify performance of the proposed methods. These were developed and implemented in MATrix LABoratory (MATLAB) for its ability to quickly and efficiently develop prototype programmes compared to third-generation programming languages like C++. Results from these experiments show that the two new local descriptor formation methods, as well as the SVM matching method are improvements over existing methods, with improvements of approximately 10% and 20%, respectively. To the best of the author's knowledge, these methods have not been presented in the literature. The contributions for each of the objectives listed are discussed below.

1.3.1 Performance Evaluation of Local Descriptor Methods

To understand the performance of recent local descriptor methods, an evaluation study was carried out. The study compared these methods under different image transformations,

including: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint changes. This study allowed for a better understanding of the suitability of these methods and was a crucial step in improving the performance of local descriptor methods.

1.3.2 Colour and Hybrid Local Descriptor Methods

Two new local descriptor formation methods were developed and are referred to as colour local descriptor and hybrid local descriptor methods. These are improvements over existing methods by using colour models instead of greyscale images. Colour images contain much more information than their greyscale counterparts and, using colour models computed from colour images, these methods can produce local descriptors that are more unique. This results in local descriptors which are more invariant, in another words, they remain unchanged, against various image transformation and illumination condition changes. Two feature-reduction methods were also developed to address the issue of an increased computation time which arose with the use of colour images.

1.3.3 Local Descriptor Matching with Support Vector Machines

Out of the three different stages of local descriptor processes, local descriptor matching was the least developed stage. Existing methods discard valuable information from the difference vectors of local descriptor pairs and by doing so, the matching accuracy of local descriptors are often undesirable when similar local descriptors from different regions of an image exist. The developed SVM-based method is a new approach for matching local descriptors that does not discard this valuable information and instead, the information is used in assisting the matching process. By using all the available information, the robustness of matches can be improved significantly. The computation time was a minor drawback of the developed method, and to further improve the practicality of the technique, two feature-reduction methods were experimented with to reduce the computation time required.

1.3.4 Assisted Image Registration

An user-assisted image registration programme was developed, which is a semi-automatic instead of an automatic approach to the registration of images. By using the advantages of the human eye, it is possible to accurately register images even when large magnitudes of image transformations exist in a simple and yet effective and efficient manner.

1.3.5 Māori artefacts

Even though the artefacts were not the centre of attention in this research, there were two advantages in using these as case studies. First, the artefacts posed additional challenges in

addition to dealing with large magnitudes of image transformations, namely: (a) repetitive regions; and (b) combination of both feature-rich regions and those that lack distinct features. These additional challenges were useful as it meant that more robust methods were required to deal with these images, thus making them even more suitable to a wide variety of applications. Secondly, by introducing a new approach for registering images of Māori artefacts, these images can be used for the 3D reconstruction of the artefacts in future research. This is a contribution towards preserving information of valuable artefacts.

1.4 Outline of Thesis

The remaining chapters of this thesis introduce the methods developed in the course of this research, as well as the experimental work conducted to verify the performance of these methods. Chapter 2 presents an overview of published research in related fields. First, a review of various image registration techniques, in particular, a detailed discussion on recent work in local descriptor processes is presented. A review of 3D reconstruction methods is also presented, as well as results from a recent evaluation study that demonstrated the capabilities of 3D reconstruction using computer vision. Following the literature review, a discussion on the relationship of image registration and 3D reconstruction, and how the information provided by image registration can be used for 3D reconstruction is presented. Lastly, research projects on reconstructing artefacts from museums around the world relevant to this thesis which are related to the application concerned are presented.

An evaluation study which studied the performance of various local descriptor processes is presented in Chapter 3. This chapter includes a discussion on the experimental design used for all the experimental work conducted, the image pre-processing steps used to ensure fair and consistent controlled experiments were carried out, and a detailed discussion on the results from the evaluation study. Based on the results from this evaluation study, issues in different stages of the local descriptor process are addressed. The direction for the development of new methods is finally discussed.

In Chapter 4, first a detailed discussion on the limitations of existing local descriptor formation methods, and the shortfalls of using greyscale images compared to colour images are presented. Two methods for computing local descriptors based on colour images were developed and experimental work conducted show that these methods are superior and more robust compared to existing methods. To address the issue of an increase in computation time due to the use of colour images, two feature-reduction methods were developed and applied. The results show that it is possible to reduce the computation time, while maintaining gains in matching accuracy over existing methods.

Chapter 5 discusses the current methods for matching local descriptors and based on close analysis of these methods, a new approach that matches local descriptors using SVM

is presented. The performance of the SVM-based local descriptor matching method was verified through a set of experiments. Two feature-reduction methods were introduced to deal with the increased computation time, while retaining the performance through the reduction in dimensionality of the difference vectors of local descriptor pairs used for matching. The versatility and robustness of the method was tested by using three different machine learning algorithms in place of SVM, and integrating the new matching method with different types of local descriptor formation methods.

In Chapter 6, the two sets of methods discussed in the two previous chapters are combined, and a complete local descriptor process is presented. Experiments were conducted and the results show that the combined approach can register in a more robust manner compared to existing methods. In addition to combining these two sets of methods, Chapter 6 also presents an assisted image registration approach which allows end-users to interact with the image registration process, and demonstrates how a minimal amount of user input greatly improves the robustness of image matches.

Chapter 7 presents the concluding marks of the thesis. This is followed by a summary of the important contributions made in this research, and the reasons that these contributions are significant. A list of future work is finally discussed, which presents potential improvements to the work presented in this thesis.

Chapter 2

Literature Review

This chapter presents a comprehensive literature review in the relevant fields of image registration and 3D reconstruction. As the focus of the research was on image registration, in particular local descriptor processes, much attention was given to this field. 3D reconstruction algorithms were reviewed and a recent evaluation study is discussed to show that these algorithms can be used to construct 3D models of the objects used as case studies. Lastly, the relationship between image registration and 3D reconstruction was investigated and discussed.

In order to design and develop image registration methods suitable for the application concerned for this research, it was important to first understand the various methods available and what makes these types of approaches suitable or unsuitable. The main aim of this research was to develop methods for registering images when large magnitudes of image transformations exist [13], for example when only 16 images were used for constructing 3D models of objects [3]. This means that the algorithms would need to handle image transformations of approximately 22.5°. The main focus of the research was on providing the necessary information for 3D reconstruction algorithms using image registration techniques instead of specialised equipment. In addition to reviewing image registration methods in computer vision, 3D reconstruction methods were also reviewed.

There are three primary objectives of this chapter, and the chapter is structured according to these objectives. Firstly, to present a review of image registration methods with the main focus on local descriptor processes which formed the basis for the work presented in this thesis. The literature review on image registration algorithms is presented in Section 2.1 and local descriptor processes in Section 2.2. Secondly, a review of 3D reconstruction methods and their performance is presented in Section 2.3. Lastly, Section 2.4 presents the relationship between image registration and 3D reconstruction algorithms and how the information from image registration can be used for 3D reconstruction. In addition to these three objectives, an overview of the work carried out by various universities and research

Figure 2.1: Overview of the image registration process.

institutes around the world which share similar goals to the EPICS team are presented in Section 2.5. The chapter is concluded in Section 2.6.

2.1 Review of Image Registration Methods

As it has been indicated in [9], there are a vast number of papers published in this field and it is therefore impractical to provide a comprehensive review on all the techniques developed over the years. This section aims to first provide an overview of the different types of image registration methods, before focusing on the group of image registration methods which form the basis of methods developed in this thesis.

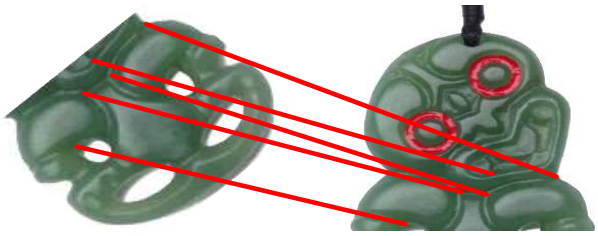
Figure 2.1 shows the processes involved in registering a pair of images, and an example is shown in Figure 2.2. First, a set of features in different forms is detected for each image, resulting in two sets of features for the image pair. These sets of features are then matched in order to identify a set of corresponding features between the two images. Given this set of corresponding features, the transformation matrix, in the case of this research the homography matrix which will be defined in Section 2.4 is computed.

Many researchers in image registration have placed their focus on different aspects of image registration, with the majority being around detecting or creating a set of useful features that are easily detectable across all the images, are unique and can be easily distinguished and represented to facilitate the matching of these features. The term ‘features’ may refer to points, lines or regions in images and different types of methods developed based on the different features utilised for the particular approach or application. An important aspect in determining and utilising the right type of algorithm for image registration is the type of features in an image.

In this research, images of Māori artefacts were used as case studies and these images needed to be registered for the purpose of 3D reconstruction. One of the difficulties with these types of images is that they often contain both feature-rich regions and regions that lack distinct features. In addition, there are also regions that have repetitive or very similar features and these needed to be considered when determining a suitable approach for the task. Examples of these difficulties will be discussed in Section 3.1, where the Māori artefacts used in this thesis are discussed.

Image registration methods can be categorised as either area-based or feature-based

(a)



(b)

(c)

Figure 2.2: Example of the image registration process: (a) a set of features is detected for each image; (b) the features are matched automatically; and (c) the matched features are used to compute the homography matrix, which is then used to transform one of the images to match it onto the other image.

(a) (b) (c)

Figure 2.3: Example of the difference between area-based and feature-based image registration methods: (a) image where a region or feature is computed; (b) area-based method, which uses a region of the image; and (c) feature-based method, which computes a feature around the region.

methods depending on the type of feature used for registering images [22, 23], as shown in Figure 2.3. Area-based methods utilise regions from images directly in identifying correspondences, as shown in Figure 2.3b. For feature-based methods, a set of features are identified from the images, and the features are often simplified or represented using other means before the matching process takes place [9]. An example is shown in Figure 2.3c. An example that distinguishes between area-based methods and feature-based methods utilising region features is that area-based methods utilise a region from an image directly for matching. In contrast, feature-based methods describe the region and reduce the amount of information carried by the region, for example the centroid and momentum of a region might be utilised instead of the region itself for matching.

2.1.1 Area-Based Methods

Area-based methods have traditionally been popular methods for their simplicity and ability to deal with any kind of images. The regions used for area-based methods range from regions detected using region detectors, a window of a pre-defined size or even the entire image. Therefore they are most suitable for images that lack distinct features, thus making it difficult for feature detectors to extract features.

A classic area-based approach is to utilise methods like cross-correlation, which directly compares the intensity values of two regions from an image pair. As these methods do not take into account factors such as intensity changes or different sensor types, they are prone to noise and can fail when surrounding conditions change. Another downside to the use of cross-correlation methods is that due to the computation of the cross-correlation value for different regions being computationally expensive, cross-correlation methods are often slow [24].

One remedy to such a problem with normalised cross-correlation to utilise Fourier transforms. Fourier methods have traditionally been used for area-based registration methods, however, because of the similarity between area-based methods and region features, they have also been used extensively in feature-based methods [9]. The Fourier transform transforms images from the spatial domain to the frequency domain, and the properties of the images in the spatial domain such as translation, rotation, reflection, distributivity and scale all have their counterparts in the frequency domain [25]. The images are usually transformed to the frequency domain either via hardware or use of the fast Fourier transform algorithm. The original image registration method in the frequency domain is based on the Fourier shift theorem [26] and was only proposed for translations. This method is robust against correlated and frequency dependent noise, as well as time-varying illumination disturbances, as the matching is performed in the frequency domain, making it simple to filter out the effect of these noises. This was enhanced in [27] to handle

rotation by analysing the phase correlation. When the image involves translation, rotation or scaling, [28] proposed a combination of the Fourier-Mellin transform and the cepstrum filter to handle these changes. The work was originally proposed to process medical images, in particular intraoral radiographs, and showed excellent results [29].

When using Fourier methods, one important factor is that noise in images is usually represented by high frequency signals in the frequency domain. As a result, low-pass filters are often used to deal with images in the frequency domain. This approach fails if the object in the image appears as high frequency signals when transformed to the frequency domain, as the object itself would be filtered out. Also because images are matched in the frequency domain, this kind of approach fails when there is frequency dependent noise in the images.

To further enhance the speed of normalised cross-correlation in the Frequency domain, Hii *et al.* [30] proposed a significantly faster method for calculating the normalised cross-correlation by using rectangular approximations in place of randomly placed landmarks. These approximations serve as an optimal set of basis functions that are automatically detected, and experiments from their work on Digital Image-based Elasto-Tomography show that the method is 37-150 times faster than the Fast Fourier transform-based normalised cross-correlation.

A relatively new type of area-based methods are the mutual information methods which originated from information theory and are a measure of statistical dependency between two data sets. The aim is to maximise the mutual information between two images and are typically used in multimodal applications, or the registration of images taken from different imaging sensors, such as a charge coupled device sensor and MRT images [31]. Over time, different methods have been developed to effectively maximise the mutual information methods, including the Marquardt-Levenberg method [32], the hierarchical search strategy together with simulated annealing [33], and the multiresolution hill climbing algorithm [34]. Mutual information has been primarily used in medical imaging due to the need to register multimodal images, for example in [35] where MRI, CT and positron emission tomography images of human brains were registered. MR images of the breast have been registered in [36] and muscle fibre images were registered in [37, 38].

The biggest downside to area-based methods lies in the fundamental approach of the methods. As many area-based methods compare the intensity of the pixels directly, the reference and sensed images must have similar intensities, either identical or near-identical in the case of cross-correlation methods or be statistically dependent in the case of mutual information methods. In terms of image transformations, because this type of methods compare pixels directly, only translation and small amounts of distortion such as rotation and tilt changes can be easily accommodated for. While it is theoretically possible to deal with distortion and scale changes, it is impractical to resample the images repeatedly for the different combinations of distortion of images due to viewpoint and scale changes and hence

the computation time is too high. On the other hand, this disadvantage does not render area-based methods useless in real-life applications, as they possess one important advantage over feature-based methods, which is the ability to deal with images where feature-rich regions do not exist.

2.1.2 Feature-Based Methods

Unlike area-based methods, feature-based methods first extract a set of features from the images and these features, instead of the areas or regions, are used for identifying correspondences. Features used in image registration can be divided into three groups: points, lines and regions. The discussion on both the features and the matching of these features in this section is categorised accordingly.

Features

Points are the simplest primitives and have been used widely in image registration since they are easily identifiable by humans. However, while these features are intuitive for humans, points are difficult to define mathematically. Traditionally points have been defined as line intersections [39], centroid of a closed-boundary region [40], or local maxima/minima of the wavelet transform [41, 42].

One of the earliest point detection algorithms was developed by Moravec [43], who defined a corner to be a point with low self similarity. Each pixel in the image is tested by considering how similar a patch is to the nearby patch, with the pixel being the centre of one patch. The similarity is calculated by the sum of squared differences of the two patches. A large sum of squared differences means that the patch and its neighbour is sufficiently different to be an edge [43]. One issue with the Moravec corner detector is that this algorithm is non-isotropic, that is, if an edge is not along the direction of the neighbouring patch, then the algorithm will consider it to be a line instead of a point. To overcome this problem, Harris and Stevens [44] considered the differential of the corner with respective to the direction, instead of using patches. The corner score is computed by taking the second derivative of the sum of squared differences between the two patches which is called the Harris matrix. A corner is present if both eigenvalues of the Harris matrix are large. Other popular corner detectors include the Kanade-Tomasi corner detector [45], the smallest univalue segment assymilating nucleus detector [46], and the features from accelerated segment test detector [47].

Line features in image registration are not restricted to edges and can represent any elongated structure, such as roads [48], coastal lines [49, 50], or anatomical features in medical applications [51]. Intuitively, edges and lines can be easily distinguished by human eyes, since it often involves a change in the intensity of pixels. However the problem lies

in how to define the threshold for the difference in intensity, which is application specific and often non-trivial. Amongst the many edge detectors developed over the last 20 or so years, the Canny detector [52] is the most commonly used detector today. Even though a large number of papers have been published on edge detectors, none have shown substantial advantages over the Canny detector [53, 54, 55]. The Canny detector first convolves the image with a Gaussian mask to reduce the noise level, four different masks are then used to detect vertical, horizontal and diagonal edges. This produces an intensity gradient for each pixel which also includes the direction of the edge because of the four masks used. The intensity gradients are then used to find the lines. The high intensity gradients are more likely to be edges that can be traced through by utilising the intensity gradients and their directions.

Region features are usually closed-boundary regions with or without high contrast between the inner and outer regions. One characteristic often used for representing a region is its centroid, since it is invariant to many forms of transformation including rotation, skewing and scaling [9]. Different examples of regions include lakes [56, 57, 58], buildings [59] and shadows [60]. Region feature extraction is usually a combination of edge detection and segmentation [61]. One of the most common region detector is based on the Laplacian of Gaussian (LoG). The LoG first convolves an image by a Gaussian kernel to give a scale-space representation of the image, the Laplacian operator is then computed, resulting in strong positive responses for dark regions and strong negative responses for bright regions [62].

Feature matching

Given the vast number of feature-based methods that have been developed over the years, a comprehensive review of these methods is not only impractical, but also overwhelming. Instead, the focus is placed on a representative few. These are discussed which aim to cover a wide variety of methods targeting different types of features available for feature-based methods. The first group of methods is the clustering technique which estimates the transformation parameters by a voting process. Initially, six accumulating vectors are defined which are used to estimate the transformation parameters. These accumulating vectors are initially zero, and by testing each combination of three point pairs between the images for correspondence, the transformation parameters can be defined, and their corresponding entries in the accumulating vectors are incremented by one. After the different combinations have been tested, the entries with the highest counts are used as the transformation parameters. This technique was first proposed in [39] which was originally used for matching image features to maps or models. Local errors in the images do not influence the global consistency of the registration process. Although the registration method was originally proposed to deal only with rigid transformations, it has been extended to deal with other forms of simple transformations [25].

A drawback of the original clustering registration method is the time complexity of the

algorithm which is $O(n^4)$ [57]. To deal with this issue, [57] suggested the use of a selection of points instead of all the feature points on the images. A subset of points are selected on the convex hulls of the images, with the assumption that there are common feature points on the convex hulls. In [58] it was proposed that the centre of gravity of region features are used as control points, and correspondences are searched between the control points. The matching is iteratively refined and the registration parameters are determined by the least squares method to achieve sub-pixel accuracy. Instead of the centre of gravity of region features, the images were divided into triangular regions in [63], which were formed by triangulating the control points. A linear mapping function was determined by registering the pairs of corresponding triangular regions.

The clustering technique has been implemented and improved in a number of studies. One notable area where the clustering registration technique has been applied is in biological and medical applications with multimodal images [64, 65]. [66] proposed a generalised framework for multimodal image registration based on the clustering of the feature space. The technique registers images by plotting all corresponding pairs on an x-y point which corresponds to the intensities of the points from the two images. [67] looked at the registration of biological images, such as histological sections, autoradiographs and cryosections. Given a pair of images, the dense similarity field is first computed with a block matching algorithm, hierarchical clustering method is then used to automatically partition the field into different categories, where the independent corresponding pairs are obtained. The transformation parameters can be estimated by using the Earth mover's distance [68] and a modified least squares method.

Clustering methods are generally less sensitive to uncorrected local variations as the spatial relationships between the control points of the images are used, and the methods consider all possible matches based only on supporting evidence. However, an important note is that the search space is usually large for clustering methods and to overcome this issue, *a priori* knowledge of the camera placement is usually required to reduce the computation time [25]. To overcome this issue, the performance of the clustering methods was improved in [69] by reducing the dimension of the parameters from four, including scaling, rotation and two translations to just scaling and rotation. This eliminates the noise effects of the translation components, making the algorithm more robust and more efficient. A method was also proposed in [70] to improve the performance of the clustering technique which is a coarse-to-fine approach based on [58]. The technique first estimates the scaling and rotation parameters between the images, the translation is then determined based on these information. The result is refined iteratively by matching the convex hull vertices. This approach also allows for image registration between images of different resolutions.

A method designed to deal with line features is the chamfer matching algorithm first presented in [71]. Chamfer matching was used to find the model of a coastline in segmented

aerial images. This approach works by first binarising an object in an image and the image itself, then using the sum of distances between the corresponding object and image points as the match-rating. The algorithm iteratively shifts the object and at each position, the sum of distances between the closest points is computed. The shift with the smallest sum of distances is used to find the true transformation of the images. Chamfer matching can be used when the local differences of objects are small but the global differences between images are large. An issue with the chamfering matching method is that sometimes the solution would converge incorrectly to a local minima, depending on the initial starting position of the algorithm. To overcome this, one approach is to use an exhaustive global search which has a high computation time due to the large search space [72], however does not guarantee the correct convergence. An improvement to this problem was presented in [73] by using a better, more generalised distance measure, as well as implementing a coarse-to-fine approach which significantly reduces the computational load. A recent study compared the shape context method with chamfer matching and it was demonstrated that chamfer matching is more robust in clutter compared to the shape context method, however is prone to local minimas even when using a global search [74].

Two groups of methods based on region features are the moment invariants and relaxation methods. The moment invariants method is a statistical measure of the characteristics of regions that may or may not specify a location in the image. This method was introduced in [75] and it was shown that the method is independent of translation, rotation and scaling changes. It was initially developed as a versatile way of describing and recognising patterns but have since evolved into a method for matching images. In essence, image moments are weighted averages or moments of the pixel intensities. Some properties of an object such as the area, total intensity, centroid and orientation can be identified using moment invariants. To match two images using moment invariants, the similarity between the values of the moments in the images are maximised. One of the problems of moment invariants is the computation time of the algorithm. This was especially problematic when the algorithm was first introduced as the computation power was significantly weaker compared to computers today. To overcome this problem, it was proposed in [57, 76] to use lower order moments for the matching process. Many variants of the moment invariant method have been proposed for different applications. Interesting algorithms which have been developed over the years include a way of dealing with blurred images [77] and 3D images [78].

The relaxation method has been around for a long time in mathematics, however it was only introduced into computer vision some 20 years ago [79]. This group of methods determines the point matching and transformation of images simultaneously, which is different from the majority of the other approaches, where the matching between images is determined before the transformation parameters are computed. The algorithm assigns, for each feature from the sensed image, a label and a feature from the reference image with

Figure 2.4: Overview of the three stages of the local descriptor process.

the same label. Each possible match of a feature defines a rating according to how closely other pairs would match under this transformation. The algorithm then iteratively adjusts the weight of each pair of points based on their ratings, with those matches whose locations are close to the actual location having a higher rating, thus having a bigger influence on the final match.

Relaxation methods are capable of handling global and local variations, however they can only handle translations. In order to deal with rotation and scaling, a uniformly distributed direction and size needs to be defined. This increases the computational complexity to $O(n^4)$ and is normally not suitable for real-time applications. The performance was improved in [40] to $O(n^3)$ by taking advantage of the properties of the features as well as the relative displacements and the use of two-way matching. This work was demonstrated using Landsat images in [40]. It was also suggested in [40] that in order to improve the performance when rotation or scaling is involved, *a priori* information of the scene should be used.

In contrast to area-based methods, features, instead of the intensity of the pixels in images are used for identifying correspondences in feature-based methods, therefore image intensities do not significantly affect feature-based methods. It is, however, still the best to normalise the images prior to registration. Because feature-based methods describe the images in terms of features and these feature representations are often small and efficient, it is possible to deal with distortion in images due to viewpoint or scale changes. On the other hand, because feature-based methods need to first identify a set of regions before features can be identified and used for matching, it is crucial that feature-rich regions exist in the images.

2.2 Review of Local Descriptor Processes

In addition to the methods discussed in the previous section, a subset of feature-based methods, local descriptor processes, are reviewed. Unlike many other feature-based methods which use the features extracted from the images directly to identify correspondences, local descriptor processes first extract a set of interest regions, local descriptors are then computed for each of these regions and these local descriptors are used for matching. This is a relatively new group of methods and the main advantage of these methods is that they have more distinguishing power compared to other traditional feature-based methods, as

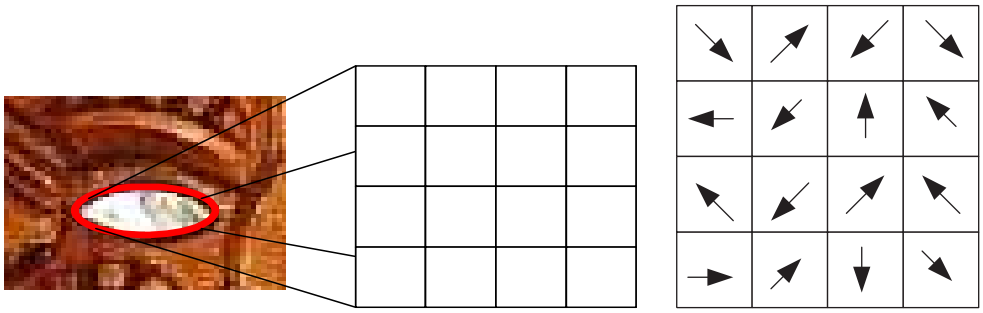


Figure 2.5: Example of a local descriptor: (a) a region is detected using a region detector, and is divided into a number of sub-regions; and (b) for each sub-region, descriptors are computed. The number of sub-regions and descriptors in each sub-region are dependent on the type of local descriptor method used.

the local descriptors formed for the interest regions are generally more unique, depending on the method used for computing the local descriptors. A recent study of various feature-based methods showed that local descriptor processes can register images of the Māori artefacts used as case studies with limited success [80]. Feature-based methods were of interest, as for 3D reconstruction it is desirable to have a wide coverage of the object and as it will be shown in Section 2.3.2. Current state-of-the-art 3D reconstruction algorithms can successfully reconstruct objects with few images, meaning that the change in viewpoint involved can potentially be large in order to reduce the effort required to take images for 3D reconstruction. Area-based methods cannot easily deal with large transformations [9], and due to the need to register images with varying degrees of image transformations, they were not suitable for the requirements of the application concerned.

Local descriptor processes can be divided into three stages as shown in Figure 2.4, and an example is presented in Figure 2.5: (a) region detection; (b) local descriptor formation; and (c) local descriptor matching. The last stage, local descriptor matching is not discussed as very little work has been carried out. It should be noted that throughout this thesis, the term ‘local descriptor process’ refers to the complete approach that consists of detecting a set of interest regions, computing local descriptors for each interest region and matching these local descriptors, whereas ‘local descriptor method’ refers to the computation of local descriptors given a set of interest regions. The two terms are defined explicitly as the three components of local descriptor processes were studied individually in order to perform controlled experiments to verify the performance of the developed methods.

2.2.1 Region Detectors

Region detectors aim to identify a set of regions in images and are designed to be invariant under a variety of image transformations such as rotation, scale, translation and

viewpoint changes. It is vital that a set of robust regions is detected, as each image may undergo different transformations and the algorithm must be sufficiently robust in order to accommodate for these changes. Since the quality of the regions determine the quality of the local descriptors formed, and this ultimately affects the robustness of the registration process, a vast amount of work has been put into perfecting the region detectors, improving both the computation speed and robustness. In particular, because translation changes do not affect the local descriptors formed, and changes in rotation are handled when the local descriptors are computed, the important area which determines the usefulness of a detector lies in its ability to detect a set of region which exists in different scales and viewpoints, and many of the region detectors utilise a method known as the scale-space theory [81, 82, 83].

One of the first region detectors introduced was the LoG. One problem with LoG is that it is inefficient due to the Laplacian operator, and to overcome this issue, the Difference of Gaussian (DoG) was introduced in [62] which is based on the LoG, however instead of using the Laplacian operator, the difference of the successive Gaussian-blurred images is computed. This variation is more efficient than the Laplacian approach, since the DoG can be computed by simple image subtraction instead of the Laplacian. Based on the DoG approximation, Bay *et al.* pushed the approximation of the LoG even further by using box filters [84]. Their work is based on the Determinant of Hessian (DoH) matrix and aimed at improving the efficiency of region detectors. Various measures have also been used to increase the computation efficiency, including keeping the weights applied to the filters simple and with the use of box filters, eliminate the need to iteratively apply the same filter to the output of a previously filtered layer, instead the filters are applied directly to the original image.

In addition to the LoG-based methods, four region detectors based on the Hessian and Harris detectors were proposed in [85, 86]. The Harris-Laplace detector is based on the Harris detector [44], and the regions are extracted by first searching for a set of interest regions in an affine Gaussian scale-space based on the LoG operator [83]. The interest points are initialised using the multi-scale Harris detector, which is followed by an iterative procedure that modifies the position, scale and shape of the region. The Harris-affine detector is similar to the Harris-Laplace detector, however instead of using the LoG operator, the Harris-affine detector is based on the second moment matrix and local extrema of the normalised derivatives. The Hessian-Laplace and Hessian-affine detectors are similar to the Harris-Laplace and Harris-Affine detectors, respectively. Instead of using the Harris detector, the Hessian corner detector is used for both these methods. The Hessian-Laplace is claimed to be more stable than the DoG operator proposed in [62], since the DoG detects edges and corners which are often unstable in the presence of large image transformations.

Methods that take different approaches to the LoG-based or Harris detectors are also present. The maximally stable extremal region detector [87] selects a set of extremal regions

which is a set of regions where all the pixels inside the regions have either higher or lower intensity than all the pixels on its outer boundary. The 'stable' regions are selected in the threshold selection process of the algorithm, which identifies those regions that are unlikely to change under various types of transforms. The regions are selected by identifying monotonic changes in the intensity of the pixels. Since this monotonicity is preserved when images undergo affine transformation, a matching extremal region will exist in the reference image, assuming the region is visible in both images.

The use of the probability density function of the intensity values of pixels to detect a set of salient regions was proposed in [88]. This region detector involves a two-step process. First, for each pixel, the entropy of the probability density function is computed over the parameters of ellipses, namely the scale, orientation and the ratio of the major to minor axes. The entropy extrema and the corresponding ellipse parameters are then recorded. After all the salient regions have been computed, they are ranked according to the magnitude of the probability density functions with respective to the scale and a threshold is defined depending on the number of regions required.

The edge-based region detector [89] is one of the earlier region detectors, which aimed at detecting affine invariant regions for image registration. This approach is based on the robustness of edge features [25], which are partially invariant to changes in image transformations and illumination conditions. The advantage of using edge features is the reduced amount of data that needs to be processed. The edge-based region detector starts off from a point detected using the Harris detector, and a close by edge using the Canny edge detector under different scales. From the corner detected, two points are moved away from the corner along the edge, and the relative speed of these two points are computed. These two points, along with the detected corner, form a parallelogram, and the points move away until the area of the parallelogram reaches a defined value. The parallelogram's centre of gravity of the region, intensity and the location of the point opposite of the original corner detected are used as the parameters describing the invariant region which can be described as a one-dimensional (1D) region as a function of the two vectors, defined by the original corner and the two points which moved away from the corner along the edge.

A similar method to the edge-based region detector later proposed in [90] is the intensity extrema-based region detector. This approach represents regions radially and replaces irregularly-shaped regions with ellipses. The method starts by detecting local intensity extremums at different scales and the intensity functions along rays emanated from each extremum are studied. For each local extremum, each ray is analysed by a function, which considers the intensity value at all points along the ray, and the point at which this function reaches an extremum is considered to be invariant under affine transformations. The points along all the rays of the local extremum which correspond to maxima of the function above are linked together to form an affine invariant region. This region is then replaced with an

ellipse having the same second order shape moments as the original, irregular-shaped region.

In the evaluation study by Mikolajczyk *et al.* [91], the Harris-Laplace detector and maximally stable extremal region detector had high scores under different scenarios in both the repeatability and accuracy of the algorithms. The maximally stable extremal region detector has been used successfully in [92], where license plates, faces and fibres in paper were tracked. The salient region detector was applied in object retrieval [93], video tracking [94] and detection of sea-surface targets [95], and the edge-based region and intensity extrema-based region detectors were used in image retrieval from large databases and servoing [89] and wide baseline image matching [96].

2.2.2 Local Descriptor Formation

After a set of regions is detected using a region detector, local descriptors are computed for each region using one of the following local descriptor formation methods. The uniqueness of a local descriptor determines how well a region can be distinguished from other regions which do not come from the same part of the object in 3D space. Many local descriptor methods have been proposed over the years in an attempt to produce local descriptors which best describe a region using the smallest amount of data possible. While it may seem that local descriptors containing a large amount of information will out-perform those of smaller dimensionality, it is a trade-off between the dimensionality of the local descriptors and the computation time in both forming and matching the local descriptors. It has been shown in many cases that the advantages gained in over-increasing the local descriptor dimension is not worth the extra computation time involved [97, 13, 84].

The Scale-Invariant Feature Transform (SIFT) proposed by Lowe [62] is one of the earliest and most well-known local descriptor method, which consists of three steps. (a) Low contrast regions at edges are eliminated using the Taylor expansion of the DoG scale-space function and the Hessian matrix, respectively; (b) orientations are assigned using a Gaussian-smoothed image; and (c) the local descriptors are formed which are 3D histograms of the gradient location and orientation, where the location is quantised into a 4×4 location grid and the gradient quantised into eight orientations. This results in $4 \times 4 \times 8 = 128$-dimension (128D) local descriptors and the contribution of each pixel in the histogram is weighted by the gradient magnitude and a Gaussian with respective to the scale of the region.

Based on the SIFT local descriptor method, various extensions have been proposed, one of such is the Principal Component Analysis-SIFT (PCA-SIFT) [97] which samples at $39 \times 39 \times 2$ locations. PCA is applied in order to reduce the dimensionality of the local descriptors for it to be usable in real-life applications. Another extension to SIFT is the Gradient Location Orientation Histogram (GLOH) [13] method which was designed to increase the robustness and distinctiveness of local descriptors. Instead of Cartesian

coordinates, polar coordinates are used and PCA is applied to reduce the dimensionality of the local descriptors down from 272 produced in 17 bins. The Colour SIFT (CSIFT) [98] is yet another extension to SIFT, where instead of greyscale images, colour images are used. To achieve colour invariance, the Kubelka-Munk colour model [99] is used and SIFT local descriptors are then computed from this colour model. It was shown to be more robust against colour and photometrical variations compared to SIFT.

The Speeded-Up Robust Features (SURF) local descriptor method [84] has advantages over existing techniques, in particular, it focuses on improving the repeatability, distinctiveness and robustness and at the same time, decrease the computation time required. The method is based on 2D Haar wavelet responses [100] and makes use of integral images efficiently. Experimental work has shown that the SURF local descriptor method has shown significant advantages in terms of both performance and computation time when compared with the SIFT local descriptor method [84, 80]. Local Energy-Based Shape Histogram (LESH) [101] is a recently proposed local descriptor method, which is built on a local energy model of feature perception. It encodes the underlying shape by accumulating local energy of the underlying signal along several filter orientations. Several local histograms from different parts of the image are generated and concatenated together into a compact spatial histogram. The LESH local descriptor method was originally developed for face recognition across different poses, however can also be applied in applications such as image retrieval, object detection and pose estimation.

Other types of local descriptors, which are not directly related or derived from the SIFT local descriptor method, have also been developed over the years. Shape context [102] has a similar approach to SIFT. It is based on edges extracted using the canny edge detector instead of image gradients. Shape context is a histogram of edge point locations and orientations containing 36D by using a log-polar grid. Shape context has been shown to give good results in the comparison carried out in [91], however since it is based on edges of images, when dealing with textured scenes or images where edges are not easily detected, the performance is reduced. The intensity-domain spin image approach proposed in [103] is based on spin images [104]. This approach computes a histogram of pixel distances from the centre point and intensities which are invariant under a number of transformations. The proposed approach utilises ten bins for distance and ten bins for intensity value, resulting in 100D local descriptors. Complex filters [105] rely on the use of a rotational invariant filter, based on the Gaussian filter. This generates a set of invariant local descriptors normalised to have a radius of one and samples in a grid of size 41×41. Rotation of the image changes the phase of the local descriptors, but the magnitude remains constant. The largest coefficients of the responses are used to orient all the descriptors to eliminate the effect of rotation of images. Another local descriptor method was developed in [106]. The method is based on moments and the moments of the images are computed for derivatives of an image patch as a function

of the order, degree and image gradients in the x and y directions. The moments describe the shape and intensity of an image patch, which can be easily compared in the case that geometric transformations exist. Although it is possible to compute moments of any order, it has been shown that high order moments are sensitive to small changes and distortions and should therefore be avoided [13].

2.3 3D Reconstruction Methods

Research in 3D reconstruction has been carried out for many years and there are a large number of methods developed. Given that the main focus of this research is on image registration, it is impractical to provide a comprehensive review in this area and instead, a specific type of 3D reconstruction method is presented to provide an understanding of the capabilities of 3D reconstruction methods. In addition, in order to understand the performance of state-of-the-art 3D reconstruction algorithms, a comparison of recent 3D reconstruction methods, along with their reconstruction accuracies using the same dataset is presented [3].

3D reconstruction algorithms usually utilise one of three types of methods for representing the reconstructed 3D models [1]: (a) point clouds; (b) triangular meshes; and (c) voxels. Point clouds is the simplest form of representing 3D models, where the models consist of a set of points. These points can be used to either represent the surface of an object only, or in the case of more complexed objects, the points can be used to represent the solid components of objects. Point clouds can be desirable in certain applications, since it is simple to fit the points to the data acquired, and the points can be simply defined by their 3D coordinates. In the case that a colour representation is desired, the point clouds are represented by their 3D coordinates and colour information. A downside to using point clouds is that as the approximation error is proportional to the square root of the inverse of the number of points, a large number of points is often required in order to accurately represent the object [107, 108].

Triangular meshes is the most common method utilised today for representing 3D models. The method is capable of representing complex surfaces in an efficient manner. The main advantage of triangular meshes over point clouds is that meshes are flexible and the storage size required depends heavily on the geometry of the surface of the object represented. For a flat surface, only a few meshes are required for a large volume while for complex features, the density of the meshes increases thus increasing the accuracy in representing these surfaces, whereas point clouds requires the same number of points regardless of the type of surface that needs to be represented. In addition, due to the popularity of triangular meshes, many graphics hardware, for example video display cards, have the capability of processing these meshes, reducing the computation time required by

software.

Voxels, or volume pixels, are volume elements representing values on regular grids in 3D space. Voxels are frequently used in medical data, for example data from Computed Tomography (CT), Magnetic Resonance Imaging (MRI) or ultrasonic scans. Voxels have not been utilised nearly as much in other computer vision applications, mainly due to the success of triangular meshes. However, their simplicity and ease of understanding, as well as these methods' ability to work directly in 3D space meant that occlusion can be easily handled [109].

2.3.1 Voxel Colouring

Traditionally, many 3D reconstruction algorithms are based on stereovision, which rely on computing disparity maps from stereo camera setup. One of the earliest work done in this field was by Marr and Poggio [110] who introduced and developed the fundamental theories of computing disparity maps based on triangulation. For the past 30 years numerous algorithms and techniques have been developed in an attempt to perfect 3D reconstruction process. The 3D reconstruction process can be roughly categorised into one of the following four groups [1]: (a) computing a cost function on a 3D volume, and extracting a surface from this volume; (b) iteratively evolve a surface to match the scene; (c) stereo-based methods to compute disparity maps; and (d) extract feature points from the scene and fit a surface by interpolating between the feature points. The methods presented in this section belong to the second group. This group of algorithms work by iteratively evolving a surface to decrease or minimise a cost function and the different types of approaches in this category include voxel-based, level sets and surface meshes.

Visual Hull

The visual hull [111] is a method in the shape-from-silhouette class of 3D reconstruction. It can be thought of as a 3D version of the 2D silhouette of an image, where only the outline of the scene is reconstructed. Unlike conventional 3D reconstruction methods, which usually reconstruct the scene by using all the pixel information available, the visual hull only uses the silhouette information of the images. Given a set of images of an object, each image is first divided into foreground and background through means of segmentation and/or thresholding. This produces a set of images, which contains silhouettes of the scene from different viewpoints. The outlines of the silhouette of the images are then projected into 3D space, and the intersection of the projection of the images is the reconstructed 3D model of the scene. Because this approach only uses information from the silhouette of the images, it is extremely fast compared to other 3D reconstruction algorithms.

This approach however suffers from one major drawback when compared to other

methods, that is it is not capable of dealing with scenes with convex features. Because the silhouette of images is used, there is no way of detecting the presence of convexity in an object, therefore regardless of the number of images used, the reconstructed model will never contain convex parts itself. The quality of reconstruction is also more dependent on the number of images compared to other methods, however apart from the need to take more images of the object, this usually does not pose a great issue, as the technique is extremely efficient and can easily deal with large number of images. The visual hull method is an important part of many voxel colouring-based methods, as it is often utilised to provide an initial bounding volume for many of the methods, since it can significantly reduce the initial volume of the object and hence reduces computation time.

Voxel Colouring

Many traditional 3D reconstruction techniques utilise stereopsis [112, 113, 114, 2, 115] which is how humans perceive objects in 3D. Instead of utilising stereopsis, Seitz and Dyer [109] took a different approach to solve the 3D reconstruction problem. Their work is similar in nature to that of [116] which, instead of the shape reconstruction problem that many stereo-based techniques have tackled, the problem is treated as a colour reconstruction problem. The voxel colouring algorithm starts with a volume of opaque voxels, which surrounds the target scene, typically a cuboid, however the visual hull can also be used to provide a bounding volume. The algorithm then traverses through the voxels one by one in a consistent manner, checking for the colour or photo-consistency of each voxel. Each voxel is then projected onto all the visible images. The colour value of the projected pixels are compared. If the colour of the projected pixels are the same, the voxel is considered to be colour consistent and remains in the volume. If on the other hand, the colour is inconsistent, the voxel is carved away, in other words they are made transparent. The algorithm progresses until all the voxels are photo-consistent.

Colour consistency, also referred to as photo-consistency, requires each voxel to project the same colour onto the pixels of all images from which they are visible from. Therefore, the photo-consistency check does not actually deal with the shape of the scene in the sense that no shape information is used for reconstruction, as it solely relies on the colour information.

Under ideal conditions, for a photo-consistent voxel, the colours of the projected pixels are exactly the same, however this is not the case in real-life scenes due to noise or imperfections in the imaging sensors. To compensate for this error, different photo-consistency measures have been developed to determine whether a given voxel is photo-consistent or not. Some examples include the use of standard deviation [109], adaptive threshold [117], and histogram [118].

Similar to other 3D reconstruction algorithms, voxel colouring makes assumptions about the scene in order to simplify the reconstruction process. First, the scene surface is assumed

to be approximately Lambertian, all surfaces are opaque, there exists constant illumination and the camera used is strongly calibrated. Lambertian surfaces mean that they have the same reflectivity, and a strongly calibrated camera is one where both the intrinsic and extrinsic parameters are known [119].

An additional constraint for the voxel colouring algorithm which is used to simplify the visibility issue, is the ordinal visibility constraint. This constraint defines the way the camera should be positioned in order to ensure the correctness of the algorithm. Two examples of camera placements are described in [109]. The first is a set of inward-facing overhead cameras traversing 360°around an object, and the second is an array of outward-facing cameras placed in a very similar manner to the first example. This assumption is however very restrictive and limits the type of object that can be reconstructed using the voxel colouring algorithm.

To solve the problem of restricted camera placement defined by the ordinal visibility constraint, Kutulakos and Seitz introduced the space carving algorithm [120]. Similar to the voxel colouring algorithm, space carving scans through the voxels one by one, checking for their photo-consistency as the algorithm progresses. In order to allow for arbitrary camera placement, the algorithm makes multiple scans of the volume, thus eliminating the constraint that the camera has to be placed in an orderly fashion. The carving process in space carving is conservative in that it never carves a voxel it should not remove, but it is likely that a voxel which is photo-inconsistent will be left in the scene due to the way the visibility of voxels are computed. When scanning through the voxels, a plane of voxels is checked at a time. Only the images from the camera viewpoints which are in front of this voxel plane are used for photo-consistency checks. A downside to this is that it is possible that a camera which is behind the voxel plane is visible to the voxel being checked, however, due to the way the photo-consistency is checked, this camera is not used for the photo-consistency check for the particular voxel.

To further improve on the voxel colouring algorithm, Culbertson *et al.* proposed the generalised voxel colouring method [121]. This is similar to the space carving algorithm [120], however the visibility of voxels is computed exactly, unlike space carving, where the visibility is only computed approximately, and therefore a more photo-consistent model can be produced compared to the space carving algorithm. Two different versions of the generalised voxel colouring algorithm were presented. The first utilises item buffers [122] and is referred to as generalised voxel colouring-item buffers, while the second utilises layered depth images [123] and is referred to as generalised voxel colouring-layered depth images. The generalised voxel colouring-item buffers records, for each pixel in all images, the surface voxel that is visible from the pixel. Once the item buffers are computed the voxels are then checked for photo-consistency in the usual manner. At the end of each iteration, the item buffers for all the pixels are computed again by rendering, using

z-buffering [124]. This process is repeated until all voxels are photo-consistent, at which point the algorithm stops and a photo-consistent model is formed.

The method used in generalised voxel colouring-item buffers ensures that all visible images are used when checking for photo-consistency for each voxel, however when a voxel is carved, the visibility of the remaining voxels change and therefore for each iteration it is required to recompute the item buffers to ensure that during each iteration, the visibility information of the voxels are correct. This is time-consuming and often only a small portion of the voxels have changed visibility, and is thus not a very efficient way of computing the visibility. The generalised voxel colouring-layered depth images approach is similar to item buffers but instead of recomputing the item buffers at each iteration, it records, for each pixel, a list of all voxels along the projection ray. The nearest voxel is considered to be the visible voxel for the pixel and is used for photo-consistency measures. When a voxel is carved, the layered depth images are updated and the next closest voxel for each pixel becomes the new surface voxel. The advantage of the generalised voxel colouring-layered depth images approach is that it does not require recomputing the item buffers each iteration. This saves a considerable amount of computation time, since it is relatively efficient to update only the surface voxel for each pixel, and does not require computing the distance of each voxel along the projection ray again. It does, however, require a much larger memory to store the information in the layered depth images compared to the generalised voxel colouring-item buffers approach.

To further enhance the performance of generalised voxel colouring, the appearance-cloning method was proposed in [125], where it was noted that while voxel colouring overcame the problem of self-occlusion found in many stereo-based methods, a new problem of photo-consistency is introduced. To overcome this issue, it was proposed that in addition to photo-consistency, the shape of the object of interest should also be utilised. From the experiments conducted in [125], it was found that the appearance-cloning method out-performed generalised voxel colouring without the use of any *a priori* knowledge about the object. It was noted that by including *a priori* information and tailoring the algorithm to be application-specific, further improvements can be achieved.

Alternative approaches that studied the problem from a statistical point of view also exist [126, 127]. Eisert *et al.* proposed a multi-hypothesis voxel colouring method [128, 129], where the process was divided into two stages: hypothesis assignment and hypothesis removal. In the hypothesis assignment stage, each voxel centre is projected onto each image. These pixel colours are compared for all pairs of views and if the voxel is consistent with at least one pair of images, a hypothesis is assigned. This is performed for all the voxels, including the interior voxels. In the hypothesis removal stage, occlusion is taken into account and for each image, the voxel space is traversed in an occlusion-compatible direction [130, 131]. A visible voxel is projected onto the image, and the pixel to which the voxel centre

projects is compared with the voxel's hypothesis. The inconsistent ones are then removed, and this process is repeated for all views. This process simplifies the visibility problem for each voxel, however at the cost of extra computation, since hypotheses need to be assigned to all the voxels, including interior ones.

Broadhurst and Cipolla introduced a statistical consistency check for the space carving algorithm [132, 133], where the global noise parameter, required for the space carving algorithm is eliminated and instead, an additional constraint that the object contains no holes is imposed. By utilising a probabilistic framework for space carving, where each voxel is assigned a probability, a new consistency function was defined which calculates the probability that each voxel exists using Bayes' rule [134].

In addition to improving the voxel colouring method by developing new methods for checking the photo-consistency of voxels, studies that focused on overcoming the different constraints posed by the original voxel colouring algorithm also exist. One such group of studies focused on dealing with different surface types. Various methods have been proposed to deal with specular highlights, such as the colouring caching method [135], the study in [136], where a method was proposed to reconstruct objects with specular highlights without the need to calibrate for lighting conditions, and [137] where multiple views of a calibrated camera were used to handle specular highlights. In addition to specular highlights, transparent surfaces were also studied in [138, 139].

A special case of 3D reconstruction using voxel colouring deals with dynamic scenes. The dynamic voxel colouring method [140] aimed to reconstruct scenes from video images by utilising texture mapping in hardware, spatial and temporal coherence and a coarse-to-fine approach. Similar spatial and temporal coherency approaches were also utilised in [141, 142, 143]. In all these studies, non-rigid bodies were studied and it was found that these methods can successfully reconstruct the scene or object using this spatial and temporal coherence in a similar fashion to optical flow methods [144].

One of the main drawbacks of the voxel colouring approach is the relatively long computation time required compared to other 3D reconstruction methods. To make the voxel colouring approach more suitable for real-life applications, one of the methods proposed in [140] is the coarse-to-fine approach which is one of the most widely used method for improving the performance of voxel colouring methods [145, 146]. The generalised voxel colouring-layered depth images [121] is another way of increasing the performance of the voxel colouring-based methods. An alternative approach to the generalised voxel colouring-layered depth images algorithm was proposed in [147]. In this study, ray traversal is used for determining voxel visibility incrementally which involves projecting a ray from each pixel, then finding the voxels on its path until it hits a surface voxel and stops. A Method for efficiently reconstructing 3D models of both static and dynamic scene from stereo images, stereo image sequences, and images captured from multiple viewpoints was

proposed in [145, 146], and is referred to as the embedded voxel colouring approach.

2.3.2 Accuracy of 3D Reconstruction Methods

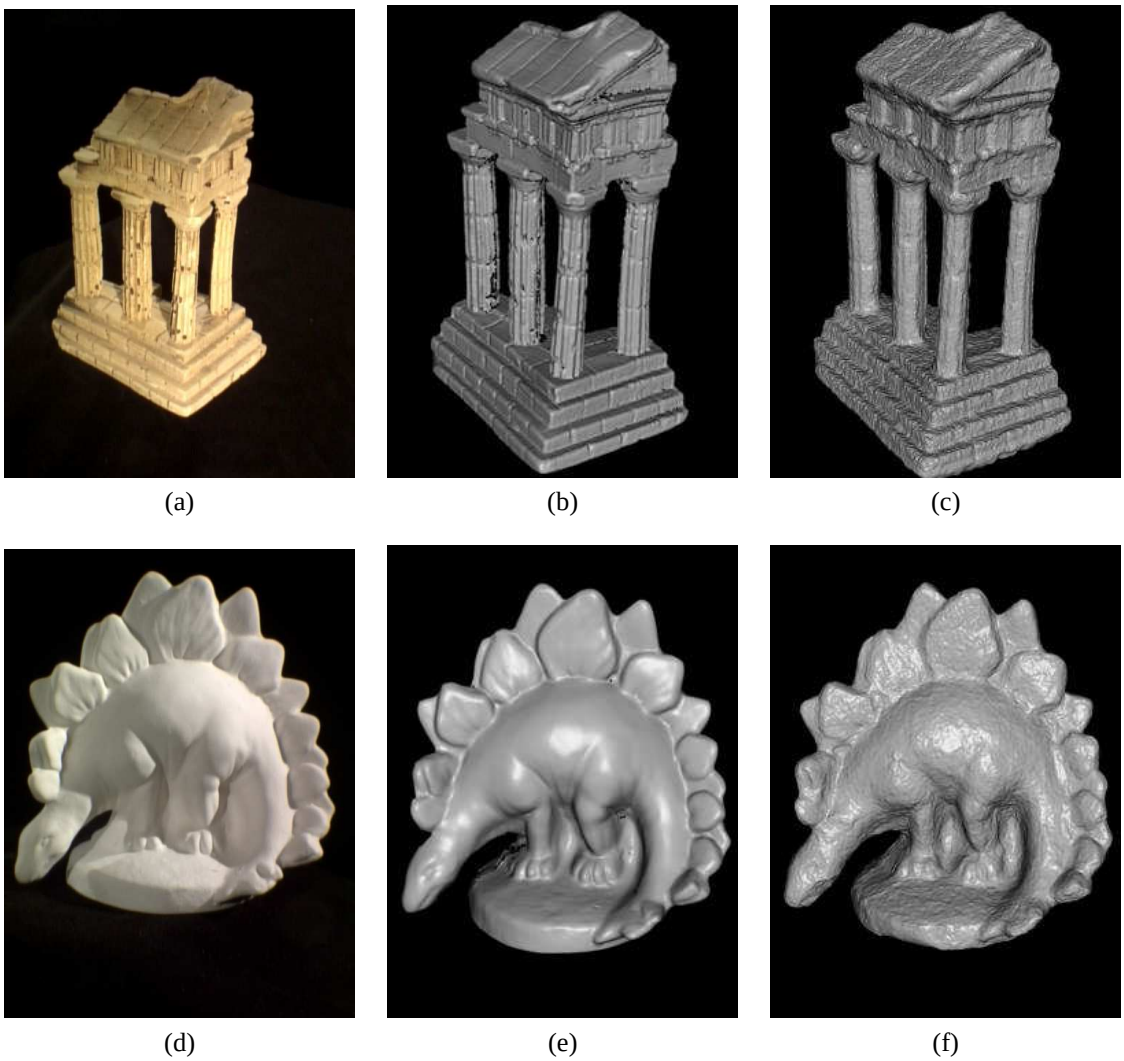
One of the common misconceptions is that computer vision methods for 3D reconstruction have a significantly lower resolution compared to laser scanners. While it is true that laser scanners can achieve a higher resolution, with the advancement of 3D reconstruction methods over the years, the difference in performance is decreasing steadily. In order to further improve current state of the art, 3D reconstruction algorithms have been compared [1]. The datasets used in [1] are also made available to future researchers for benchmarking purposes [3]. Images of two objects are included in the dataset. The first is a plaster reproduction of the Temple of the Dioskouroi, also known as the Temple of Castor and Pollux and the second is a plaster stegosaurus. These objects are shown in Figure 2.6.

To ensure fairness of comparison, the ground truth models are not made available in [1] and instead, results were submitted to the authors, which were then compared against the ground truth models. Both models have a resolution of $0.25mm$, which were captured using the Stanford spherical gantry [148]. The calibration of the images used for 3D reconstruction have an accuracy of approximately one pixel, and the images provided have resolutions of 640×480 pixels or video graphics array resolution. For each object, three sets of images are available: full, ring and sparse. The full set contains the most number of images taken from different viewpoints around the objects, the ring set contains 48 images taken on a ring around the object, and the sparse set is similar to the ring set, except only 16 images are available. The results from the full set are reproduced from [3] and discussed. The best reconstruction results from [3] are shown in Table 2.1.

In order to quantise the performance of the various methods, two terms were introduced in [1]: accuracy and completeness. Accuracy is defined as the distance, in *mm*, such that 90% of the reconstruction is within this defined distance of the ground truth, and the completeness is the percentage of points on the ground truth that are within 1.25 *mm* of the reconstructed models. Lower accuracy and higher completeness values denote better algorithms.

As can be seen in Table 2.1, the methods shown have accuracies of less than 1 *mm* and a completeness percentage of 98% or higher. Noting that the images used by all these algorithms are of video graphics array resolution, this demonstrates the improvements to 3D reconstruction algorithms in recent years, and prove that a computer vision approach to the problem of reconstructing Māori artefacts used as case studies is a viable one. As the resolution of reconstruction for computer vision approaches is highly dependent on the resolution of the images used, better results could be obtained by using images with a higher resolution than video graphics array resolution.

One thing that should be noted from the results shown in Table 2.1 is that the results



(a) (b) (c)

(d) (e) (f)

Figure 2.6: Objects used in [1] to compare the performance of 3D reconstruction algorithms: (a) image of the temple of the Dioskouroi used for 3D reconstruction; (b) 3D ground truth model of the temple; (c) 3D model of the temple by [2]; (d) image of the stegosaurus used for 3D reconstruction; (e) 3D ground truth model of the stegosaurus; and (f) 3D model of the stegosaurus by [2] (reproduced from [3]).

shown were reconstructed using the full set of images, and in most cases when using the ring or sparse sets, the results are not as desirable except in a few cases, one most notable being the work by Furukawa and Ponce [2]. The results obtained using the algorithm presented in [2] are similar for both objects when the full and sparse sets were used. This result is important as it will be shown later that by being able to use a small number of images for reconstruction, in this case only 16 images were used, the requirements for image registration is increased. This leads to a need for algorithms that are capable of dealing with these changes.

Table 2.1: Matching accuracy of various 3D reconstruction algorithms (reproduced from [3]).

	Temple		Stegosaurus	
	Accuracy	Completeness	Accuracy	Completeness
Campbell *et al.* [149]	0.41	99.9		
Furukawa and Ponce [2]	0.49	99.6	0.33	99.8
Goesele *et al.* [150]	0.42	98.2	0.46	96.7
Habbecke and Kobbelt [151]	0.66	98.0	0.43	99.7
Hernandez and Schmitt [152]	0.36	99.7	0.49	99.6
Hornung and Kobbelt [153]	0.58	98.7	0.79	95.1
Vogiatzis *et al.* [154]	0.50	98.4		
Zach [155]	0.51	98.8	0.55	98.7

2.4 Image Registration and 3D Reconstruction

So far this chapter has presented an overview of the capabilities of the state-of-the-art 3D reconstruction methods, however the relationship between 3D reconstruction and image registration has not yet been clarified. In Section 2.3, it was mentioned that for many 3D reconstruction algorithms, the camera needs to be strongly calibrated, in other words, the extrinsic and intrinsic parameters of the camera for each viewpoint need to be known. In computer vision, the extrinsic parameters refer to the pose, or location and orientation, of the camera, and the intrinsic parameters refer to the characteristics of the camera used. For simplicity as well as practical reasons, it is often assumed that the same camera configuration is used for capturing a set of images for 3D reconstruction, which is not an unrealistic assumption, as it is unlikely that different cameras would be used when taking the required images. The extrinsic parameters are often acquired by the use of specialised equipment, for example the Stanford spherical gantry. In order to utilise image registration techniques for computing the required extrinsic parameters instead of relying on specialised hardware, first the advantages of an image registration approach compared to utilising specialised hardware is compared. The relationship between image registration and 3D reconstruction is then discussed.

2.4.1 Advantages of Automatic Image Registration

While dedicated equipment like the Stanford spherical gantry is capable of traversing the camera to the required location in an accurate manner, they restrict where the equipment can be used, as well as the type of object that can be photographed. An alternative approach is the use of turntables [25, 156], an example is the one used in the Virtual Heritage Acquisition and Presentation project [6]. A limitation of this type of approach is that the size is restricted by the size of the turntable.

To avoid the need of using equipment which operates under very strict conditions, alternative approaches for acquiring images for 3D reconstruction were studied. One possible approach is to manually align images. This involves manually selecting a set of corresponding points from image pairs. However this method is labourious, time-consuming and not realistic when a large number of images are concerned. Image registration without user input is the most suitable approach, as it eliminates the afore-mentioned problems of using specialised hardware and handling issues, and instead relies on computer vision algorithms to register the images automatically. Automatic image registration using computer algorithms has numerous advantages over manual registration. For example, it has the ability to register a large number of images simultaneously, which is time-consuming and labour intensive if done manually. Another advantage is that due to the vast improvement in computation power in personal computers, it is possible for end-users to reconstruct 3D models of objects using these methods. As the methods are automatic, no special training is required and this broadens the use of computer vision algorithms, as well as attracting more researchers into the field.

2.4.2 Relationship Between Image Registration and 3D Reconstruction

Figure 2.7 shows an overview of the conventional approach to 3D reconstruction, where hardware such as the Stanford spherical gantry or turntables are normally used. When using such hardware, the extrinsic parameters, denoted by $(\mathbf{R},\mathbf{T})^1$ and $(\mathbf{R},\mathbf{T})^2$ in the figure which refer to the extrinsic parameters of the reference and sensed images, respectively, are obtained as images from different viewpoints are taken as shown in Figure 2.7. Figure 2.8 shows an overview of the relationship between image registration and 3D reconstruction. Instead of acquiring the extrinsic parameters of viewpoints when images are taken, the extrinsic parameters are computed by using image matches. Note that in both cases, the intrinsic parameters of the camera are assumed known, which is a reasonable assumption, as the calibration of the intrinsic parameters is an offline process and only needs to be performed once. This can be performed prior or after the required images are taken.

In order to compute the extrinsic parameters of images, the relationship between image pairs needs to be identified. Given a pair of images where the intrinsic parameters are known, the images can be described by the homography matrix:

$$\mathbf{x}^2 = \mathbf{H}^{12}\mathbf{x}^1 \tag{2.1}$$

Where $\mathbf{x}^1$ and $\mathbf{x}^2$ are the pixel locations in the reference and sensed images, respectively, defined as:

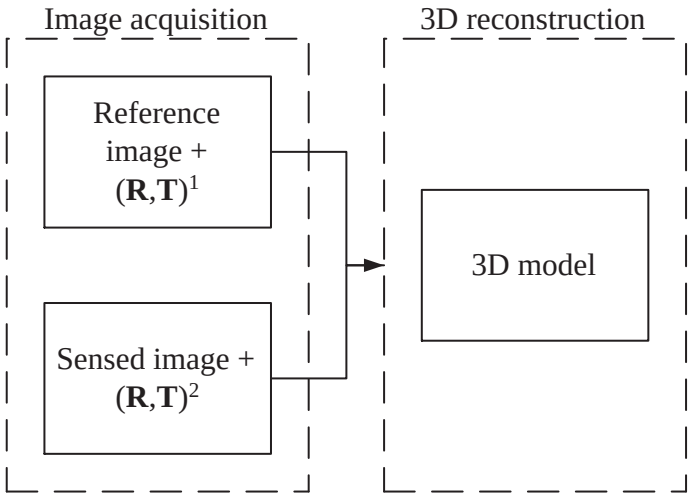


Figure 2.7: 3D reconstruction using specialised hardware for acquiring the extrinsic parameters of each viewpoint.

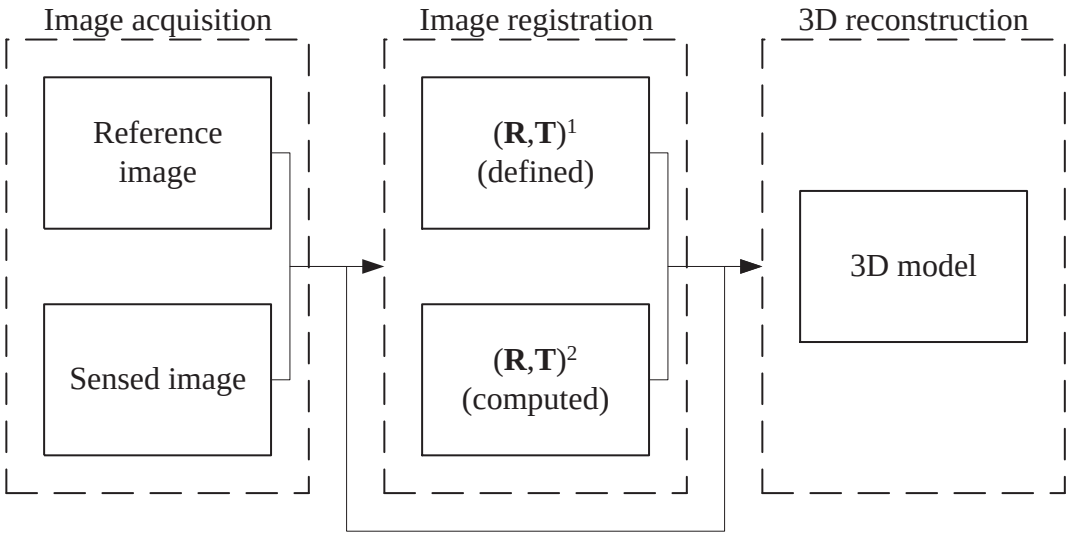


Figure 2.8: Relationship between image registration and 3D reconstruction.

$$\mathbf{x}^1 = \begin{bmatrix} \hat{x^1} \\ \hat{y^1} \\ l^1 \end{bmatrix} \tag{2.2}$$

The location of a pixel (x^1, y^1) in the reference image is then given by:

$$x^1 = \frac{\hat{x^1}}{l^1} \tag{2.3}$$

Similar equation holds for y^1 and $\mathbf{x^2}$. $\mathbf{H}^{12}$ is the homography matrix describing the projection of these pixels from the reference image to the sensed image [119]. The homography matrix is a 3×3 matrix and can be computed given a set of matching points from the image pair. Many methods exist for computing the homography matrix and one popular approach is the RANdom SAmple Consensus (RANSAC) algorithm [157] combined with a least squares fit approach. This combination of algorithms is capable of computing the correct homography matrix given a set of corresponding point pairs from the image pair in the presence of outliers. The ability to handle and discard outliers means that the homography matrix will be accurate, and will not be affected when outliers are present, which is inevitable in real-life applications. RANSAC is an iterative method that works by first selecting a random subset of the original data, the data are then defined as the hypothetical inliers and the hypothesis is tested as follows: first a model is fitted to the hypothetical inliers, all the other data are then tested against the fitted model, if a point fits, then the point is considered a hypothetical inlier. This fitted model is considered a good model if a sufficient number of points are classified as inliers. If the model is considered good then it is refitted using all the hypothetical inliers and the error of the model is computed using the refitted model and the set of hypothetical inliers. This process is repeated using randomly selected points as hypothetical inliers, and if the error is smaller than the previous models, the refined model is defined as the best model until a threshold or a pre-defined number of iterations have been executed.

Given the homography matrix, the next task is to relate Equation 2.1 with the extrinsic parameters. The projection equation can be utilised to achieve this and is defined by:

$$\mathbf{x} = \mathbf{K}[\mathbf{R}|\mathbf{T}]\mathbf{X} \tag{2.4}$$

Where $\mathbf{K}$ is the camera matrix, $\mathbf{R}$ and $\mathbf{T}$ are the rotation and translation matrices which form the extrinsic parameters of a camera viewpoint, and $\mathbf{X}$ is a set of 3D points on the object which corresponds to the set of image points $\mathbf{x}$. The camera matrix is also known as the intrinsic matrix and defines the intrinsic parameters:

$$\mathbf{K} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2.5}$$

Where f_x, f_y are the focal lengths along the x- and y-axes, c_x, c_y is the location of the principal point and s is the distortion factor. By combining Equations 2.1 and 2.4, the following is derived:

$$\begin{aligned} \mathbf{x}^2 &= \mathbf{H}^{12}\mathbf{x}^1 \\ \mathbf{K}[\mathbf{R}^2|\mathbf{T}^2]\mathbf{X} &= \mathbf{H}^{12}\left(\mathbf{K}[\mathbf{R}^1|\mathbf{T}^1]\mathbf{X}\right) \\ \mathbf{H}^{12} &= \left(\mathbf{K}[\mathbf{R}^2|\mathbf{T}^2]\right)\left(\mathbf{K}[\mathbf{R}^1|\mathbf{T}^1]\right)^{-1} \end{aligned} \tag{2.6}$$

In other words, the homography matrix can be defined as a function of the extrinsic parameters of both images as well as the camera matrix. In order to compute the extrinsic parameters given the homography matrix, Equation 2.6 is rearranged:

$$\begin{aligned} \mathbf{x}^2 &= \mathbf{H}^{12}\mathbf{x}^1 \\ \mathbf{K}[\mathbf{R}^2|\mathbf{T}^2]\mathbf{X} &= \mathbf{H}^{12}\left(\mathbf{K}[\mathbf{R}^1|\mathbf{T}^1]\mathbf{X}\right) \\ [\mathbf{R}^2|\mathbf{T}^2] &= \mathbf{K}^{-1}\mathbf{H}^{12}\left(\mathbf{K}[\mathbf{R}^1|\mathbf{T}^1]\right) \end{aligned} \tag{2.7}$$

From Equation 2.7, it can be seen that if the extrinsic parameters of the reference image are known, then the extrinsic parameters of the sensed image can be computed given the homography matrix which is computed from the matching of image pairs using image registration. Note that the pose of each viewpoint is relative to each other, it becomes obvious that the extrinsic parameters of the first reference image can be manually defined and all subsequent extrinsic parameters of images can be computed using the first image as the reference.

2.5 Review of 3D Reconstruction of Artefacts

2.5.1 Current Projects Around the World

Over the past decade or so, work has started to reconstruct artefacts around the world. Examples include the Canadian Heritage Information Network [158], the Virtual Heritage Acquisition and Presentation [6] in Europe, the Salzburg Research Institute [159] in Austria, the Statue of Liberty in New York by Texas Tech [160], the virtual Monticello [4] by the

University of Virginia, the Cuneiform tablets [15] by Johns Hopkins University in Baltimore, the plastico di Roma antica [16] of Rome and the David statue by Michelangelo [17, 5] by Stanford University and the University of Washington.

In addition to these work, various universities and museums have also started exploring the field in a more technical manner or initiated digitisation programmes, including the University of Stanford, the University of Ireland, Museum of the City of New York [161], Royal Ontario Museum [162], Museum of Science in Boston [163] and the American Museum of Natural History [164]. Museums have been successful in establishing 3D virtual museums on the internet to allow the general public easy access to the resource. Examples include the Canadian Museum of Civilisation [165, 166], the New Orleans Museum of Art [167] and the Victoria and Albert Museum [168] in the United Kingdom. Some of the notable examples are presented to provide an understanding of what has been achieved in the 3D reconstruction of artefacts around the world.

Monticello

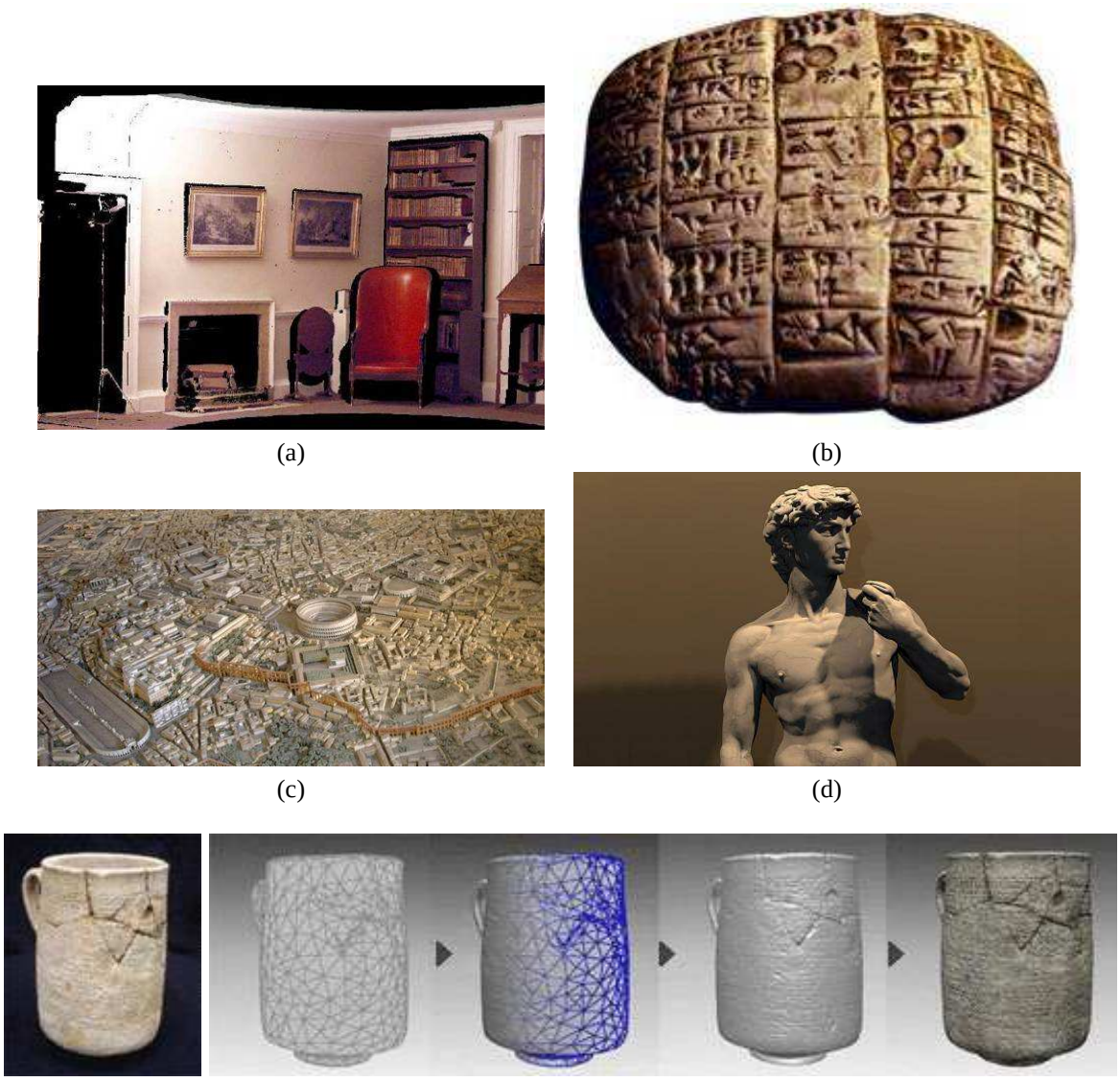
Monticello was the home of the third president of the United States, Thomas Jefferson. In 2002, Luebke and a team from the University of Virginia's Computer Science department in conjunction with the University of North Carolina scanned the estate using a laser scanner [14]. The 3D model constructed by the laser scanner was then combined with colour images to provide coloured 3D models. The reconstructed building is showcased in the Virtual Monticello at the New Orleans Museum of Art [167]. An example of the scanned and reconstructed model of Thomas Jefferson's library is shown in Figure 2.9a.

Cuneiform Tablet

The Cuneiform script is one of the earliest known forms of written language and Cuneiform tablets are tablets with the Cuneiform script carved onto them. Kumar and a team of students from the Johns Hopkins University in Baltimore scanned the tablets in 2003 using a laser triangulation scanner with a regular grid pattern at a resolution of 0.025 *mm* [15]. However despite the resolution the laser scanner is able to achieve, it was concluded that the Cuneiform tablets are difficult to scan due to the complexity of the pattern which exists on the surface of these tablets. An example of a Cuneiform tablet is shown in Figure 2.9b.

plastico di Roma antica

The plastico di Roma antica is a model of ancient Rome and Guidi *et al.* [16] scanned the model using a modulated light scanner, supplemented by a triangulation scanner. The modulated light scanner was utilised as neither triangulation-based methods nor time of flight methods were satisfactory, due to the object containing both large and small details.



(a) (b)

(c) (d)

(e)

Figure 2.9: Examples of 3D models computed by research institutes around the world: (a) Thomas Jefferson's Virginia home (reproduced from [4]); (b) an example of a Cuneiform tablet; (c) model of ancient Rome, plastico di Roma antica; (d) David statue by Michelangelo (reproduced from [5]); and (e) cup digitised by the Virtual Heritage Acquisition and Presentation project (reproduced from [6]).

The modulated light scanner was able to overcome this by providing both high accuracy for the small details and a wide range for the large details. The plastico di Roma antica model is shown in Figure 2.9c.

David Statue

Stanford University and the University of Washington digitised a number of sculptures and architectures of Michelangelo in 1999 using laser scanners [17], and the David statue, one of the most famous sculpture today was part of the collection of sculptures digitised. The objects were scanned at a resolution of 0.25 *mm* and was detailed enough to see Michelangelo's chisel marks. The work is presented at the 'The Digital Michelangelo Project' [5]. An example of the digitised David statue is shown in Figure 2.9d.

Virtual Heritage Acquisition and Presentation Project

Virtual Heritage Acquisition and Presentation was a project funded by the European Union and was designed to increase public awareness of Europe's most precious artefacts and documents. The project aimed to develop new tools for 3D scanning and acquisition of visually rich 3D models, post-processing and virtual heritage tools for presentation and navigation. Artefacts of varying sizes have been scanned. In particular, the project focused on the digitisation of ceramic artefacts, and these have been digitised using a 3D laser scanner with an accuracy of 0.008 *mm*. The scanner was mounted on a track and the scanning process was automated through the use of turntables. The digitised objects are displayed online [6]. An example of a cup which was digitised, along with the process in digitising the object is shown in Figure 2.9e.

2.5.2 EPICS

The EPICS project at the University of Auckland aimed at creating accurate 3D models of various Māori artefacts from the Auckland War Memorial Museum with the goal of utilising these 3D models for archival documentation, historical conservation, online exhibition, replication and educational purposes [169]. The project started in 2006, where comprehensive research was undertaken to investigate a suitable digitisation approach for these artefacts. Various laser scanning technologies and devices were examined. Based on the research, a low resolution Polhemus FastSCAN laser scanner [170] was first utilised. The scanner has a resolution of 1 *mm* and was used to scan artefacts in greyscale, as the scanner was not capable of scanning in colour. A higher resolution colour laser scanner [171] was later purchased which allows for the scanning and construction of fully colour rendered 3D models. Various improvements were made over the years to improve on the scanning of these artefacts, including the use of a KUKA robotic arm to overcome one of the

(a) (b)

Figure 2.10: Some artefacts scanned by the EPICS team: (a) canoe prow; and (b) wahaika club (reproduced from [7]).

problems faced due to the inconsistency in the movement of the laser scanner operated by hand [18, 20, 21]. The EPICS project website [7] contains a number of 3D models of artefacts scanned throughout the years, as well as a detailed discussion on the process involved in the reconstructing these artefacts. Part of the collection of artefacts reconstructed by the team are shown in Figure 2.10.

2.5.3 Issues with Current Approach

As with all real-life applications, the EPICS project was not without problems. The main problems include: (a) need for specialised equipment; (b) intrusiveness of the approach; (c) scanning time and labour intensity; and (d) reflective surfaces. It should be noted that while these issues are discussed in terms of the problems faced by the EPICS team, they are definitely not unique and are common problems faced by many groups attempting 3D reconstruction using laser scanners.

Specialised Equipment

It is no secret that specialised equipment like laser scanners are expensive and often require some form of training in order to properly utilise the equipment. This is no exception in this case and as described in [18], the original Polhemus FastSCAN laser scanner utilised suffers from a number of drawbacks including accuracy issues and the inability to scan colour information from the artefacts, and as a result a new laser scanner was required. While the purchase and use of specialised equipment are often not a major issue for research institutes or industries, the use of these equipment meant that the scanning of the artefacts or any other types of objects are restricted as it is often not viable to have multiple scanners to share the workload.

Intrusiveness

Even though laser scanners are non-contact 3D scanners, a certain amount of hand-handling is often required for various reasons, such as transporting the objects to the scanning equipment or positioning the object for scanning. This is particularly problematic when dealing with artefacts which have high historical values and are often fragile due to their age. One issue that was noted by the EPICS team was that the artefacts that the team had access to were restricted due to the amount of handling that is required for scanning the objects. Many of the artefacts available from the Auckland War Memorial Museum are one-off items and it was not possible to scan these artefacts due to fear of damage to the artefacts.

Scanning Time

Due to the nature of laser scanners which can often only scan a single line at any given time, the scanning of objects is done in a sweeping manner, where the laser scanner is moved from one end of the artefact to the other. Due to hardware restrictions, this sweeping movement often needs to be slow and in addition, even though the laser scanner is attached to a coordinate measurement machine, which measures and records the movement of the scanner as scanning is being performed, it was found that best results were observed when the sweeping motion is at a constant velocity for the duration of the scan [21]. This meant that in order to perform a scan, an end-user is required to hold the laser scanner and slowly and consistently sweep across the different surfaces of the artefact one at a time, until all the surfaces have been scanned, at which point a post-processing software is required to combine these scanned surfaces into a 3D model. This is a very time-consuming and labourious process and due to the need to move the laser scanner at a slow and consistent motion, it is difficult to achieve by humans operators. Even though a KUKA robotic arm was utilised to assist in the scanning process, it is still a tedious and slow task and often took days or weeks to scan a single artefact to a satisfactory level.

Reflective Surfaces

Another problem with many laser scanners is that they do not function well with shiny, reflective surfaces, as laser scanners rely on the time of flight of laser beams and when reflective surfaces were concerned, it was found that the lights were scattered and could not be properly scanned. This meant that a large number of artefacts such as those made of pounamu, or greenstones, could not be scanned as the results from scanning these objects are poor and could not be used in the construction of 3D models. An alternative approach is to use contact sensors, however this is undesirable for many objects that have high historical values. Computer vision is an attractive alternative, since reflectivity of surfaces is often not an issue for these methods.

2.6 Conclusions

This chapter presented a comprehensive review in related fields. First, an overview of various image registration methods proposed over the years was discussed before a detailed discussion on local descriptor processes, a subset of feature-based methods that showed promising performance for registering images when large magnitudes of image transformations exist were discussed. Large magnitudes of image transformations can be defined based on the evaluation study in [3], where as few as 16 images were used for the 3D reconstruction of objects [3]. Based on this definition, the algorithms would need to handle image transformations of approximately 22.5°. This detailed discussion on the various local descriptor processes provided a good background knowledge and collection of local descriptor processes that needed to be evaluated, in order to determine which method would be suitable for registering images when the discussed issues exist.

Because the goal of the application concerned was to construct 3D models of objects, using image registration to provide the necessary information, a review of 3D reconstruction methods and a comparison of the performance of the state-of-the-art techniques in this field was investigated. From the review of 3D reconstruction methods, it was clear that while computer vision methods still trail methods like laser scanners in accuracy, this difference is constantly reducing due to advancements in imaging sensors and 3D reconstruction algorithms. It was concluded that there are advantages to a computer vision approach for 3D reconstruction, such as the wide coverage of computer vision methods, and does not require specialised equipment. The discussion on how the required parameters for 3D reconstruction can be obtained by image registration methods demonstrated that the two fields can be efficiently integrated, and eliminate the need for specialised hardware to capture images for 3D reconstruction.

Lastly, projects that aimed at reconstructing artefacts or cultural and historical sites around the world were discussed. The challenges the EPICS team at the University of Auckland were faced with in dealing with the reconstruction of Māori artefacts were also investigated.

Chapter 3

Performance Evaluation of Local Descriptor Methods

In order to understand the issues which exist with current local descriptor processes, an evaluation study was carried out and presented. Based on the results from the evaluation study, issues in two of the three stages of the local descriptor process were identified, namely the lack of uniqueness in the local descriptors computed, and discarding important information from from the local descriptors when they are matched. Based on these issues, new algorithms were proposed to improve local descriptor methods.

Based on the discussion of the advantages of local descriptor processes over both area- and feature-based methods presented in the literature review in Chapter 2, an evaluation study was required. The evaluation study was necessary in order to determine the most suitable local descriptor process for registering images to deal with large magnitudes of image transformations, as defined in Chapter 2. The aim of this research was not to simply utilise existing methods, but instead, the algorithms presented in existing studies serve as the basis for the development of new methods. It was therefore vital that the performance of these methods were well understood and evaluated.

There are three project objectives described in this chapter. The first is to present a discussion on the artefacts used as case studies and the types of features found on these artefacts. The features found in these objects pose additional challenges in the registration of images of the objects, and it was therefore crucial that a thorough understanding of the features was gained. The experimental setup also needs to be discussed, as the same experimental setup was used for all the experimental work presented in the remaining chapters of this thesis. This discussion includes the various equipment utilised and the pre-processing steps carried out to ensure controlled experiments can be conducted. The last important piece of work is an in-depth performance study of various local descriptor processes proposed over the years and based on the results, the suitability of these methods

for the application of concern was determined.

This chapter is structured as follows. The Māori artefacts used in the various experimental work in this thesis and the reasons for choosing these artefacts, as well as a detailed discussion on the types of features found on these artefacts are presented in Section 3.1. The definitions of the two terms used to describe the results presented, accuracy and robustness, are presented in Section 3.2, followed by the experimental design in Section 3.3 which provides an insight into the hardware setup used to capture the required images. In addition, the four image transformations studied are also discussed in this section. The pre-processing steps used to process the images to ensure that fair and controlled experiments can be conducted are presented in Section 3.4, followed by the results from the evaluation study in Section 3.5. Following the evaluation study, the areas where further research is required are presented in Section 3.6, and the chapter is concluded in Section 3.7.

3.1 Māori Artefacts

Four artefacts were used for experimental work, composing of three wooden carvings and a greenstone pounamu tiki. These were selected as they are representative of the typical artefacts often found in Māori culture. These four artefacts are shown in Figure 3.1.

The wooden flute shown in Figure 3.1a is a replica of a typical flute found in Māori culture and was used as it consists of highly complex geometry on the surface. This complexity increased the difficulty of registering images of the flute, making them ideal for evaluating the robustness of the image algorithms developed in this research. It was of interest to develop algorithms which are sufficiently robust to deal with the various imaging conditions, as well as the difficulties posed by the artefacts which will be presented in the next section. The patu and wahaika in Figures 3.1b and 3.1c are different types of short striking weapons used by the Māori in battles [172]. These two weapons were chosen as they consist of different surface make-ups compared to the flute. The patu has a relatively smooth surface, however with a high level of texture, both from the grains of the wood used to carve the artefact as well as the drawings found on the surface. In contrast, the wahaika consists of both highly complex geometrical surface features as well as smooth regions, which increases the difficulty of image registration when different distinct features are involved. Lastly, the pounamu tiki was chosen due to its importance in the Māori culture where they are considered as taonga, or treasure. In today's society, tikis are often presented as gifts to visitors. In addition, since this type of objects have shiny, reflective surfaces, it was difficult for the laser scanner approach adapted by the EPICS team [18, 20, 21] to construct 3D models of these objects. It would therefore be of great significance to develop a new approach which is capable of reconstructing these objects in a simple and yet efficient manner.

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

Figure 3.1: The four Māori artefacts used for the set of experiments conducted in this research: (a) flute; (b) patu; (c) wahaika; and (d) tiki.

These artefacts were selected as they were readily available due to the relative simplicity in making these artefacts compared to larger or rarer and more valuable ones. While it would have been desirable to work with rare artefacts, this was avoided for practical and security reasons. Regardless, the four artefacts provided a good representation of the different types of artefacts typically found in the Māori culture and were ample for evaluating and verifying the performance of local descriptor processes.

3.1.1 Features on the Māori Artefacts

Unlike many objects which consist of either feature-rich surfaces or surfaces which lack distinct features, many Māori artefacts consist of both these types of surfaces. Feature-rich surfaces are best dealt with by feature-based methods while those that lack distinct features often require area-based methods, and the combination of these two types of surfaces meant that it is difficult to use one method for all the objects involved. An example of this is the wahaika artefact shown in Figure 3.1c. The bottom of the artefact in the figure shows intrigue carvings found in typical Māori artefacts, while the remaining surface of the artefact is smooth, with the only detail being from the grains of the wood used to carve the artefact. This combination of different surface structures makes it difficult for existing image registration methods to correctly register these images. Feature-based methods will struggle to register the smooth surface found on the majority of the surface of the artefact, while area-based methods cannot efficiently register the surfaces which are feature-rich, such as the part of the artefact containing the intrigue carvings.

Another issue with existing image registration methods for registering these images is the repetition of features found on the artefacts, such as the flute shown in Figure 3.1a. An example of the repetitiveness of features is shown in Figure 3.2, where there exist many swirl-like features on the object. This set of swirl-like shapes is another typical carving style found in many Māori artefacts and is found throughout the surface of the flute. Due to the repetitiveness of these patterns, it is very easy for image registration methods to mis-identify features from images, as many regions from the reference image may appear similar to a region from the sensed image. Another example is the tiki shown in Figure 3.1d which,

Figure 3.2: Example of the repetitiveness of certain features in Māori artefacts.

despite being hand-carved, is very symmetrical and similar to the flute. It causes problems when attempting to match images of these objects as the local descriptors representing different regions of the objects appear similar, therefore mismatches are more likely to occur. These two types of features posed challenges in the registration of the images of these objects, in addition to dealing with images with large magnitudes of image transformations. To overcome these issues, local descriptor methods capable of producing more unique local descriptors, and hence simplify the task of registering images using local descriptors were needed. This ensured that the local descriptor methods developed will not only perform well for the objects used as case studies, but in addition, due to the more unique local descriptors required, can perform favourably for other objects also.

3.2 Definitions of Accuracy and Robustness

Prior to presenting the evaluation work, the terms ‘accuracy’ and ‘robustness’ need to be formally defined to allow for a better understanding and interpretation of the results presented in this thesis.

3.2.1 Accuracy

The matching accuracy describes how accurate the matching of local descriptors is, and is defined in terms of the number of correct matches:

$$\text{accuracy} = \frac{\text{number of correct matches}}{\text{total number of matches}} \tag{3.1}$$

It will be shown in a later section that the results presented in this chapter are described in terms of the 1-precision, or 1-accuracy, which are two different names used in literature that

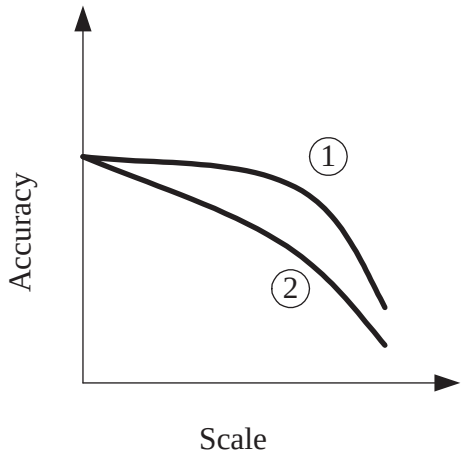


Figure 3.3: Difference between accuracy and robustness.

refer to the same term. The 1-precision was used in the evaluation study, as this format was utilised in existing studies [13]. The use of 1-precision therefore allowed for a comparison of results with existing literature. In later chapters, the precision, or accuracy, instead of the 1-precision, was computed for all experiments conducted as intuitively, accuracy is easier to understand. Because the recall values are not computed for reasons to be discussed later in the chapter, there are no advantages in using the 1-precision. The matching accuracy is a quantitative term and is used to describe how well a method has performed for a specific magnitude of image transformation, for example for a rotation change of 5°.

3.2.2 Robustness

The robustness of methods refers to the overall performance for an image transformation, such as the accuracy of a method for images with scale changes in the range of $[1.1, 1.5]$ and is a qualitative measure. The main difference between accuracy and robustness is the type of measure, where accuracy is a quantitative term and robustness is a qualitative one. This is best illustrated in Figure 3.3. The figure shows two possible results, described by the matching accuracy and plotted against scale changes. At a given scale, the second result has a lower matching accuracy compared to the first result. As a whole, the second result degrades quicker than the first, even though the initial performance of the two is the same. From the figure, it can be seen that the first result is more robust over the range studied due to higher matching accuracies for all scale values. In the discussion of results however, both terms provide a good understanding of the results presented and these two terms should therefore not be treated with great difference.

3.3 Experimental Design

In order to conduct well-controlled experiments to evaluate and further understand the suitability of local descriptor processes, an experimental setup was designed to capture the necessary images used for the experimental work. As the aim of this research was to develop methods to robustly register images for the purpose of 3D reconstruction, images encompassing the objects studied were required to provide a good coverage of the objects from all viewpoints. Due to this requirement, image pairs with large transformations, as defined in Chapter 2, needed to be registered. The experimental setup was designed with this in mind and covered image transformations along all three primary axes.

Four image transformations were considered: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint changes. These image transformations are shown in Figure 3.4. These image transformations include the rotation of the camera about the three primary axes, as well as a fourth transformation which studied the effects of the change in scale of the size of the object in images. Translation changes were not considered as it is a well-known fact that translation changes are often easily dealt with by both area- and feature-based image registration algorithms [9]. The translation of the camera does not cause significant distortion of the object in images and this distortion is often negligible for pure translation, depending on the size of the movement of the camera relative to the distance of the camera from the object. This is shown in Figure 3.5, where C_o^1 and C_o^2 are the optical centres and d^1, d^2 are the distances from the object to the reference and sensed images, respectively. θ is the angle and $\Delta\mathbf{T}$ the translation change between the reference image and sensed image, and O is the object concerned. The angle of change can be defined as:

$$\theta = \tan\left(\frac{\Delta\mathbf{T}}{d^1}\right) \tag{3.2}$$

If $d^1 \gg \Delta\mathbf{T}$, then:

$$\begin{aligned} \theta &\approx \tan(0) \\ &\approx 0 \end{aligned} \tag{3.3}$$

Since the angle of change is approximately zero, then the distortion of the object in the image due to viewpoint changes is approximately zero or in other words, negligible.

In order to provide a consistent lighting for all the images taken for the controlled experiments, a device was constructed to provide diffused lighting as shown in Figure 3.6. This also dealt with the problem of reflective surfaces, as diffused lighting meant that there were no specular highlights in the images. This problem is normally not encountered in a museum environment where the artefacts used as case studies are typically stored, as diffused

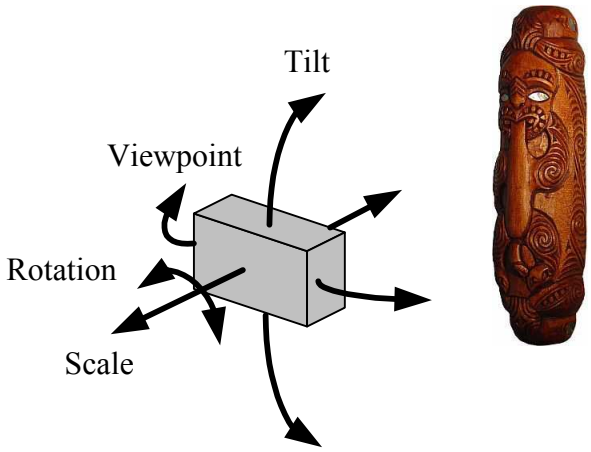


Figure 3.4: The four different image transformations utilised in the experimental work in this research: rotation, scale, tilt and viewpoint changes.

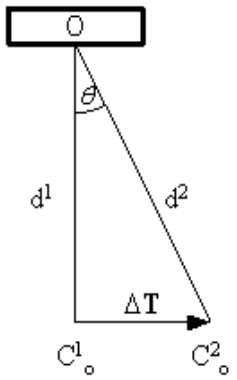


Figure 3.5: Effects of translation changes on the object in images.

lighting already exists, and was therefore not considered a constraint in this research. The same device was also used in Chapter 4 when studying the effects of different illumination colours and intensities. A data projector was used instead of standard light bulbs to provide the necessary illumination condition changes.

Examples of images taken for each of these four image transformations are shown in Figure 3.7. Figure 3.7a is the reference image for images in Figures 3.7b-3.7e, which represent the images taken when rotation, scale, tilt and viewpoint transformations exist, respectively. The four local descriptor processes evaluated in this chapter are: SIFT [62], GLOH [13], PCA-SIFT [97] and SURF [84], which were discussed in detail in Section 2.2.2. These four methods were evaluated as they showed good performance in previous studies [173, 13].

Figure 3.6: Experimental setup used to capture images in the Auckland War Memorial Museum for the experimental work conducted in this research. The device was constructed to allow for consistent, diffuse lighting to be projected onto the object and eliminated the problem of specular highlights in images.

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

Figure 3.7: Images showing examples of the four different image transformations utilised in the experiments in this research: (a) reference; (b) rotation; (c) scale; (d) tilt; and (e) viewpoint changes.

3.3.1 Rotation Changes

The rotation changes describe the rotation of the images about the optical centre of the camera, as shown in Figure 3.4. In order to take images of the object under rotation changes, the rotation of the camera, instead of the object, was utilised. This approach was taken as it was a much simpler task to rotate the camera, which was mounted on a tripod. Depending on the size of the object concerned, it is a more difficult task to rotate the object about the optical centre of the camera due to difficulties in mounting the object to a rotatable fixture in a safe manner. The centre of rotation was aligned approximately with the optical centre of the camera. As the alignment was not perfect, in addition to rotation, small amounts of translation also existed. This was however not an issue as the small translations had no effect on the images as it did not distort the object's appearance in the images, since it was assumed that the translation was very small given that the distance of the camera from the object was significantly bigger than the translation caused by the misalignment of the centre of rotation and the optical centre of the camera as shown in Figure 3.5. For the evaluation study, images used to study the effects of rotation changes were taken in the range of $[0^\circ, 90^\circ]$ at intervals of 5°. This range and interval, and the ones to follow, were chosen to provide a good coverage of the transformation concerned in order to evaluate the performance of the methods.

3.3.2 Scale Changes

Scale changes refer to the change of the size of the object in the images without the rotation of the camera or object in any of the three axes. To capture images for scale changes, three options were available: (a) utilising the zoom lens of a digital camera; (b) compute images of different scales from a reference image; and (c) physically move the camera or object to change the distance between the two. While the first two approaches are relatively simple, they are not without problems. Depending on the type of zoom lens used, the focal length is changed when the zoom of the camera is changed, which affects the intrinsic parameters of the camera, thus making the post-processing of images more difficult as the intrinsic parameters needs to be computed for each scale, making the approach difficult and time-consuming. Changing the scale of the object in images can be done easily in almost any image manipulation software, however this often introduces artifacts in the images and furthermore, using the same base image to generate a set of images of different scales, resulting in complications when registering these images. Due to these drawbacks, the images for scale changes were captured by mounting the camera on a slide rail which is capable of traversing in a linear fashion away or towards the object. The scale of an object in an image can be defined as:

$$S = \frac{h_O^2}{h_O^1} \tag{3.4}$$

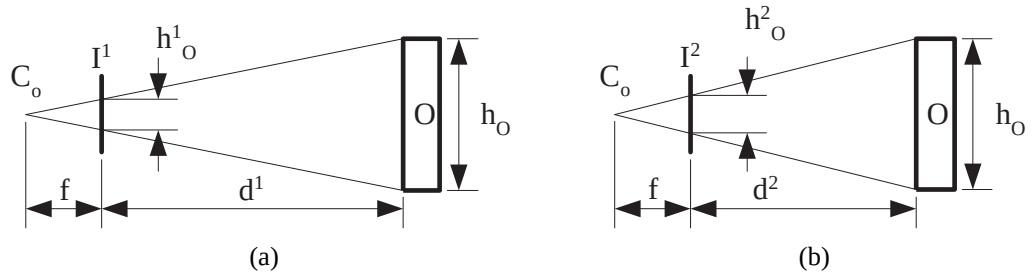


Figure 3.8: Effects of scale changes on the object in the images.

Where h_O^1 and h_O^2 are the height of the object, I^1 and I^2 are the image planes in the reference and sensed images of scale *S*, respectively. From Equation 3.4, it can be seen that the scale is proportional to the height of the object in the sensed image, as the height of the object in the reference image is constant. This is best illustrated in Figure 3.8, where Figure 3.8a shows the reference image, and Figure 3.8b shows a new image with scale *S*. Using properties of triangles, the following relationship was identified:

$$\frac{h_O^1}{f} = \frac{h_O}{f + d^1} \tag{3.5}$$

Where h_O is the actual height of the object and f the focal length of the camera. Similar equation holds for h_O^2. By rearranging Equation 3.5 in terms of h_O^1 and similarly for h_O^2 and substituting these into Equation 3.4, the distance between the camera and the object for a required scale can be obtained:

$$\begin{aligned} S &= \frac{h_O^2}{h_O^1} \\ &= \frac{f h_O}{f + d^2} \frac{f + d^1}{f h_O} \\ &= \frac{f + d^1}{f + d^2} \end{aligned} \tag{3.6}$$

In other words, the scale of an image is a function of the focal length of the camera and the distance from the reference and scaled images to the object. For the evaluation study, images used to study the effects of scale changes were taken in the range of $[1,3]$ at intervals of 0.5.

3.3.3 Tilt Changes

In order to take images of tilt changes while maintaining a constant distance of the camera to the object, a turntable was used. The turntable was constructed based on a phonograph turntable and is shown in Figure 3.9. As various Māori artefacts needed to be placed on the

Figure 3.9: Turntable used to capture images for tilt and viewpoint changes.

turntable, safety of the artefacts was considered carefully in the design and construction of the turntable. A rubber surface ensured that the objects would not slip during the rotation of the turntable, and therefore no fixture was required to prevent the artefacts from moving. In addition, by using a rubber surface, it reduced possible damages due to the contact of hard surfaces.

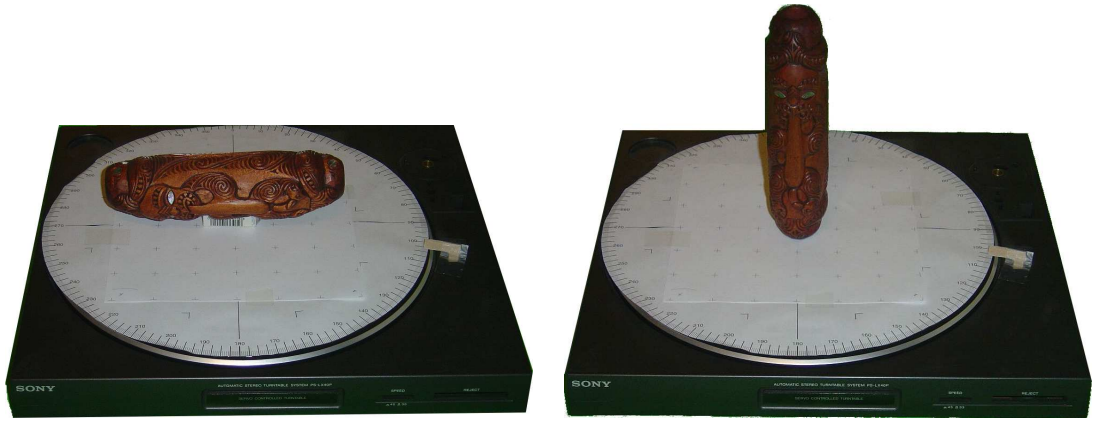
The images for tilt changes were acquired by placing the artefacts on the turntable, but rotated 90°. This approach was chosen over moving the camera as it was difficult to maintain a constant distance between the camera and the object if the camera was moved. The use of a turntable is a much more efficient method compared to the design and construction of a rig similar to the Stanford spherical gantry [148] to move the camera in the desired trajectory.

3.3.4 Viewpoint Changes

Similar to images for tilt changes, the images for viewpoint changes utilised the turntable shown in Figure 3.9. Instead of placing the objects on the turntable and rotated 90° as it was done for the images for tilt changes, the objects were placed up-right. The hardware setup for capturing images for both tilt and viewpoint changes is shown in Figure 3.10. The same reason applied for the choice of the turntable over a rig, which moved the camera around the object in the desired trajectory. Previous studies have shown that viewpoint changes pose the biggest challenge for image registration algorithms [9], and this was therefore the main focus of attention in the evaluation study, as viewpoint changes account for the majority of image transformations when taking images for the purpose of 3D reconstruction. For the evaluation study, images used to study the effects of viewpoint changes were taken in the range of $[-90^\circ, 90^\circ]$ at intervals of 2.5°.

3.4 Image Pre-Processing

In order to conduct controlled experiments, it was desirable to eliminate as much variance as possible in order to focus on the evaluation of the performance of local descriptor processes. To achieve this, it was necessary to pre-process the images before these were used for



(a) (b)

Figure 3.10: Hardware setup for capturing images for: (a) tilt; and (b) viewpoint changes.

experimental work. Three pre-processing steps were involved, including: (a) background removal; (b) image undistortion; and (c) computation of the homography matrix. These steps ensured that the results obtained are fair and consistent for all the image registration algorithms evaluated and developed. The homography matrices also allowed for the verification of image matching results.

3.4.1 Background Removal

The backgrounds of all the images were removed for the following reasons. First, these backgrounds may either have a positive or negative contribution towards the performance of the methods discussed. By providing additional features, it either simplifies the task of registering images, or makes the registration process more difficult by introducing features similar to those found on the objects. Another reason was that for the images taken using the turntable, the location or orientation of the object changes with respective to the background when these images were taken. As a result of this, the relationship of two images described by the background in the images and the relationship of the two images described by the object in the images do not agree, thus affecting the accuracy of methods. By removing the background in the images, these issues were avoided. Since the background did not contribute towards the matching accuracy of images, it was possible to concentrate on the matching accuracy of images which were solely dependent on the features found on the objects alone. To ease the removal of background in images, a chroma key was used in the image acquisition stage, and the backgrounds were later removed and checked using image manipulation software.

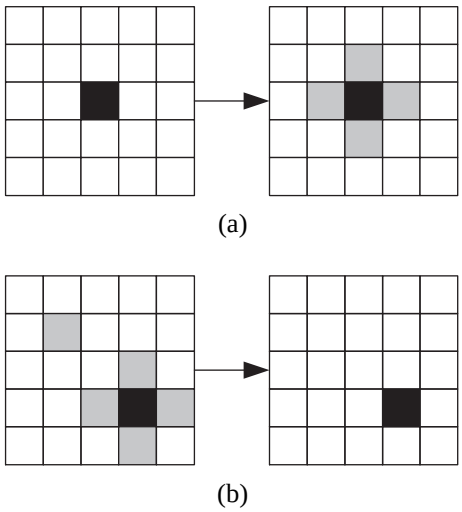


Figure 3.11: Processes used for removing noise introduced from imaging sensors: (a) dilation, which adds pixels to fill in gaps or missing pixels; and (b) erosion, which removes noise introduced from the imaging sensor.

3.4.2 Noise Removal

Because the image registration method of concern in this thesis is a feature-based method, noise in images such as the salt and pepper noise are not of great concern. However, as it will be discussed in Section 4.4, the hybrid local descriptor method developed is a combination of both area- and feature-based methods, and as such, noise removal was performed in order to enhance the performance of the method. This is achieved by the use of a median filter, as well as dilating and eroding the images [174]. Dilation and erosion are demonstrated in Figure 3.11. In Figure 3.11a, the image is dilated by adding surrounding pixels to existing ones to fill gaps or missing pixels after background removal. In Figure 3.11b, the image is eroded by removing the border pixels and in the case of individual pixels, as shown in the top left of the first image in Figure 3.11b, this is removed completely as it is considered to be a salt and pepper noise.

3.4.3 Image Undistortion

Images taken by digital cameras are often distorted both radially and tangentially, due to the imperfection of the camera lens, affecting the geometry of the features of objects in images [119, 175]. To compensate for this distortion in the camera lens, the images needed to be undistorted before being used for experimental work. This was achieved by computing the intrinsic parameters of the camera, a 3×3 matrix describing the internal characteristics of the camera as shown in Equation 2.5 and using the distortion factor to undistort the images.

It has been shown that tangential distortion can often be ignored for industrial applications, and often only one radial distortion term needs to be considered, as more elaborate

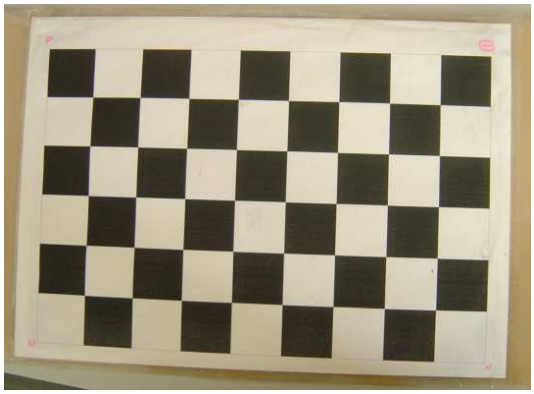


Figure 3.12: Calibration grid used to calibrate for the intrinsic parameters of the camera used for capturing images used for the experimental work in this thesis.

models would often result in numerical instability [175]. The intrinsic parameters were obtained by using a calibration grid, which contains grids of known geometry, as shown in Figure 3.12. A series of images of the calibration grid were taken at different locations, and for each image, the location of the corners in the grid were automatically extracted and these corners were used to compute the intrinsic parameters by using the Open Source Computer Vision Library [176]. The images can be undistorted by:

$$\hat{x} = x_c + s(x - x_c) \tag{3.7}$$

Where x is the measured location of a pixel along the x-axis, x_c is the radial distortion centre along the x-axis and $\hat{x}$ is the corrected pixel location. A similar equation holds for $\hat{y}$.

3.4.4 Homography Matrix

The homography matrix defines the spatial relationship of two images in terms of rotation and translation. The homography matrix for all image pairs acquired in this research needed to be computed, as the homography matrix allows for checking of the correctness of matches of local descriptors from image pairs. This was achieved by projecting the location of local descriptors from one image onto another and comparing the location of the projected and matched local descriptors. In order to compute the homography matrix, for each image pair, at least four corresponding point pairs were manually selected. A corresponding point pair refers to two points, one from each of the reference and sensed image, that describe the same point in the 3D space. As the homography matrix is a 3×3 matrix defined up to scale, a minimum of eight pieces of information is required [119] in order to compute the matrix. This means that at least four point pairs are required, as each pair consists of two pieces of information: the location of the points in the x and y directions, respectively. In practice, more than four pairs are often used to ensure the accuracy of the homography matrix. To compute the homography matrix from an over-fitted equation, the RANSAC algorithm in

combination with a least squares fit approach was utilised which iteratively checked whether a point pair should be used in computing the homography matrix or not.

3.5 Results and Discussion

The performances of local descriptor processes were evaluated using the criterion discussed in [13], which are based on the number of correct matches, mismatches and correspondences for an image pair. A plot of the recall versus 1-precision was computed for each image pair used for the experiments. In addition, the recall value for each image pair was plotted against the viewpoint changes to study the change of performance of local descriptor methods as the image viewpoint changed. Recall is a measure of how well the local descriptor processes performed, based on the ratio of the number of correct matches and the total number of corresponding regions, determined by the overlap error [85]:

$$\text{recall} = \frac{\text{number of correct matches}}{\text{total number of correspondences}} \tag{3.8}$$

An overlap of 50% was used in this research, as suggested in [85]. The 1-precision is a measure of accuracy and is defined as:

$$\text{1-precision} = \frac{\text{number of incorrect matches}}{\text{total number of matches}} \tag{3.9}$$

These two measures were used in evaluating the performance of local descriptor processes in the evaluation study, as these measures are the common choice for existing studies. They allow for the comparison of the performance of local descriptor methods with images used in this thesis and images of other objects used in previous studies. However, the recall measure takes into account the performance of region detectors. Because the local descriptor process was divided into three separate stages and studied separately in this research, and as a result of this, the performance of the region detector methods was not of interest and therefore the recall measure was of no practical use. The recall measure was therefore discarded for the experiments conducted in later chapters.

In addition, the precision, instead of 1-precision, was utilised in later experiments since intuitively, precision is a better representation of performance. The precision is also referred to as the ‘matching accuracy’ in later chapters. Also note that the effects of tilt changes were not studied in the evaluation study, as it was felt that due to its similarity in nature with viewpoint changes, an in-depth study of the effects of tilt changes for the evaluation study was not required. The effects of tilt changes were, however, studied for the experiments discussed in later chapters, and it was therefore introduced in this section along with the other three image transformations.

The three image transformations that were studied for the artefacts are shown in

Figure 3.1, and the results are shown in Figures 3.14-3.17. Due to the similarity in the trend of the results, and because the study aimed to compare the methods involved against each other, the discussion of the results in the following sections refer to the results for the flute artefact. Similar trends were observed for the other artefacts and the discussions presented are applicable to the results presented in Figures 3.15-3.17.

3.5.1 Recall versus 1-Precision Plot

Before discussing the results from the evaluation study, it is first necessary to explain how the recall versus 1-precision plot is useful in analysing the performance of local descriptor processes. Figure 3.13a shows three types of possible outcomes from the recall versus 1-precision plot. The ideal result is a vertical line along the y-axis, indicating that regardless of the number of correspondences, which is a function of the overlapping error, the number of correct matches is always the same as the number of total matches. In other words, the precision is always one, or that the 1-precision value is always zero. This is the ideal case and is almost never observed in real-life applications, as inaccuracies will almost always exist, regardless of how small they may be.

The worst case scenario is a horizontal line along the x-axis, meaning that the number of correct matches is always zero, in other words regardless of how many regions the algorithm identifies, none of these regions will be correctly matched. In this case the recall value does not change as the recall is defined as the number of correct matches over the number of correspondences, and since the number of correct matches is always zero, the recall is subsequently always zero. While rare, this is possible in real-life applications where the algorithm fails to identify any correct matches.

The typical result is a curved line that increases in value as the 1-precision increases. This is the most common result and often, but not always, contains a curve which has a steep initial gradient which then flattens and increases slowly as the 1-precision increases. One of the possible reasons for the change in gradient of the curve is the similarity of the features in the different regions of the images, thus making it difficult to distinguish and identify correct matches in later parts of the registration process [13].

Figure 3.13b shows an example of how the recall versus 1-precision plot should be interpreted. In general, the more the plot shifts towards the upper left corner, in other words towards the ideal curve, the better performance can be expected. On the other hand, if the plot consists of low x values for most of the curve, as observed in some results shown in Figure 3.14, then the performance of the local descriptor process is undesirable and improvements are needed in order to make these methods more robust.

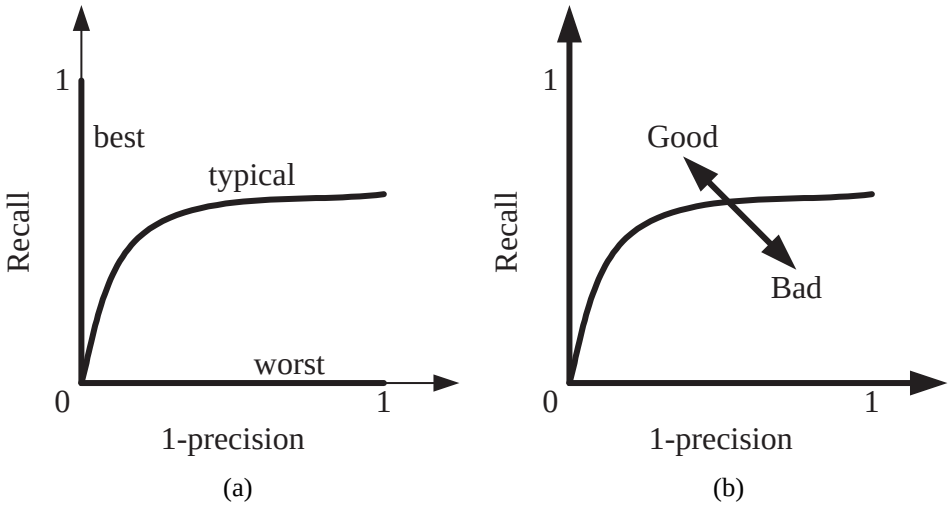


Figure 3.13: Recall versus 1-precision plot showing: (a) the three possible outcomes; and (b) how the plot should be interpreted.

3.5.2 Rotation Changes

Figure 3.14a shows the recall plot for rotation changes. The figure shows minor changes in the performance of local descriptor processes as the angle of rotation increased. An explanation to these changes is that due to the way the local descriptor processes evaluated are constructed [84], the recall values are higher when the angle of rotation of the image is aligned with the angles at which the vectors of these local descriptors were assigned. However, due to the magnitude of changes being relatively minor, it is most suiting to conclude that the performance of these local descriptor processes were consistent for rotation changes.

3.5.3 Scale Changes

Figures 3.14b and 3.14c show the results for scale changes. The recall versus 1-precision plot shown in Figure 3.14c is near straight with a slight curve, indicating that the performance of the local descriptor processes are constant and do not change dramatically with changes in scale. The recall plot shows that initially, changes in scale have little effect on the performance, however as the scale change increased, it was found that the performance degraded sharply. Further study of the images and the local descriptors computed from these images indicated that the drop in the recall value was due to the loss of information in the images, as a result of the images being taken from a further distance away from the object. As a result of this, it was not possible for the methods evaluated to identify the same set of features from the images due to this loss of information with the increased distance, resulting in a lower recall value compared to scale changes of lower magnitudes.

The relatively good performance of local descriptor processes for both rotation and scale changes compared to viewpoint changes, to be discussed in the section to follow, is in agreement with results from a previous study [173]. In the study, it was reported that the local descriptor processes are robust against rotation and scale changes regardless of the type of object under study. The high performance of local descriptor processes for images with rotation changes was expected, as the local descriptors are normalised in terms of orientation prior to being used to identify correspondences in image pairs. As such, rotation changes should, in theory, have no effect on the robustness of the algorithms. As the local descriptors are also normalised for scale, robustness against scale changes was also expected and this is partly true, where for scale changes of up to two the performance did not vary significantly. The poor performance of the algorithms for scales of higher values was identified to be due to the removal of backgrounds in the images, and the object does not cover the entire image for images of larger scales. This means that the size of the object in the images is smaller, and therefore less information from the images are available compared to the evaluation studies in [13, 91].

3.5.4 Viewpoint Changes

Viewpoint changes often posed the greatest challenge for image registration algorithms in general and this is reflected in the results obtained. Figures 3.14e and 3.14f show the recall values plotted against the 1-precision values for two different viewpoint changes: 5° and 10°. In Figure 3.14e, a slowly increasing curve is shown which indicates that the performance was influenced by the degradation of images, in particular, changes in viewpoint angle. The near-horizontal curve in Figure 3.14f indicates that the performance was limited by the similarity of the features of the objects due to distortion of the objects in images, and the local descriptor processes can no longer distinguish between these features. From the results obtained for viewpoint changes of higher magnitudes, it is clear that as the angle of viewpoint change increased, the smaller the gradient of the curves became, as a result of the difficulties the local descriptor processes had in registering images under these circumstances.

Figure 3.14d shows the recall values for the different viewpoint angles, similar to Figures 3.14a and 3.14b. As can be seen, the recall values degraded rapidly as the viewpoint angle increased or decreased away from zero, indicating the poor performance of the local descriptor processes for dealing with images consisting of viewpoint changes. Note that the recall value for a zero viewpoint change was computed from two images of the artefacts taken from very close, however not identical, viewpoints. Close inspection of the matched local descriptor pairs show that the reason for a relatively low recall value for this zero viewpoint angle change is due to the way local descriptors were matched. For a low threshold value for matching using the threshold matching method which will be discussed in detail

in Chapter 5, not all the local descriptors which describe the same point in the 3D space were matched correctly, however if the threshold value was increased, then the number of mismatches increased which also decreased the recall value.

Mismatches are particularly problematic in images of objects such as the flute and tiki artefacts, due to the repetitive regions which exist on the surface of the objects. By analysing the mismatched local descriptor pairs, it was found that there are two issues associated with the mismatches. The first is that the local descriptors are not distinct enough and as a result, for regions which have similar appearances, the local descriptors computed are very similar, making it nearly impossible to distinguish between two different regions. Another issue is due to the way local descriptors were matched. By using the threshold matching method, all the features which exist in the difference vectors of local descriptor pairs were reduced down to scalar values and the comparison of local descriptor pairs were made based on the scalar values computed from the difference vectors. This method removed a lot of potentially useful information and led to a lower matching accuracy.

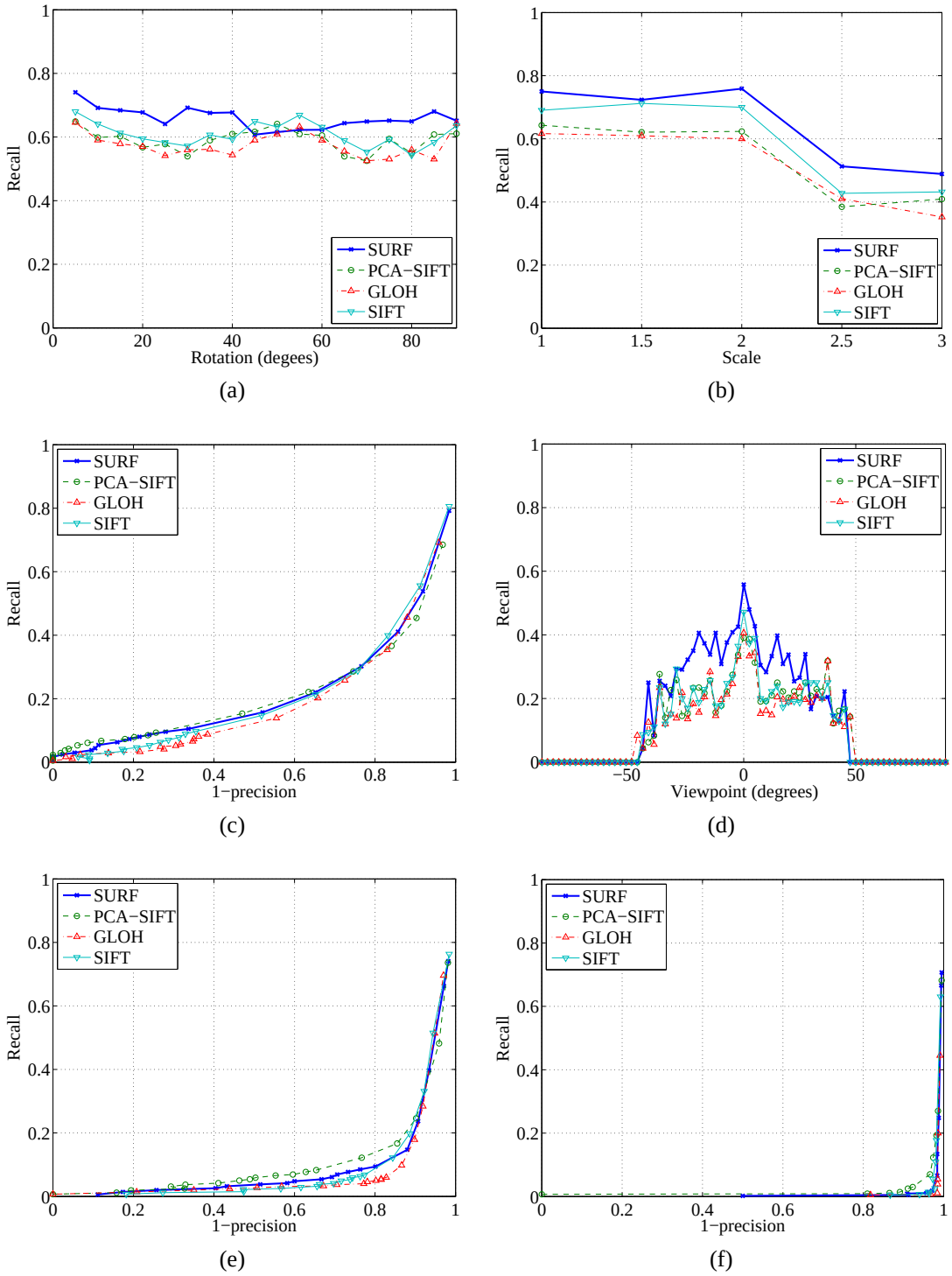


Figure 3.14: Image matching results using four different local descriptor processes for the flute artefact: (a) recall values for rotation changes; (b) recall values for scale changes; (c) recall versus 1-precision plot for scale changes; (d) recall values for viewpoint changes; (e) recall versus 1-precision plot for a 5° viewpoint change; and (f) recall versus 1-precision plot for a 10° viewpoint change.

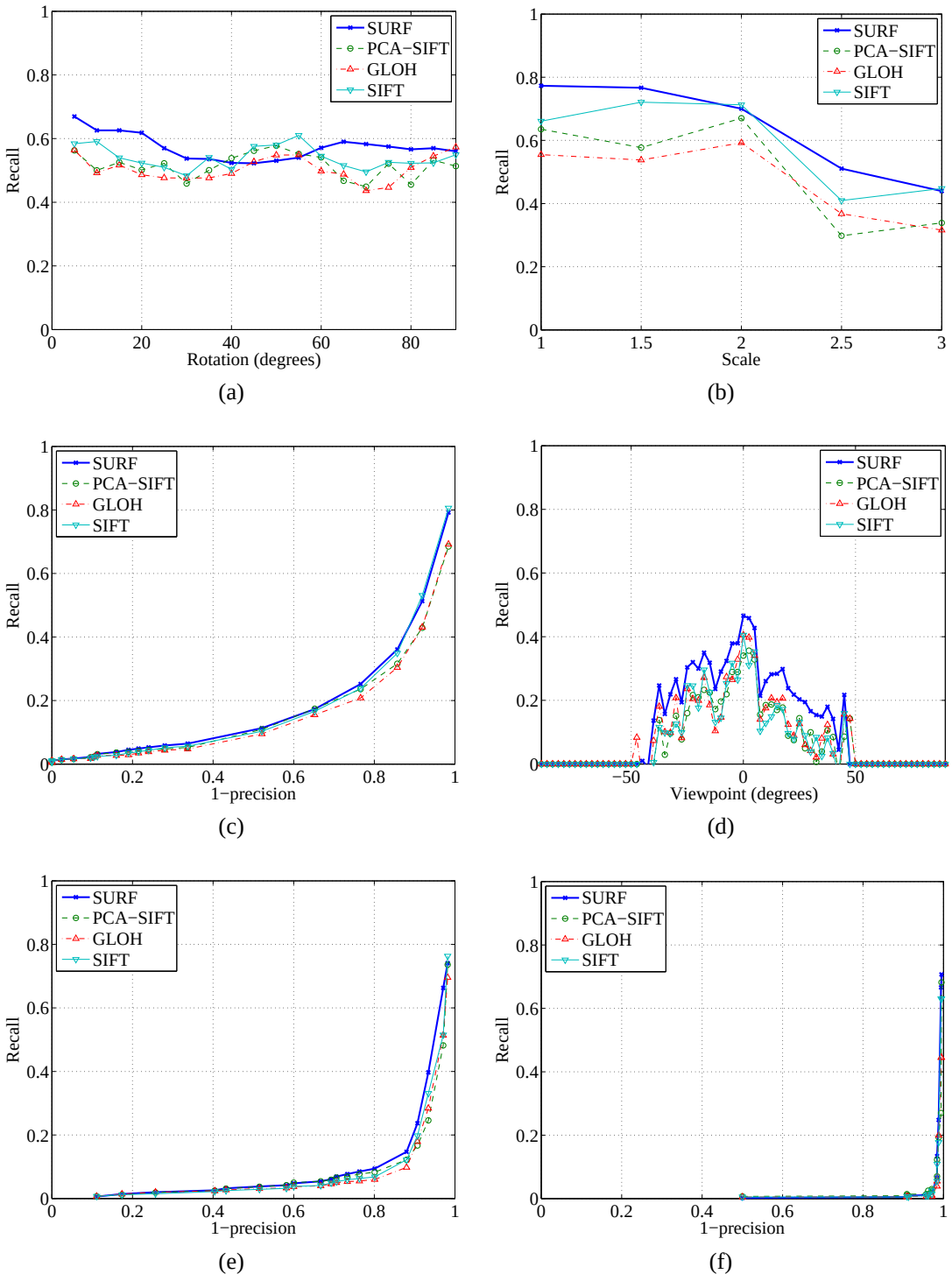


Figure 3.15: Image matching results using four different local descriptor processes for the patu artefact: (a) recall values for rotation changes; (b) recall values for scale changes; (c) recall versus 1-precision plot for scale changes; (d) recall values for viewpoint changes; (e) recall versus 1-precision plot for a 5° viewpoint change; and (f) recall versus 1-precision plot for a 10° viewpoint change.

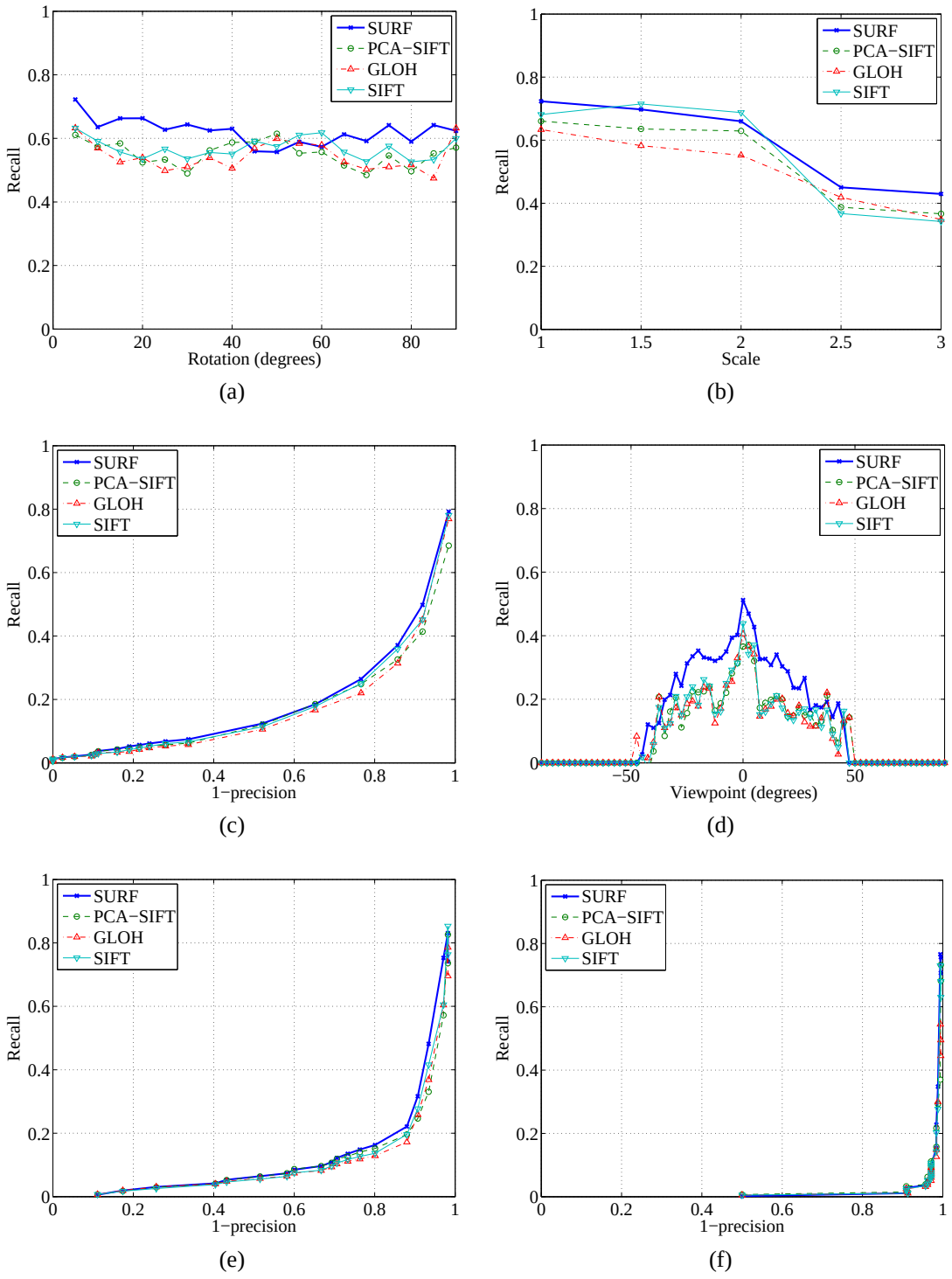


Figure 3.16: Image matching results using four different local descriptor processes for the wahaika artefact: (a) recall values for rotation changes; (b) recall values for scale changes; (c) recall versus 1-precision plot for scale changes; (d) recall values for viewpoint changes; (e) recall versus 1-precision plot for a 5° viewpoint change; and (f) recall versus 1-precision plot for a 10° viewpoint change.

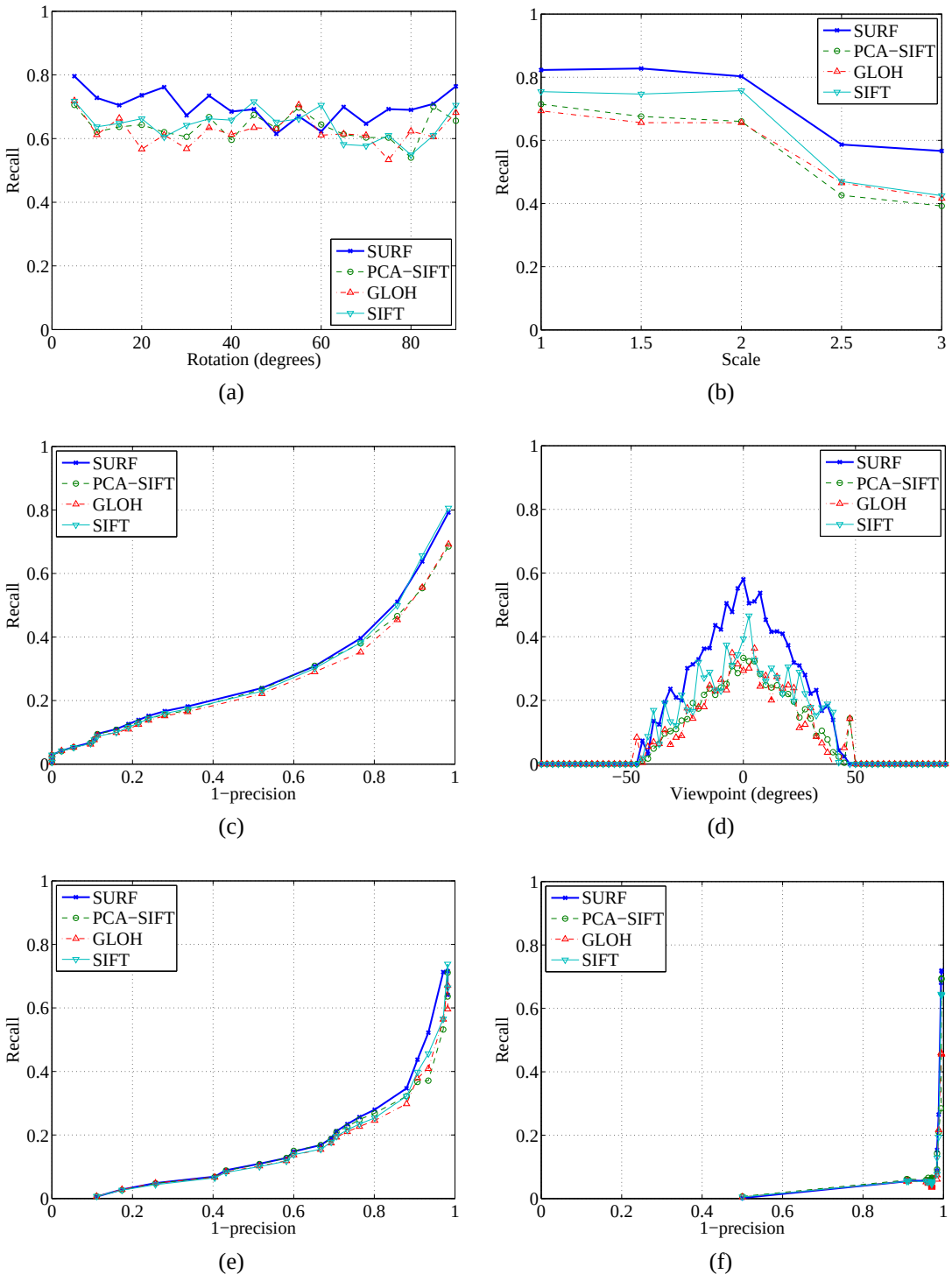


Figure 3.17: Image matching results using four different local descriptor processes for the tiki artefact: (a) recall values for rotation changes; (b) recall values for scale changes; (c) recall versus 1-precision plot for scale changes; (d) recall values for viewpoint changes; (e) recall versus 1-precision plot for a 5° viewpoint change; and (f) recall versus 1-precision plot for a 10° viewpoint change.

3.5.5 Ranking of the Algorithms Based on Accuracy

For the three types of image transformations studied, the four local descriptor methods compared had similar performances, however, it was clear that SURF out-performed the other three. This was in contrast to some previous studies which suggested that SIFT was the best local descriptor method. The improvement was particularly noticeable in Figure 3.14d which shows the performance of the local descriptor processes for viewpoint changes which, as discussed earlier, was the main image transformation of concern due to the nature of the images required for 3D reconstruction purposes.

From the results, the ranking of the local descriptor methods is as follows: SURF out-performed the other methods, with SIFT following close in second, followed by PCA-SIFT and GLOH which have similar performance for the image transformations studied. In addition to the higher accuracy achieved with the SURF local descriptor method, it was significantly faster than all the other methods, including SIFT. This is because SURF was designed with low computation time in mind, and due to the smaller number of features found in the SURF local descriptors which is 64D as opposed to SIFT and GLOH which are both 128D, the matching of these local descriptors by using the threshold method was faster. Even though PCA-SIFT is 36D, the computation time for computing the local descriptors was significantly higher than SURF and as a result, was slower than SURF overall. It should however be noted that the computation time of local descriptors was not of the main concern, as the aim of the research was to develop robust algorithms for registering images which can then be used for 3D reconstruction of the objects. In addition, as the registering of images for 3D reconstruction is an offline process, similar to the majority of image registration applications, real-time performance was not the main concern.

3.6 Issues with Existing Methods and Development of New Algorithms

Based on the results from the evaluation study, issues with existing local descriptor processes were identified. Close inspection of local descriptor processes led to the division of the process into three separate stages: (a) region detection; (b) local descriptor forming; and (c) local descriptor matching, as shown in Figure 2.4. By dividing the local descriptor process into three stages, it was possible to analyse each stage individually and identify the problems which exist in these stages. A close inspection of the fundamental theory of region detectors, a subset of feature detectors, suggested that it was difficult to improve on the performance of these techniques for objects without distinct features, and in these cases, other approaches such as laser scanners or manual alignment of the images would be more suitable approaches [18, 20].

From the issues identified in the evaluation study, it was concluded that more unique local descriptors were needed and the matching process should be improved. From this, The focus of the research was turned to improving the second and third stages of the process, namely local descriptor forming and local descriptor matching. This is discussed in Section 3.6.1. In addition, as it was shown from the evaluation study, local descriptor processes performed poorly for images which underwent significant viewpoint changes. This is consistent with previous studies which concluded that image registration algorithms are susceptible to changes in viewpoints. As a result, experiments were also conducted which studied alternatives to pure image registration techniques for registering images by means of an assisted image registration approach, while minimising the amount of end-user input as much as possible. This is discussed in Section 3.6.2.

3.6.1 Improvements to Local Descriptor Processes

Local Descriptor Formation

The aim of developing new techniques for computing local descriptors was to increase the uniqueness of the local descriptors formed. The uniqueness of the local descriptors affect how easy it is for the matching algorithm to accurately match local descriptors, and is considered the most crucial stage of the local descriptor process. From the evaluation study, it was clear that the mismatches were attributed with similarities of local descriptors from the same image which originated from different regions. By developing new algorithms that can produce more unique local descriptors, this issue can be reduced and ideally, eliminated. In contrast to conventional methods which compute local descriptors from greyscale images, it was proposed that colour images be used, as there are numerous benefits from using colour images compared to greyscale images. This will be discussed in detail in Chapter 4, where the developed algorithms are presented, and experimental work conducted that show the developed algorithms out-performed existing methods.

Local Descriptor Matching

The second issue identified in existing local descriptor processes was the way local descriptors were matched. By using the commonly utilised threshold matching method, features from the difference vectors of local descriptor pairs are reduced to scalar values and the correctness of matches of local descriptor pairs are based on these scalar values. Despite this downside, the matching of local descriptors has been a much neglected area in the local descriptor process. It is not hard to imagine the amount of information that are lost in this transition from vectors to scalar values, and therefore this research explored new methods for matching local descriptors. Instead of reducing the dimensionality of the difference vectors of local descriptor pairs to scalar values and effectively discarding potentially useful

information, the difference vectors were instead fully utilised. A local descriptor matching method utilising SVM was developed and will be discussed in detail in Chapter 5.

3.6.2 Assisted Image Registration

In addition to improvements made to local descriptor processes, another area which was explored was to develop an 'assisted' image registration method. As it is a well-known fact that image registration methods have limited success when dealing with large magnitudes of viewpoint changes [9], it was of interest to determine how much improvement can be gained by having an assisted approach, where the end-user can assist in the alignment of images, while reducing the effort and specialised knowledge in computer vision required by end-users.

To this end, an user-assisted programme was developed which requires a minimal amount of input from the end-user, however with a significant increase in the matching accuracy of images even in the presence of large magnitudes of viewpoint changes. This will be discussed in detail in Chapter 6.

3.7 Conclusions

The evaluation study highlighted issues with existing local descriptor processes. New algorithms needed to be developed, keeping in mind the challenges posed by the objects studied, resulting in a more robust approach for registering images. The two terms used to discuss the results in the experimental work conducted throughout this research, namely accuracy and robustness, were defined which allows for a better understanding and interpretation of the results presented in this thesis.

Prior to presenting the results from the evaluation study, the experimental setup which was used for all the experiments conducted was presented. The evaluation study was then presented and from the results, it was concluded that viewpoint changes pose the biggest challenge for image registration algorithms in general, and is in agreement with existing literature. Based on the evaluation study, the areas which needed attention were discussed and it was concluded that the different stages of local descriptor process should be tackled separately. The focus of the research was placed on the local descriptor formation and matching stages. In addition to improvements to local descriptor processes in these two stages, an assisted image registration was also proposed to overcome the common issue of the inability of image registration methods in dealing with large viewpoint changes. This approach aimed to combine image registration algorithms with the human eyes and provide a more robust method for registering images while only requiring a minimal amount of user input.

The evaluation study carried out and discussed in this chapter is an important piece of work, as it provides a detailed understanding of the capabilities of existing algorithms. Without this understanding, it is difficult to improve on existing algorithms, since the issues that exist cannot be fully understood. Due to this importance, it can be concluded that this is an important contribution towards improving local descriptor methods.

Chapter 4

Colour and Hybrid Local Descriptors Methods

This chapter investigates local descriptor methods utilising colour images instead of greyscale images to improve the uniqueness of the local descriptors. A colour model was utilised in order to be invariant to illumination condition changes. Two local descriptor methods, namely colour local descriptor and hybrid local descriptor methods, were developed. Results from the experiments conducted show that these methods are more robust against a variety of image transformation and illumination condition changes compared to existing methods.

The world we live in is filled with many colours, and the technology for acquiring colour images have existed for over a hundred years, with the first known permanent colour photography taken by Maxwell in 1861 [8] as shown in Figure 4.1. Intuitively, it therefore makes sense that colour images, instead of greyscale images, should be used for computer vision applications. This is, however, not the case in many real-life applications and for many years the focus has been placed on greyscale images, whether it is for object recognition, image registration, 3D reconstruction or other aspects of computer vision [177, 178]. The main drawback of colour images and reason for the popularity of greyscale images has been the lack of computation power and the incapability of computers to process colour images, however, with the growing computation power of modern computers [179] it is now possible to develop a whole new range of methods for image processing applications. Integrating colour images with existing techniques is however not a straight-forward process, as colour images often suffer from illumination issues, which affect the appearance of object in images, thus affecting the performance of image processing techniques [180]. In addition, many image processing techniques have been designed only with greyscale images in mind and are not versatile enough to be used with colour images without significant modifications.

Existing studies that make full use of colour images for image registration have been scarce, and only a handful of work exist for local descriptors utilising the advantages of

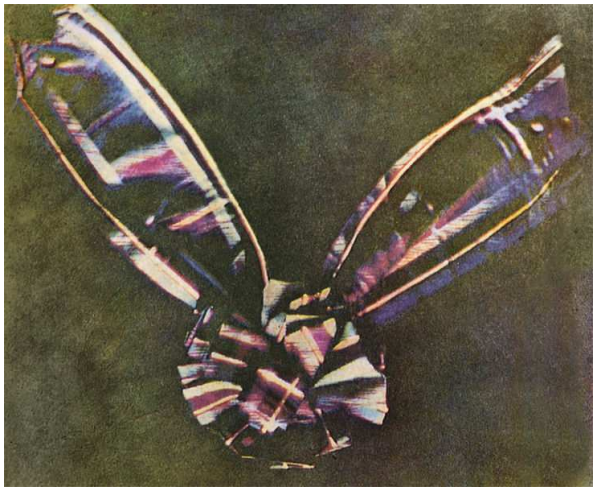


Figure 4.1: The first known permanent colour photograph, taken by James Clark Maxwell in 1861 (reproduced from [8]).

colour images. There is clearly a need for further research in this area to take full advantage of the computation power of modern computers today and the amount of additional data available from colour images.

This chapter presents two new local descriptor methods by integrating local descriptors with colour images and in doing so, the local descriptors are made more unique and easier for local descriptor matching methods to identify correct local descriptor pairs from image pairs. A colour model instead of colour images is utilised to handle changes in illumination conditions. The performance of the two methods were verified using images of Māori artefacts presented in Chapter 3, which contain both regions with repetitive features as well as regions that have a lack of distinct features.

This chapter is structured as follows. The limitations of greyscale images are discussed in Section 4.1 which forms the basis for the need for local descriptors utilising colour images. This is followed by a literature review in Section 4.2 on existing methods for integrating colour information with local descriptors, the issues faced and reasons why improvements were needed. The two new local descriptor methods, referred to as 'colour local descriptors' and 'hybrid local descriptors', are presented in Sections 4.3 and 4.4 which include an in-depth discussion on the various colour models and why colour models were preferred over using standard colour images. In addition, two feature-reduction methods to reduce the computation time of the colour local descriptor method are also developed. An uniqueness test for the local descriptor methods, as well as experiments conducted to verify the performance of the methods and the experimental setup are presented in Section 4.5. The results for these experiments, as well as a detailed discussion on the results are presented in Section 4.6 and the chapter is concluded in Section 4.7.

4.1 Limitations of Greyscale Images

Results from the evaluation study presented in Chapter 3 suggest that local descriptor methods have limited success in registering images of objects used as case studies in this research. Close inspection of the registration results shows that the matching accuracy can be further improved, since many local descriptors were mismatched due to the ambiguity of these local descriptors. This is best illustrated in Figure 4.2, where local descriptors of four different regions from an image pair are shown graphically. Figure 4.2a shows a SURF local descriptor computed from a region in the sensed image, and this local descriptor is incorrectly matched to Figure 4.2b, a SURF local descriptor from the reference image. Figure 4.2c shows the local descriptor from the reference image, which is the correct match for the local descriptor in Figure 4.2a. As can be seen in the figures, the local descriptors in Figures 4.2b and 4.2c appear very similar and it is difficult to distinguish between the two. A more unique local descriptor method, which has more variations in the individual vectors of the local descriptor is shown in Figure 4.2d. The issue faced in Figure 4.2 was not unique, and can be found throughout the matched local descriptor pairs in any given image pair in the experiments conducted in Chapter 3.

In order to increase the uniqueness of the local descriptors, the images used in the experimental work were carefully examined. By comparing the original, *RGB* images with their greyscale counterparts, it is clear that a large amount of data is lost in the conversion from the *RGB* colour space to a greyscale one. A *RGB* image is typically converted to its greyscale counterpart by:

$$Y = 0.3 \times R + 0.59 \times G + 0.11 \times B \tag{4.1}$$

Where R, G and B are the pixel intensity values of the red, green and blue channels in the *RGB* image, respectively, and Y is the pixel intensity value in the greyscale image. It should be noted that the weights for computing a greyscale image presented in Equation 4.1 are typical values used, and the exact weights are dependent on the choice of *RGB* primaries [181]. From Equation 4.1 it is clear that this method of conversion is prone to ambiguity in the greyscale image computed as multiple combinations of *RGB* values can result in the same value of Y. An example of this is shown in Figure 4.3. Figure 4.3a is the original *RGB* image and Figure 4.3b shows the greyscale image computed using Equation 4.1. These figures demonstrate that a wide variety of colours are represented by the same greyscale representation and while the figures shown is an extreme case, it demonstrates the incapability of greyscale images in truly representing the colour captured by modern cameras. This limitation meant that a search for better local descriptors that utilise colour images was required.

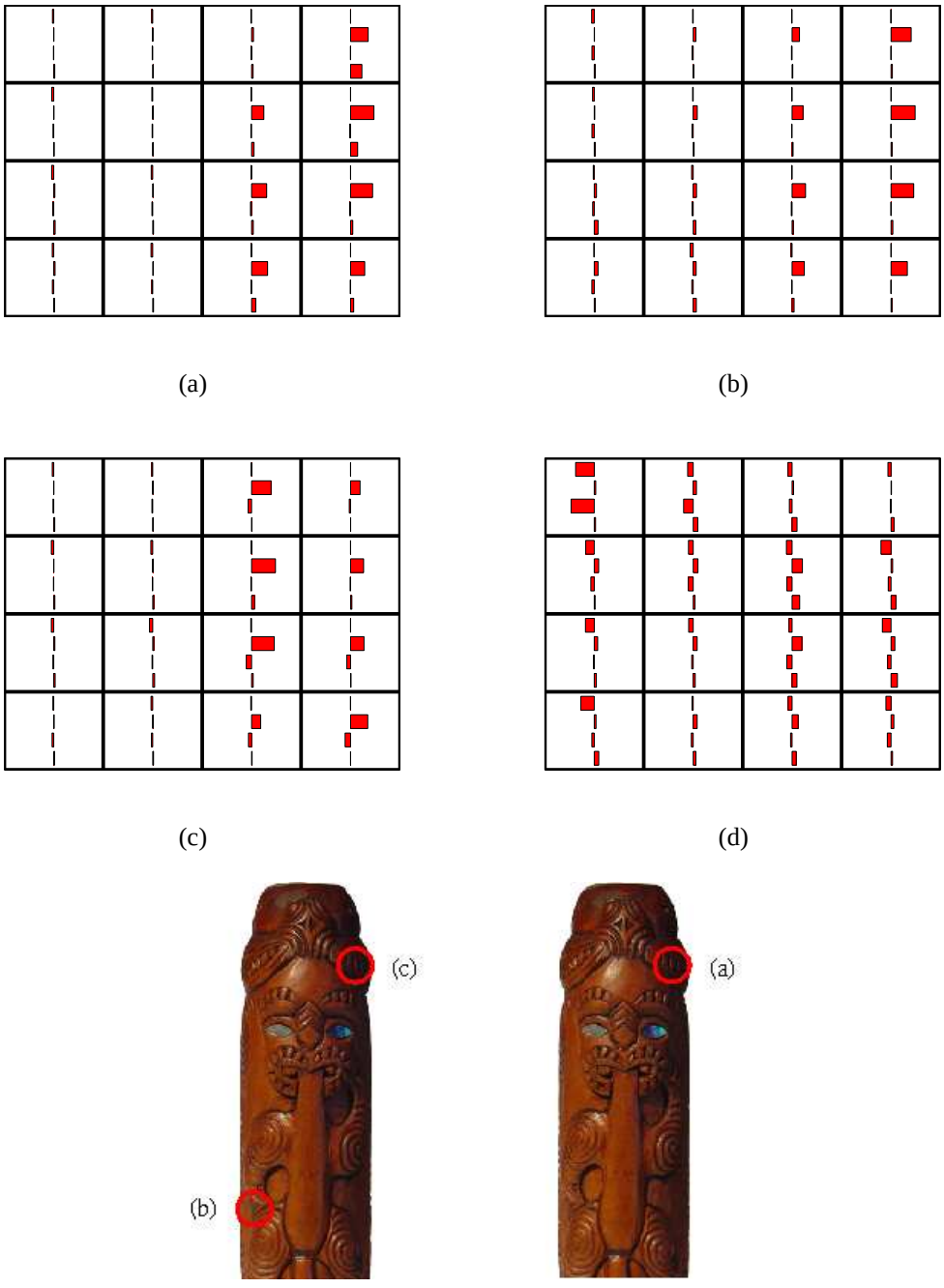


Figure 4.2: SURF local descriptors for: (a) local descriptor from the sensed image; (b) local descriptor from the reference image to which (a) is incorrectly matched to; (c) local descriptor from the reference image which is the correct match for (a); (d) an example of a more unique local descriptor; and (e) images containing the local descriptors shown in (a)-(c).

(a) (b)

Figure 4.3: The incapability of greyscale images for representing colour images: (a) the different colours which appear as the same shade in a greyscale image; (b) greyscale image of (a).

4.2 Review of Local Descriptors Based on Colour Images

Published research studies on the effects of combining local descriptors with colour images have been scarce. The primary focus of development in local descriptor methods is still very much in using greyscale images. However, previous studies have shown that by combining the additional data available from colour images with local descriptors, the performance can be improved and as such, it was of interest to pursue and further develop local descriptor methods utilising colour images. A comprehensive review of local descriptor methods has already been presented in Chapter 2, and this section discusses two more studies on local descriptors based on colour images.

Abdel-Hakim and Farag [98] proposed the Colored SIFT or CSIFT local descriptor method which is an extension to the SIFT local descriptor method. CSIFT combines SIFT with a colour model based on the Kulbelka-Munk theory [99] which models the reflected spectrum of coloured bodies. To detect a set of interest regions, the colour invariant images from the colour model are used and the extrema in the difference of the Gaussian pyramid are used for the interest points. Instead of the gradients of greyscale images as used in the SIFT local descriptor method, the gradients of the colour invariants are used to construct the CSIFT local descriptors. Experiments conducted showed an improvement over the SIFT local descriptor method in the repeatability of the features under different illumination directions and intensities [98].

Weijer and Schmid [182] also studied the effects of combining local descriptors with colour images. Different colour models were studied including normalised *RGB*, hue, opponent angles, spherical angles and comprehensive colour image normalisation [183] based on four criteria: (a) photometric robustness; (b) geometric robustness; (c) photometric

stability; and (d) generality. Performance gains were observed in all the experiments conducted. Based on the results, different colour models were suggested depending on the images concerned. For scenes with saturated colours, it was suggested that the hue model should be used, and with less saturated colours, the opponent angle approach should be used.

One of the main issues in these studies was that from the evaluation study and results presented in Chapter 3, it was found that SURF out-performed SIFT for all image transformations studied, and it would be of interest to base the work on the SURF local descriptor method instead of the SIFT local descriptor method to take advantage of the better performance observed. Another issue was that the effect of changes in illumination colour was not studied, and this was a concern in the registration of image of Māori artefacts, as the illumination colour in the Auckland War Memorial Museum, where a vast number of Māori artefacts are stored and will be digitally reconstructed is not always of the same colour and this would have an impact on the performance of the local descriptor method used.

To overcome these issues, two new approaches combining local descriptors with colour images have been developed based on the SURF local descriptor method and a colour model invariant to changes in various illumination changes including colour and intensity. The colour local descriptor method uses colour models for computing the local descriptors, while the hybrid local descriptor method is a hybrid method consisting of both area- and feature-based methods. The two methods make use of the colour information available from the original colour images, as shown in Section 4.1. The colour model contains much more information than their greyscale counterparts and does not suffer from the issues which arose from the use of greyscale images, such as the incapability of distinguishing from two different colours due to the way greyscale images are constructed.

To provide a fair comparison between conventional local descriptor methods such as the SURF local descriptor method based on greyscale images and the two developed methods in this thesis based on colour images, the interest regions used in the experiments conducted in this chapter were computed from the same source for each image. The Harris-Laplace detector [184] was utilised using the greyscale images of the artefacts for the experiments. The Harris-Laplace was chosen due to its high performance in previous studies [182, 80] and by using the same set of interest regions, controlled experiments were possible. The comparisons were made between the performance of the local descriptors constructed using greyscale and colour images only and are independent of the interest regions computed.

4.3 Colour Local Descriptors

Figure 4.4 shows an overview of conventional local descriptor methods, which compute both the interest regions and local descriptors from the greyscale image, often converted from the original colour image. As can be seen, the colour information of the scene from the original

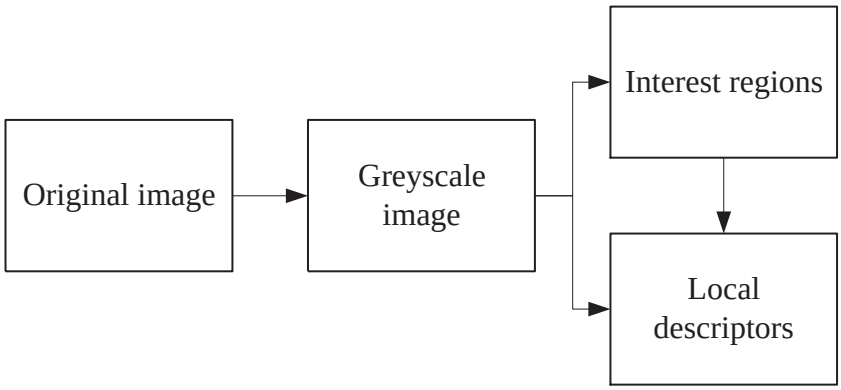


Figure 4.4: Overview of conventional local descriptor methods. Both the interest regions and local descriptors are computed from the greyscale image which in turn is computed from the original, colour image.

image is not utilised since the image is converted to its greyscale counterpart before any computation takes place. The first method in this chapter aimed at utilising this discarded data and is referred to as the colour local descriptor method, which are local descriptors computed from colour models instead of greyscale images. This approach was motivated by previous study [98], however improvements were made to further enhance the capabilities of the local descriptor method in dealing with illumination condition changes. The colour local descriptor method was developed with the aim of having a better representation of specific regions from the images, achieved by increasing the uniqueness of local descriptors by using colour images which contain more information about the scene compared to greyscale images. In addition, by integrating local descriptors with colour models, changes in illumination conditions can be properly handled.

An overview of the colour local descriptor method is shown in Figure 4.5. The major difference between the developed method and conventional local descriptor methods shown in Figure 4.4 is that the colour local descriptor method makes use of a colour model to fully utilise the information available from the original colour images. The colour local descriptor method is described as follows. First a reference image is converted to two images: (a) a greyscale image; and (b) a colour model of the original image. After the two images are computed, a set of interest regions are then detected using the Harris-Laplace region detector on the greyscale image. This process is repeated for the sensed image of the same image pair.

Once the interest regions have been computed for both the reference and sensed images, the next step is to construct the colour local descriptors from the colour model in the following manner. For the $m_1m_2m_3$ model used that will be discussed shortly, there are three colour channels, namely m_1, m_2 and m_3. For each colour channel, the local descriptor is computed by first defining a square region around the interest region with the orientation of the square region defined by the region detector. The square region is then divided

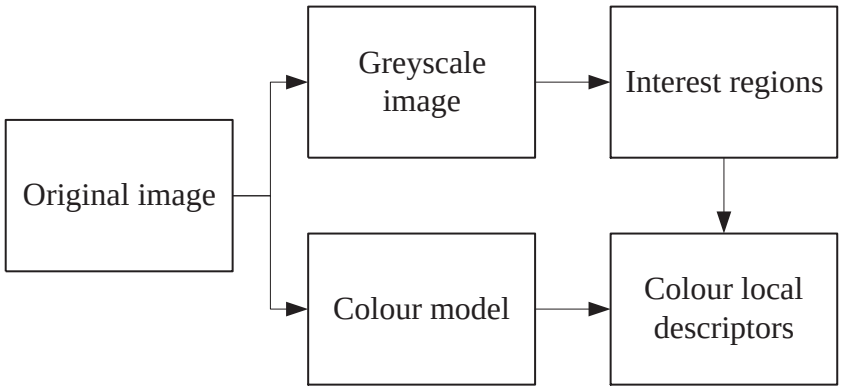


Figure 4.5: Overview of the colour local descriptor method. The interest regions are computed from the greyscale image while the colour local descriptors are computed from the colour model of the original, colour image.

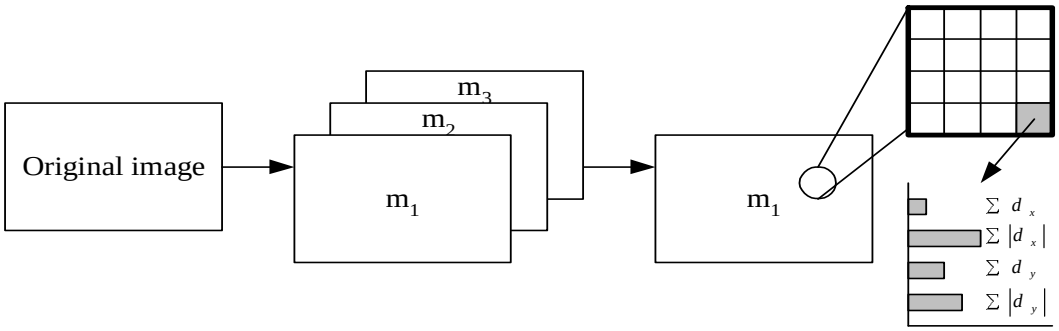


Figure 4.6: Overview of the computation of the colour local descriptor method. For each region in a colour channel, it is divided into $4 \times 4 = 16$ sub-regions. The Haar wavelet responses are then computed for each sub-region. The wavelet responses from each sub-region are combined to form the colour local descriptor.

into $4 \times 4 = 16$ sub-regions. For each sub-region, the Haar wavelet responses d_x and d_y are computed along the horizontal and vertical axes, which are defined in relation to the orientation of the square region. To increase the method's robustness, the wavelet responses are then weighted with a Gaussian centred at the centre of the interest region. Given these weighted wavelet responses, the following are computed: Σd_x, Σd_y, $\Sigma |d_x|$ and $\Sigma |d_y|$. This results in four vectors for each sub-region, resulting in a total of $4 \times 4 \times 4 = 64$D [84]. Once the three sets of local descriptors are computed, one for each of the three colour channels, these local descriptors are then combined to form the colour local descriptor for the interest region concerned. Given three colour channels, this results in colour local descriptors of $64 \times 3 = 192$D for each interest region. This process is shown graphically in Figure 4.6.

After sets of colour local descriptors have been computed for the reference and sensed images, these local descriptors are then matched to identify a set of corresponding local descriptors and this is achieved using the threshold matching method with Euclidean distance measure. The Euclidean distance measure is one of the most common measure for comparing local descriptors and is defined as:

$$L^2 = \left(\Sigma_{i=1}^{n} |x_i^1 - x_i^2|^2\right)^{1/2} \tag{4.2}$$

Where x_i^1 and x_i^2 are the i^{th} vectors of local descriptors of v dimensions from the reference and sensed images, respectively. In this case $v = 192$. L^2 is the Euclidean distance of the difference vector of a local descriptor pair. To determine whether a colour local descriptor pair is a correct match or not, a threshold is applied. For each colour local descriptor in the sensed image, the Euclidean distance of colour local descriptors from the reference image are computed and for any pair that has an Euclidean distance equal to or smaller than the defined threshold, the pair is considered to be a correct match:

$$L^2 \leq \tau_{L^2} \tag{4.3}$$

Where τ_{L^2} is the threshold for the Euclidean distance measure.

4.3.1 Colour Models

The main area of concern for the colour local descriptor method is the use of a colour model. One of the reasons colour images are not used in many image processing applications, aside from the increased computation requirement of the images, is that they are prone to illumination changes. In order to integrate colour information with local descriptors, first a method for utilising the colour information must be determined. This was affected by the many external factors when taking images of Māori artefacts which cannot always be controlled perfectly, even in a laboratory environment, which is not the case where many Māori artefacts are stored in the Auckland War Memorial Museum. The common conditions generally present real-life applications include changes in [180]: (a) illumination colour; (b) illumination direction; (c) illumination intensity; (d) specular highlights; and (e) viewing direction. The simplest method for utilising colour information for computing local descriptors is to compute the local descriptors using colour images directly instead of greyscale images as it has been done in many image registration algorithms. This is, however, unreliable in many cases as the *RGB* colour space cannot handle the various imaging conditions listed above. To gain a better understanding of the capabilities of various colour models, a list of common colour models, as well as their strengths and weaknesses are shown in Table 4.1.

The *RGB* colour space is the most widely known and commonly used colour space due to its simplicity and wide usage both in industry and the consumer world. Each colour can be defined by a combination of *R*, *G* and *B* values, for example black is represented by $(R,G,B) = (0,0,0)$ and white by $(R,G,B) = (M,M,M)$, where M is the maximum value for *R*, *G* and *B*. While the *RGB* colour space is very simple to work with, it suffers from the incapabilities in dealing with the various illumination conditions, and is therefore not

Table 4.1: List of the common colour models and the imaging conditions which they are invariant to. + denotes the colour model is invariant to the particular imaging condition and − denotes the colour model is susceptible to changes in the imaging condition.

	Illumination colour	Illumination direction	Illumination intensity	Specular highlights	Viewing direction
RGB	-	-	-	-	-
Norm. *RGB*	-	+	+	-	+
Intensity	-	-	-	-	-
Hue	-	+	+	+	+
Saturation	-	+	+	-	+
$c_1c_2c_3$	-	+	+	-	+
$l_1l_2l_3$	-	+	+	+	+
$m_1m_2m_3$	+	+	+	-	+
Opp. ang.	-	-	+	+	-
Spher. ang.	-	+	+	-	+

suitable for many image processing applications. A remedy for the issues in the *RGB* colour space is the use of the normalised *RGB* colour model. By normalising the *RGB* colour space, it is possible to achieve invariance to changes in illumination direction and intensity, as well as changes in viewing directions. The normalised *RGB* colour model is defined as:

$$r(R,G,B) = \frac{R}{R+G+B} \tag{4.4}$$

Where $r(R,G,B)$ is the normalised representation of the R channel. Similar equations hold for $g(R,G,B)$ and $b(R,G,B)$, the normalised G and B channels, respectively. Despite the simplicity of the *RGB* colour space and the normalised *RGB* colour model, however, for many computer vision applications colour images are often not used. The intensity image $I(R,G,B)$ is used. Intensity is the combined value of the *RGB* colour space and is computed by:

$$I(R,G,B) = R+G+B \tag{4.5}$$

This can be considered a special type of greyscale images computed using Equation 4.1. While a typical greyscale image has weights for the R, G and B channels of 0.3, 0.59 and 0.11, respectively, an intensity image has weights of $0.3\dot{3}$ for each of the three channels. Similar to greyscale images, intensity images suffer from the ambiguity issue when representing colour images.

Aside from intensity images, other popular methods for representing images include hue and saturation. Hue refers to colour impressions which are often described by names such as ‘red’, ‘green’, and ‘blue’. Hue is computed from the *RGB* colour space using the equation:

$$H(R,G,B) = \arctan\left(\frac{\sqrt{3}(G-B)}{(R-G)+(R-B)}\right) \tag{4.6}$$

As a colour model, hue is often used due to its simplicity and ability to deal with almost all types of illumination conditions as shown in Table 4.1, lacking only the ability to deal with changes in illumination colour. The other popular choice next to hue is saturation, which defines the colourfulness of a colour, and is the difference between a colour against grey. Saturation can be computed from the *RGB* colour space by:

$$S(R,G,B) = 1 - \frac{\min(R,G,B)}{R+G+B} \tag{4.7}$$

Similar to hue, saturation is invariant to various illumination conditions including changes in illumination direction, illumination intensity and viewing direction. Unlike hue, it is not invariant to specular highlights. In addition to the more common normalised *RGB*, hue and saturation colour models, many more sophisticated colour models have been presented over the years. One of the colour models proposed in [180] is the $c_1c_2c_3$ model, which is invariant to illumination changes for matte, dull surfaces:

$$c_1 = \arctan\left(\frac{R}{\max(G,B)}\right) \tag{4.8}$$

$$c_2 = \arctan\left(\frac{G}{\max(R,B)}\right) \tag{4.9}$$

$$c_3 = \arctan\left(\frac{B}{\max(R,G)}\right) \tag{4.10}$$

The assumption that objects consist of only matte, dull surfaces, however, is unrealistic in practice, and in order to accommodate for more object types, the effect of surface reflection, or highlights, are considered. The $l_1l_2l_3$ model was presented as an improvement over the $c_1c_2c_3$ colour model, and is invariant to illumination intensity and direction changes, as well as specular highlights for matte and shiny surfaces:

$$l_1 = \frac{(R-G)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \tag{4.11}$$

$$l_2 = \frac{(R-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \tag{4.12}$$

$$l_3 = \frac{(G-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \tag{4.13}$$

While this model is invariant to all the different illumination conditions that have been dealt with by other colour space and models discussed previously, one factor that was

not considered by many methods including the $l_1l_2l_3$ model is the change in illumination colour. To counter the effect of illumination colour changes in images, the effects of illumination colour changes was studied and based on this, a new model called $m_1m_2m_3$ was proposed [180]. This colour model aimed to be invariant to illumination colour changes, as well as the more obvious issues of illumination intensity and direction changes:

$$m_1 = \frac{R_{\mathbf{x}_1} G_{\mathbf{x}_2}}{R_{\mathbf{x}_2} G_{\mathbf{x}_1}} \tag{4.14}$$

$$m_2 = \frac{R_{\mathbf{x}_1} B_{\mathbf{x}_2}}{R_{\mathbf{x}_2} B_{\mathbf{x}_1}} \tag{4.15}$$

$$m_2 = \frac{G_{\mathbf{x}_1} B_{\mathbf{x}_2}}{G_{\mathbf{x}_2} B_{\mathbf{x}_1}} \tag{4.16}$$

Where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the image locations of the two neighbouring pixels for each given pixel. Without loss of generality, m_1 is used to derive the results that also hold for m_2 and m_3. By taking the logarithm of both sides:

$$\log\left(m_1\left(R_{\mathbf{x}_1} R_{\mathbf{x}_2} G_{\mathbf{x}_1} G_{\mathbf{x}_2}\right)\right) = \log\left(\frac{R_{\mathbf{x}_1} G_{\mathbf{x}_2}}{R_{\mathbf{x}_2} G_{\mathbf{x}_1}}\right) \tag{4.17}$$

and expanding (4.17):

$$\log\left(m_1\left(R_{\mathbf{x}_1} R_{\mathbf{x}_2} G_{\mathbf{x}_1} G_{\mathbf{x}_2}\right)\right) = \log\left(\frac{R_{\mathbf{x}_1}}{G_{\mathbf{x}_1}}\right) - \log\left(\frac{R_{\mathbf{x}_2}}{G_{\mathbf{x}_2}}\right) \tag{4.18}$$

Using (4.18), the colour ratios can be represented as the difference of two neighbouring pixels, $\mathbf{x}_1$ and $\mathbf{x}_2$:

$$d_{m_1}\left(\mathbf{x}_1, \mathbf{x}_2\right) = \left(\log\left(\frac{R}{G}\right)\right)_{\mathbf{x}_1} - \left(\log\left(\frac{R}{G}\right)\right)_{\mathbf{x}_2} \tag{4.19}$$

By considering these differences in a particular orientation between the neighbouring pixels, d_{m_1}, the finite difference differentiation is obtained, which is invariant to changes in illumination direction, intensity and colour. Similar equations can be obtained for m_2 and m_3 using the same approach. A disadvantage to this approach, however, is that unlike the $l_1l_2l_3$ model and the hue image, it is not invariant to specular highlights.

Two colour models discussed in [182] are the opponent angle and spherical angle models. The opponent angle is invariant to specularities in the case of white illumination and is defined by:

$$ang^{O_1} = \frac{R-G}{\sqrt{2}} \tag{4.20}$$

$$ang^{O_2} = \frac{R+G-2B}{\sqrt{6}} \tag{4.21}$$

This is however unstable around the grey axis and therefore an error analysis is applied, resulting in:

$$\delta ang^{O} = \frac{1}{\sqrt{(ang^{O_1})^2 + (ang^{O_2})^2}} \tag{4.22}$$

The spherical angle on the other hand is invariant to change in illumination directions [182] and is defined by:

$$ang^{S_1} = \frac{G_{\mathbf{x}}R - RG_{\mathbf{x}}}{\sqrt{R^2+G^2}} \tag{4.23}$$

$$ang^{S_2} = \frac{R_{\mathbf{x}}RB + G_{\mathbf{x}}GB - B_{\mathbf{x}}R^2 - B_{\mathbf{x}}G^2}{\sqrt{(R^2+G^2)(R^2+G^2+B^2)}} \tag{4.24}$$

Similar to the opponent angles, an error analysis is applied for stability:

$$\delta ang^{S} = \frac{1}{\sqrt{(ang^{S_1})^2 + (ang^{S_2})^2}} \tag{4.25}$$

In [182], it was suggested that the opponent angle should be used in the case of diffused lighting, and the hue image should be used in cases where saturated colours exist. While there exists many more colour models, it is impossible to cover all of these and therefore a selected few have been presented, and the selection of a suitable colour model was made from the list of models discussed.

In order to fully utilise the invariant properties of the colour models discussed above, it was important to first consider the imaging conditions that may be encountered. As many Māori artefacts are of high historical value, the images often need to be captured in environments where excessive lighting is not present so as to prevent damage to the artefacts. This is due to the fact that many artefacts are fragile and sensitive to strong illumination. Because of this, specular highlights are unlikely as diffused lighting is often used to evenly distribute the light onto the objects, avoiding concentration of light in one region. Also, in a museum environment it is not uncommon to have different colours and intensities of lighting present in order to showcase the artefacts fully across different locations of the museum. Since the aim was to develop a robust method that can be applied not only to images of Māori artefacts, but is also versatile and can be applied to other objects with similar features,

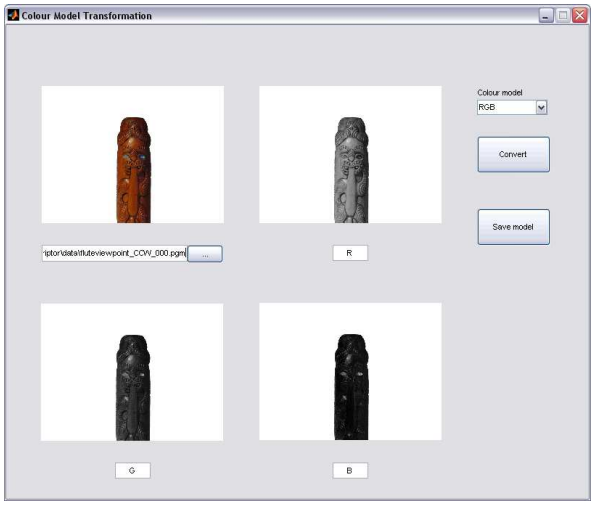


Figure 4.7: MATLAB programme for converting *RGB* images into various colour models.

it was desirable to acquire images of the artefacts without disturbing or moving the artefacts greatly if possible, in other words, the various illumination colours and intensities needed to be dealt with. Based on the conditions faced, it was determined that the $m_1m_2m_3$ model is the most suitable as this colour model is invariant to changes in illumination colour and intensity and in addition, is invariant to almost all other illumination conditions, only missing the ability to deal with specular highlights which, as discussed previously, is unlikely to be encountered and is therefore the most suitable model.

A MATLAB programme was developed to allow for a quick and efficient way of visualising the images transformed to the various colour models discussed, and was used to assist in determining the suitability of the colour models. This programme is shown in Figure 4.7. In the figure shown, the image is separated into the three colour channels of the *RGB* colour space: red, green and blue.

4.3.2 Feature-Reduction

A disadvantage of the colour local descriptor method that does not in fact affect image registration of images in many real-life applications is the increased computation time of the developed method, due to the increased number of features in the colour local descriptors. As many image registration applications being offline processes, real-time performance is often not required and this was therefore not considered a drawback. However, it is always desirable to reduce the computation time as this leads to a reduction in the overall processing time and potential reduction in the cost associated with the application of interest.

As the colour local descriptor method makes use of the $m_1m_2m_3$ colour model and is based on the SURF local descriptor method, colour local descriptors are 192D. This is three times the size of SURF local descriptors due to the three colour channels from the colour model utilised. The computation time of local descriptors does not increase dramatically with

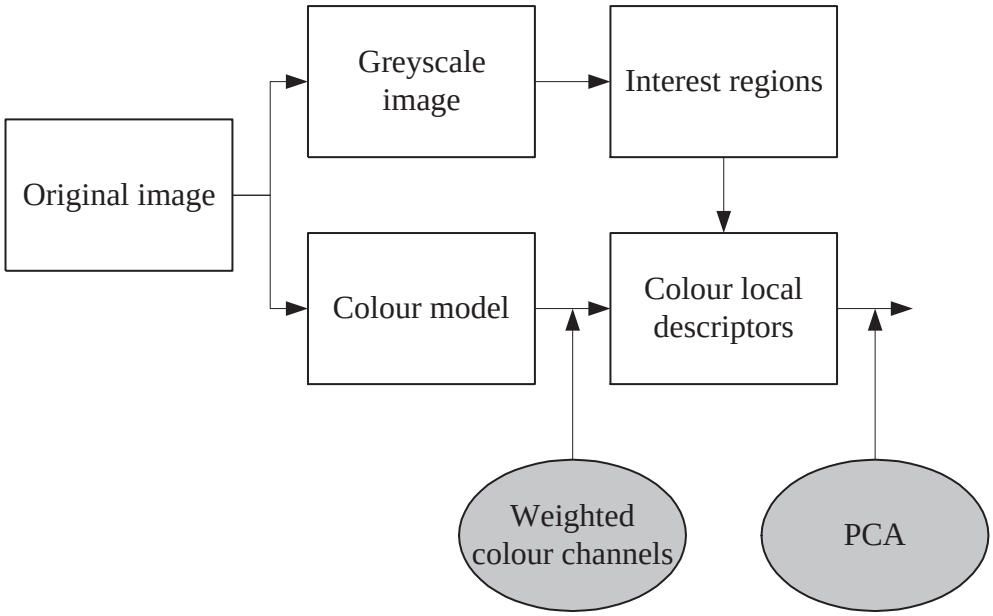


Figure 4.8: The two feature-reduction methods for the colour local descriptor method.

the increased dimensionality of the local descriptors, due to the relatively efficient method of computing [84], the focus was the placed on the matching process which can be time-consuming in comparison. By reducing the number of features in colour local descriptors, the computation time required for computing the Euclidean distance measure of difference vectors of colour local descriptor pairs can be reduced.

To reduce the size of these local descriptors, two methods were developed with the colour local descriptor method. The first method makes use of PCA, while the second is based on the weighting of the colour channels from the colour model. Note that the two dimension reduction methods discussed below only apply to the colour local descriptor method and not the hybrid local descriptor method to be presented in Section 4.4. This is due to the fact that the two local descriptor methods discussed take different approaches to constructing local descriptors using colour images, resulting in two local descriptor methods of different nature. From the experimental work conducted which will be discussed in Section 4.6, it was felt that the colour local descriptor method is the more practical approach of the two methods developed in this research, more work had therefore gone into further developing and improving the colour local descriptor method. Figure 4.8 shows how the two feature-reduction methods are integrated with the colour local descriptor method. The weighted channel method is applied after the local descriptors have been computed for each colour channel of the image, before the local descriptors are combined to form the colour local descriptors for the image. The PCA method on the other hand is applied after the colour local descriptors are formed and works with the complete colour local descriptors.

Principal Component Analysis

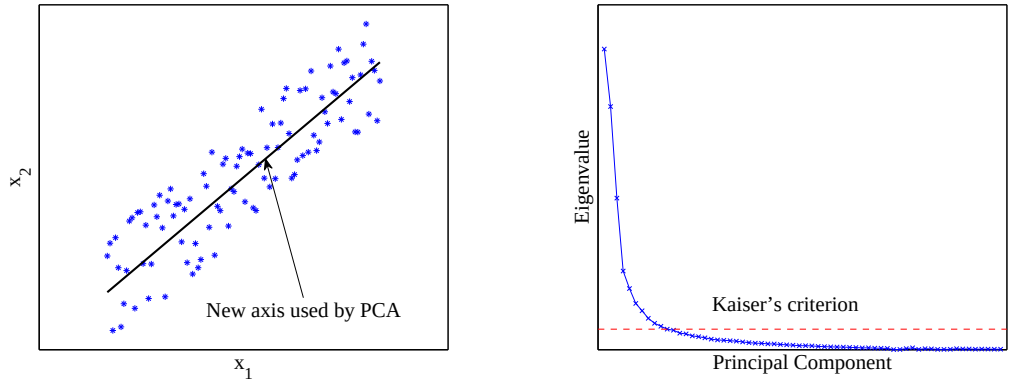
PCA is one of the most well-known and widely used feature-reduction methods in statistics for its simple, yet effective approach. PCA has been previously applied in local descriptor methods to reduce the number of features of local descriptors [97, 91], where larger-than-conventional local descriptors are reduced to be in line with the size of conventional local descriptors such as SIFT and SURF. In this research, PCA is applied in a similar manner to [97, 91] where the aim was to reduce the number of features of colour local descriptors in an attempt to reduce the computation time when comparing colour local descriptors, while maintaining the performance the original colour local descriptor method is capable of.

The process for applying PCA is as follows. Suppose each colour local descriptor has v dimensions, and there are m colour local descriptor pairs consisting of correctly matched colour local descriptors from the reference and sensed images, the data matrix $\mathbf{X}$ is then a $2m \times v$ matrix. This data matrix is mean-shifted by first computing the mean for each column, then shifting the data matrix by this vector of means. The covariance matrix of the data matrix can then be computed by:

$$\begin{aligned} \mathbf{C} &= cov(\mathbf{X}) \\ &= \frac{\mathbf{X}'\mathbf{X}}{m} \end{aligned} \tag{4.26}$$

By finding the eigenvalues and eigenvectors of the covariance matrix $\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{D}$, where $\mathbf{D}$ is the diagonal matrix of eigenvalues and $\mathbf{V}$ the eigenvectors, a set of principal component (PC) scores can be obtained, where PC scores = $\mathbf{X}\mathbf{D}$. The eigenvalues give indication on how much information the eigenvectors represent, and by analysing the eigenvalues, it is possible to determine how many PC scores or vectors should be retained in the local descriptors. Two methods for determining the number of components required are utilised, namely the Kaiser's criterion [185] and scree graph [186]. The Kaiser's criterion states that only those components that have eigenvalues greater than $\Sigma_i \frac{\lambda_i}{v}$ should be kept. One common problem with Kaiser's criterion, however, is that this method often retains too many components and therefore the scree graph is also used in assisting the component selection process. The scree graph is a plot of eigenvalues against the i-th component. If the corresponding eigenvectors of the eigenvalues are sorted in descending order, then the scree graph obtained is a declining curve which can be used to manually select the number of PCs by manually separating the 'large' and 'small' eigenvalues. This is determined by identifying where the 'elbow' of the curve is which is the point where the gradient of the curve changes significantly.

Figure 4.9a shows an example of how PCA can reduce the number of features for colour local descriptors. The data clouds in the figure are originally described by the two axes, x_1



(a) (b)

Figure 4.9: (a) How PCA is used to reduce the number of features in the colour local descriptor method; and (b) Scree-diagram used for selecting the number of PCs.

and x_2. By computing the eigenvalues and eigenvectors of the data matrix, PCA transforms the dataset to a new set of axis as shown in the figure, and it is now possible to describe the data clouds by using a single axis while maintaining the majority of the information about the dataset. Figure 4.9b shows an example of the scree graph and the Kaiser's criterion. The dashed line in the figure represents the Kaiser's criterion value $\Sigma_i \frac{\lambda_i}{v}$. As can be seen from the figure, the number of PCs from the two methods often do not match exactly and when selecting the number of PCs, it is important to experiment with different numbers of PCs suggested by these two methods.

An important factor is that in order to apply PCA for reducing the number of features in colour local descriptors, the covariance matrix of correctly matched colour local descriptors is required. Since it is not possible to obtain the covariance matrix when matching colour local descriptors, as this implies that the correct matches of the colour local descriptors are known and therefore the images would have already been registered, the covariance matrix needs to be estimated in a training phase. The training phase involves acquiring images of objects containing features similar to those found in the object to be registered, and by computing the homography matrix of the various image pairs, a set of training data consisting of correctly matched colour local descriptors can be obtained. By using this set of correctly matched colour local descriptors, it is then possible to obtain a covariance that is a good representation of the covariance matrix of the colour local descriptors to be matched in the new image pairs of the object of interest.

In the case that the object of interest is an one off, or the nature of the features found on the surface of the object is unknown, it is also possible to train the covariance matrix using images from a wide variety of objects. This approach of obtaining the covariance matrix is the same as it has been discussed in existing literature for both the PCA-SIFT [97] and GLOH [13] local descriptor methods.

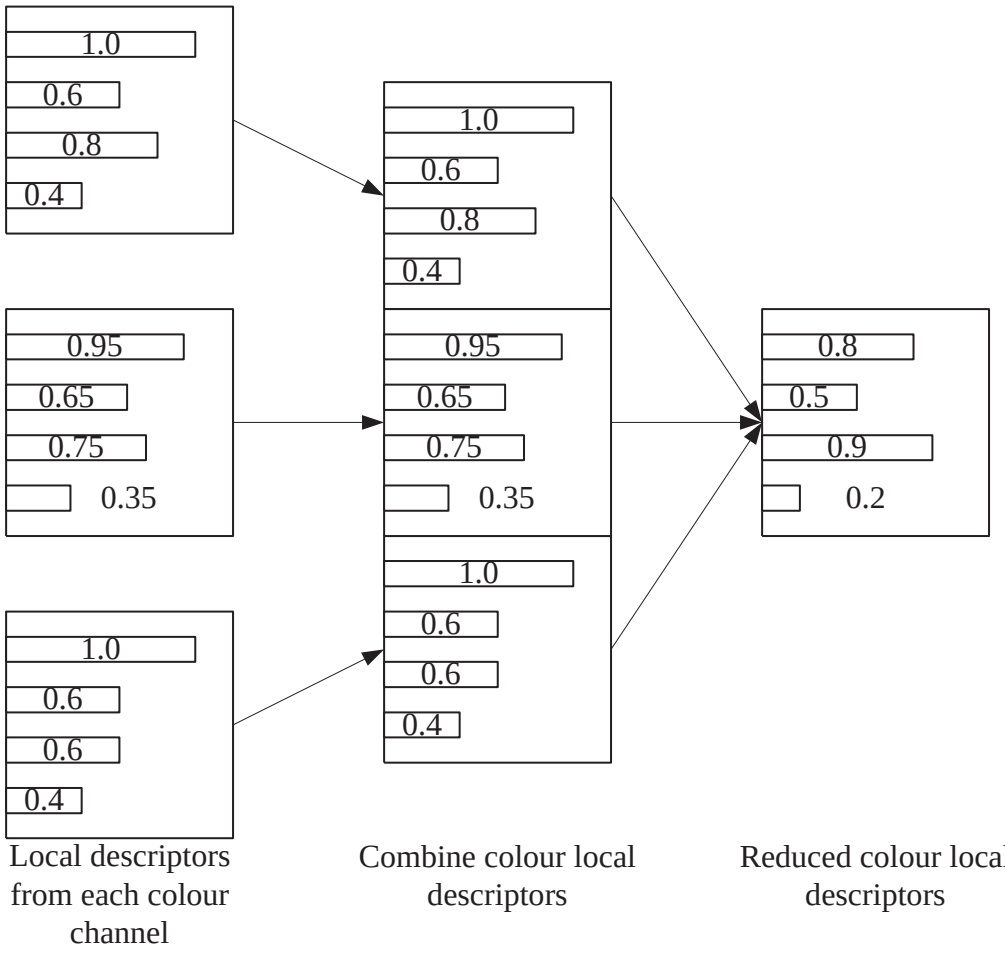


Figure 4.10: Example of how the feature-reduction method using PCA works on colour local descriptors. First, local descriptors are computed for each of the three colour channels. These are then combined to form the colour local descriptor. Finally, PCA is applied to re-represent the colour local descriptor with a reduced number of features.

Figure 4.10 shows a simplified example of how the feature-reduction method using PCA works. First the local descriptors for the colour channels are computed, these local descriptors are then combined together to form the colour local descriptors for each image, and finally, PCA is applied to reduce the number of features in the colour local descriptors.

Weighted Channels

The second method developed for reducing the number of features of colour local descriptors works by weighing the interest regions in the individual channels of the colour model, $m_1m_2m_3$, and based on the weights, the number of features of local descriptors from each channel is reduced before they are combined to form the colour local descriptors. This method works by analysing the variance of colour in each of the three colour channels, which affects the way local descriptors are computed. Local descriptors describe features in local regions, and work best when the local regions have distinct features, in other words, the

intensity of the pixels in a given region would ideally be changing across different locations of the region.

The weighted channel method is described as follows. First the interest regions and local descriptors for each colour channel of images from an image pair are computed. For each colour channel, the variance is computed for each interest region by using the pixel intensities inside the interest regions. This results in r variances for r interest regions in each colour channel. This process is then repeated for the other two colour channels. Given the three sets of variances, one for each colour channel, the weights of the colour channels can be determined by taking into account how much variability there are in each of the channels. The more variation means that the local descriptor for the colour channel is more unique and would therefore be of more use in the decision making of the correctness of colour local descriptor matches. To compare the variances from the three channels, the median of the three sets of variances is computed, and weights are assigned accordingly:

$$w_{m_1} = \frac{\tilde{x}(m_1)}{\tilde{x}(m_1)\tilde{x}(m_2)\tilde{x}(m_3)} \tag{4.27}$$

Where w_{m_1} is the weight assigned to the first colour channel, $\tilde{x}(m_1)$ is the median for the first colour channel and similarly for the second and third colour channels. Medians were used instead of means to eliminate the effects of outliers which means are prone to. The weights for the other two colour channels can be computed in a similar manner. The weights indicate how many vectors of the local descriptors from each colour channel would be used in the colour local descriptors, and the number of vectors is dependent on the size of the local descriptors desired:

$$v_{m_1} = w_{m_1} \times v \tag{4.28}$$

Where v_{m_1} is the number of vectors from the local descriptors computed using the colour channel m_1 and v the total number of vectors desired in the reduced colour local descriptor. Similar results are obtained for the other two colour channels. An example of this is as follows. If the weights for the three colour channels are $0.125, 0.625, 0.250$, then to obtain a colour local descriptor of size 64D, $0.125 \times 64 = 8$ vectors would originate from the m_1 channel, $0.625 \times 64 = 40$ from the m_2 channel and $0.250 \times 64 = 16$ from the m_3 channel.

Figure 4.11 shows a simplified example of how the feature-reduction method with weighted channels works. First local descriptors for each of the three colour channels are computed, the required number of features are then extracted from each colour channel depending on the weights assigned and these features are combined to form the colour local descriptors required.

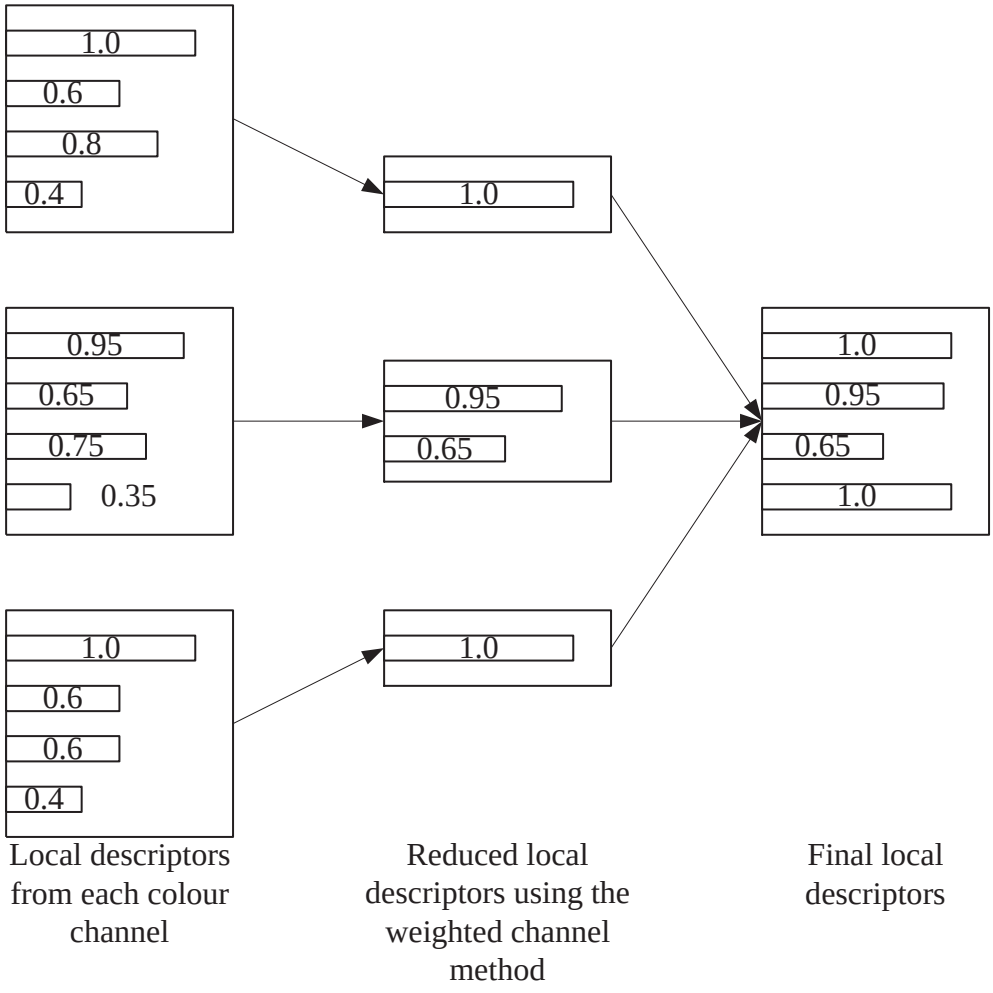


Figure 4.11: Example of how the feature-reduction method using the weighted channels method works on colour local descriptors. First, local descriptors are computed for each of the three colour channels. Weights are then applies to the local descriptors from each channel, and the required features are acquired from these local descriptors. Finally, the features are combined to form the colour local descriptor with a reduced number of features.

4.4 Hybrid Local Descriptors

4.4.1 Integration of Area-Based and Feature-Based Methods

The second method developed utilising colour images instead of greyscale images is the integration of greyscale local descriptors with colour regions. The concept behind this approach is that while feature-based methods are superior for feature-rich images, they often perform poorly in the case that images do not contain distinctive features, due to the incapability of feature-based methods to represent these regions. Area-based methods on the other hand are superior in these cases, as these methods compare the regions directly and do not attempt to condense the information available and describe the regions using a set of features. Area-based methods however do suffer from the potentially high computation time involved due to the nature of area-based methods. Since area-based methods compare each region directly, often in a pixel-to-pixel manner, when image transformations in the form of scale or rotation changes exist, a large number of comparisons need to be made by resampling the image for every comparison. As each comparison is computationally expensive, the large number of comparisons often make these methods impractical for real-life applications. By utilising the scale and orientation information available from local descriptors, however, it is possible to reduce the number of comparisons required, and hence reduce the computation time dramatically, making it a more practical approach.

An overview of the method is shown in Figure 4.12. Similar to the colour local descriptor method, the original image is first converted to a greyscale image and a colour model. Local descriptors are then computed for each interest region using the greyscale images. For each image, the region in the colour model described by each interest region is then recorded alongside its corresponding local descriptor, in the form of a colour patch. The scale and orientation of the interest regions are used to first re-orient the colour patches before these are recorded, and by doing so eliminating the need to systematically rotate the regions when comparing two patches. In order to compensate for the fact that the orientation obtained from the interest regions are not perfect, two neighbouring angles are also included, one to each side of the defined orientation from the interest regions. The colour patches are then resized to a standardised size of $p \times p$ pixels to allow for quick comparison of two regions from each of the reference and sensed images. The resizing process is achieved by first dilating the colour patches to fill gaps or missing pixels after background removal. The colour patches are then eroded to remove noises introduced from the image acquisition device [174]. The local descriptors and the colour patches are combined to form hybrid local descriptors, which are combination of both area- and feature-based methods. The disadvantages of the two methods, namely the incapability of feature-based methods to deal with regions lacking distinct features and area-based methods' high computation time are then minimised.

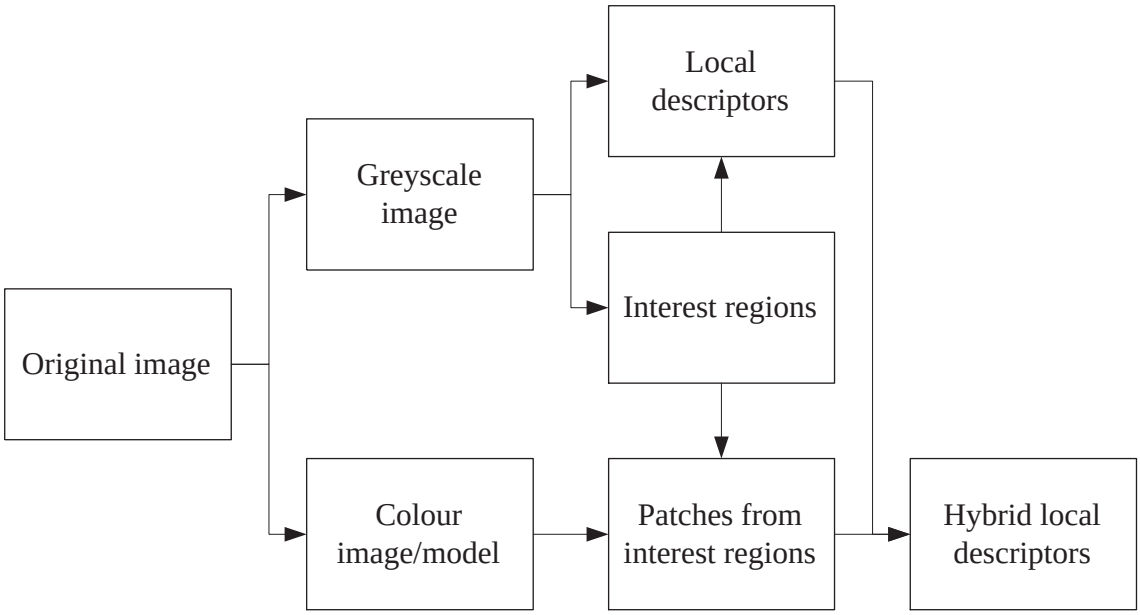


Figure 4.12: Overview of the hybrid local descriptor method. The interest regions and local descriptors are both computed from the greyscale image, while the colour patches are computed from the colour model of the original image. The local descriptors and colour patches are then combined to form the hybrid local descriptors.

4.4.2 Matching of Colour Patches

Once sets of hybrid local descriptors have been computed for both images in an image pair, the hybrid local descriptors are then matched in order to find correspondences. The local descriptors formed from greyscale images are matched using the threshold matching method defined in Equation 4.2. To match the colour patches, for each region in the sensed image, only some regions from the reference image are selected based on the results from the threshold matching method to reduce the computation time. This research took the best five matches based on the Euclidean distance measure, that is, the five regions in the reference image that have the smallest Euclidean distance values. Each of these five regions are then compared with the region from the sensed image using an area-based matching method known as normalised cross-correlation [187, 188]. The method is defined as:

$$NCC = \frac{\Sigma_{x,y}\left(\mathbf{r}^1(u+x,v+y) - \overline{\mathbf{r}^1}(u,v)\right)\widehat{\mathbf{r}^2}(x,y)}{\sqrt{\Sigma_{x,y}\left(\mathbf{r}^1(u+x,v+y) - \overline{\mathbf{r}^1}(u,v)\right)^2 \widehat{\mathbf{r}^2}(x,y)^2}} \tag{4.29}$$

Where $\mathbf{r}^1$ and $\mathbf{r}^2$ are the interest regions from the reference and sensed images of size $p^1 \times p^1$ and $p^2 \times p^2$, respectively. (x,y) and (u,v) are indices valid in $\{1,\ldots,p_2\}$ and $\{1,\ldots,p_1 - p_2\}$. $\overline{\mathbf{r^1}}(u,v)$ is the mean of the region from the reference image at location (u,v) and $\widehat{\mathbf{r}^2} = \mathbf{r^2}(x,y) - \overline{\mathbf{r^2}}$ is the mean subtracted region from the sensed image. Note that $p^1 = p^2$ since the interest regions from both images are resized to $p \times p$ regions before comparisons are

made. There is however a problem in the standard normalised cross-correlation approach. As normalised cross-correlation was designed to work with greyscale images, it is not capable of dealing with colour images. To overcome this problem, Sangwine and Ell [189] proposed a modified normalised cross-correlation method for colour images based on quaternions [190]. Quaternions contains four components, one real and three imaginary:

$$q = a + ib + jc + kd \tag{4.30}$$

Where a, b, c and d are real numbers, while i, j and k are complex operators. A quaternion is considered a pure quaternion if the real component is zero. A *RGB* image can be represented as a pure quaternion as:

$$q_I = iR + jG + kB \tag{4.31}$$

The normalised cross-correlation in Equation 4.29 is then applied to this quaternion representation of the image. Once a set of normalised cross-correlation values are obtained for all the colour patch pairs, the correctness of matches of hybrid local descriptor pairs can be defined using a weighted scheme:

$$w_{L^2} \times L^2 + w_{\text{MNCC}} \times \text{CC} \leq \tau_{\text{HLD}} \tag{4.32}$$

Where w_{L^2} and w_{MNCC} are the weights assigned to the Euclidean distance measure from the local descriptor matches and the modified normalised cross-correlation value from the colour patches, respectively. If the smallest value of the sum of these two values between the five regions for each region in the reference image is below a defined threshold, τ_{HLD}, then it is considered to be the correct match.

4.5 Experimental Design

To demonstrate the robustness of the developed colour local descriptor and hybrid local descriptor methods, extensive experiments were conducted. For each of the experiments, the effects of the the following image transformations were studied: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint changes. To study the robustness of the method against illumination changes discussed in Section 3.2, two sets of experiments were also conducted to examine changes in: (a) illumination colour; and (b) illumination intensity.

4.5.1 Uniqueness

To determine whether the local descriptors computed from colour images are indeed more unique or not, an uniqueness test was carried out. The uniqueness of local descriptors

is defined by their eigenvalues [13] and for each local descriptor method compared in this chapter, their eigenvalues were computed and the sum of these values compared. As the hybrid local descriptor method contains both area- and feature-based components, its uniqueness was not able to be determined using this measure. However, as stated in [13], the robustness of methods in registering images is equally, if not more important, as uniqueness only gives an indication of how well a method might perform. To this end, the uniqueness of hybrid local descriptors was not computed and instead, only its robustness was analysed from the various experimental work carried out.

4.5.2 Local Descriptors Based on Greyscale versus Colour Images

The first set of experiments compared the performance of the SURF local descriptor method and the colour local descriptor method. The same image transformations used in Chapter 3 were used, however, the range of the image transformations are different from those presented in Chapter 3 for practical reasons. The focus on the experiments conducted in this, as well as the chapters to follow, was on the practicality of the methods for registering images for the purpose of 3D reconstruction and as such, the range of image transformations were re-considered. For example, as it has been shown in [3], the largest transformation involved was approximately 22.5°. Also, it is unlikely there would be a large change in the scale of the object from image to image, and has therefore been reduced from $[1,3]$ to $[1.1,1.5]$. This reduction in the range of image transformations studied allowed for a more in-depth study of the changes in the performance of the methods in the more realistic ranges, as the intervals in the ranges were reduced.

4.5.3 Feature-Reduced Colour Local Descriptors

The same set of experiments as those carried out in Section 4.5.2 was conducted for the two feature-reduction methods described in Section 4.3.2, as it was of interest to make a direct comparison between the conventional methods and the colour local descriptor method and its two derivatives using the feature-reduction methods. Because of this, the number of features of the colour local descriptors after they have been reduced in dimensionality was designed to be the same as the SURF local descriptor method. This was also the case for the PCA-reduced colour local descriptors, where instead of using the scree graph and Kaiser's criterion, the number of vectors was selected to allow for a direct comparison where the number of features are the same.

4.5.4 Illumination Changes

In addition to enhancing the performance of local descriptor methods by integrating local descriptors with colour images or colour models, the main reason for utilising the $m_1m_2m_3$ colour model was to utilise its ability to handle changes in illumination colour and intensity.

The ability of the $m_1m_2m_3$ colour model to deal with changes in illumination colour and intensity has already been proven in [180], however the colour model's ability to handle these changes in conditions when integrated with local descriptors was unknown, therefore experimental work were carried out to study the effects of these changes. For both changes in illumination colour and intensity, a video projector was utilised to provide the necessary lighting. This was used as it was possible to produce lighting of different colours and intensities using the video projector, and was a much more efficient method of controlling the light projected onto the object. Using methods like different coloured light bulbs was not only difficult to control, but also limited the range of illumination colour available. The video projector was also able to provide lighting of different intensity which was significantly more efficient compared to the use of filters to reduce the intensity of the light. Five different illumination colours, as well as a wide range of illumination intensities were studied.

4.5.5 Hybrid Local Descriptors

The experiments conducted for the hybrid local descriptors utilised the same set of experiments conducted in Section 4.5.2. As a comparison of performance, the results for the colour local descriptor, hybrid local descriptor and conventional local descriptor methods are presented together for ease of discussion and comparison in Section 4.6.

4.5.6 SURF versus SIFT

The local descriptor method of choice that formed the basis of the methods discussed in this thesis is the SURF local descriptor method. This is in contrast to the majority of existing studies, which were nearly always based on the SIFT local descriptor method. The reason for choosing SURF over SIFT was based on the results obtained in Chapter 3. In the experiments discussed, it was found that the SURF local descriptor method out-performed the SIFT local descriptor method when registering images of the artefacts studied. Based on the results obtained, it was determined that the SURF local descriptor method was a more suitable approach. In order to justify this selection, a comparison of the performance of the two local descriptor methods was carried out and experiments were carried out using the same setup outlined. Colour local descriptors were constructed based on both the SURF and SIFT local descriptor methods. The methods for constructing these colour local descriptors were the same, except that where the SURF local descriptor is 64D, the SIFT local descriptor is

Table 4.2: Uniqueness of the four local descriptor methods compared in this chapter.

	$\Sigma_{i=1:10}\lambda_i$	$\Sigma_i\lambda_i$
Colour local descriptors	1.8147×10^{13}	3.6324×10^{13}
CSIFT	1.9058×10^{12}	3.0975×10^{12}
SIFT	4.1270×10^{9}	7.2785×10^{9}
SURF	7.9134×10^{9}	1.5469×10^{10}

128D, meaning that the colour local descriptors computed using the SIFT method have a size of $128 \times 3 = 384$D. The difference in the size, or the number of features, of the colour local descriptors was ignored, as this difference had been ignored in all previous experiments in Chapter 3 as well as previous studies [84, 80] and despite this difference, the SURF local descriptor method still out-performed SIFT [84, 80].

4.6 Results and Discussion

4.6.1 Uniqueness

Table 4.2 shows the uniqueness of the four local descriptor methods compared, excluding the hybrid local descriptor method. These values were computed from local descriptors from randomly selected regions of images, and the same regions were used to compute local descriptors for the different methods to ensure fairness of the results. The second column in the table shows the sum of the first ten eigenvalues and the last shows the sum of all the eigenvalues computed from these local descriptors [13]. The values in the table illustrate the amount of variance described by each local descriptor and consequently, how unique these local descriptors are. As can be seen, the colour local descriptor method has the largest sum, followed by CSIFT, SURF and SIFT. This shows that the colour local descriptors are the most unique and should, in theory, perform the best. It should be noted however that this is simply a measure of variance and while it provides a good indication of how useful these local descriptor methods might be, the matching accuracy is equally if not more important in the ranking of the performance of these local descriptor methods [13].

4.6.2 Local Descriptors Based on Greyscale versus Colour Images

The results for the comparison of SURF and the new colour local descriptor method for image transformations of rotation, scale, tilt and viewpoint changes are shown in Figures 4.13 and 4.14. For viewpoint changes the results for all four artefacts are presented. For the rotation, scale and tilt changes, only results for the flute artefact are shown in this chapter, however similar trends were observed for the patu, wahaika and tiki artefacts. Because it was of interest to study the trend, since the methods were compared and the results are

relative to each other, and as the trend of the results from the objects are similar, they are not presented in this chapter. Instead, complete results for the experiments conducted, as well as the experiments in the sections to follow, can be found in Appendix B.

For all four image transformations, notable improvements were observed for the colour local descriptor method, with gains of up to approximately 10% in matching accuracy observed, depending on the type of transformation involved. The trend of the matching accuracy of the colour local descriptor method follows closely of that of the conventional approach, which demonstrates the improvement of the colour local descriptor method under all types of imaging conditions. An interesting attribute in the results is that the gain in improvement of the colour local descriptor method is reduced as the angle of transformation increases for rotation, tilt and viewpoint increased. The gain in matching accuracy reduced to approximately 5% in these cases. This is an indication of the limitation of image registration techniques under large magnitudes of image transformations. Due to the magnitudes of transformation involved which distorted the images, it was not possible for image registration techniques to accurately register these images despite the improvements gained in the discussed method. This trend was not observed in the results for scale changes as for scale changes, no distortion of the object in the images were present, only a change of the size of the object in the images.

4.6.3 Feature-Reduced Colour Local Descriptors

Principal Component Analysis

The results for reducing the dimensionality of colour local descriptors using the PCA and weighted channel methods are shown in Figures 4.13 and 4.14. From the seven result plots, it is clear that while the performance degraded when PCA was applied, better accuracies have been achieved over conventional local descriptors. The improvement of the PCA-reduced colour local descriptors varied depending on the type of image transformation involved, with an average of approximately 5% over the local descriptors computed from greyscale images.

Experiments were also conducted to analyse the effects of different sized colour local descriptors with the number of features reduced with PCA, and it was observed that generally, the more features the colour local descriptors contained, the better matching accuracies can be achieved. This is shown in Figure 4.15. The number of PCs used in Figures 4.13 and 4.14 was 64, and is the same as the number of features found in SURF local descriptors. The Kaiser's criterion and scree graph suggested to 34, and based on this, two additional values, 30 and 68, were included for comparison. 30 was chosen as it is lower than 34, and was used to demonstrate the result when less features than suggested by these two methods was used. 68 was chosen as it was slightly higher than the number of features which exist in SURF local descriptors.

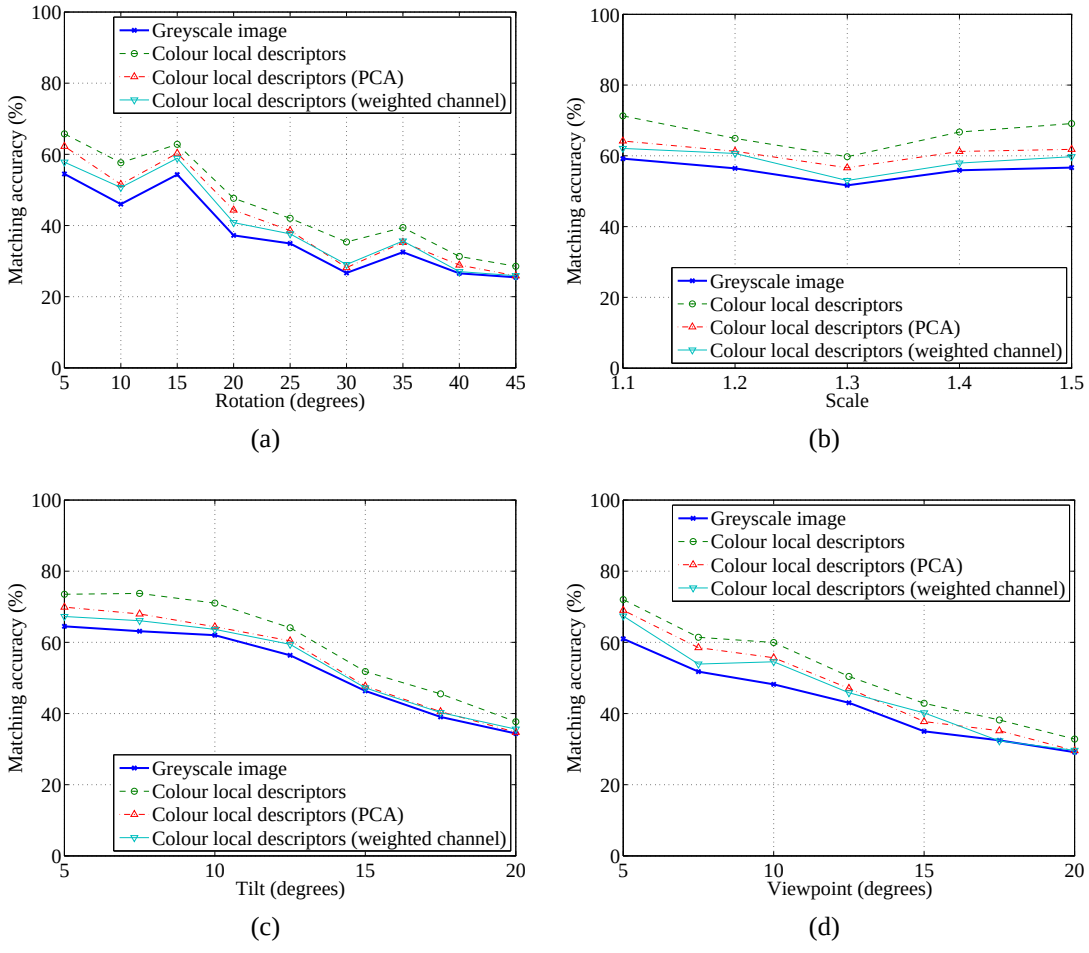


Figure 4.13: Image matching results using four different local descriptor methods for the flute artefact with different image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint.

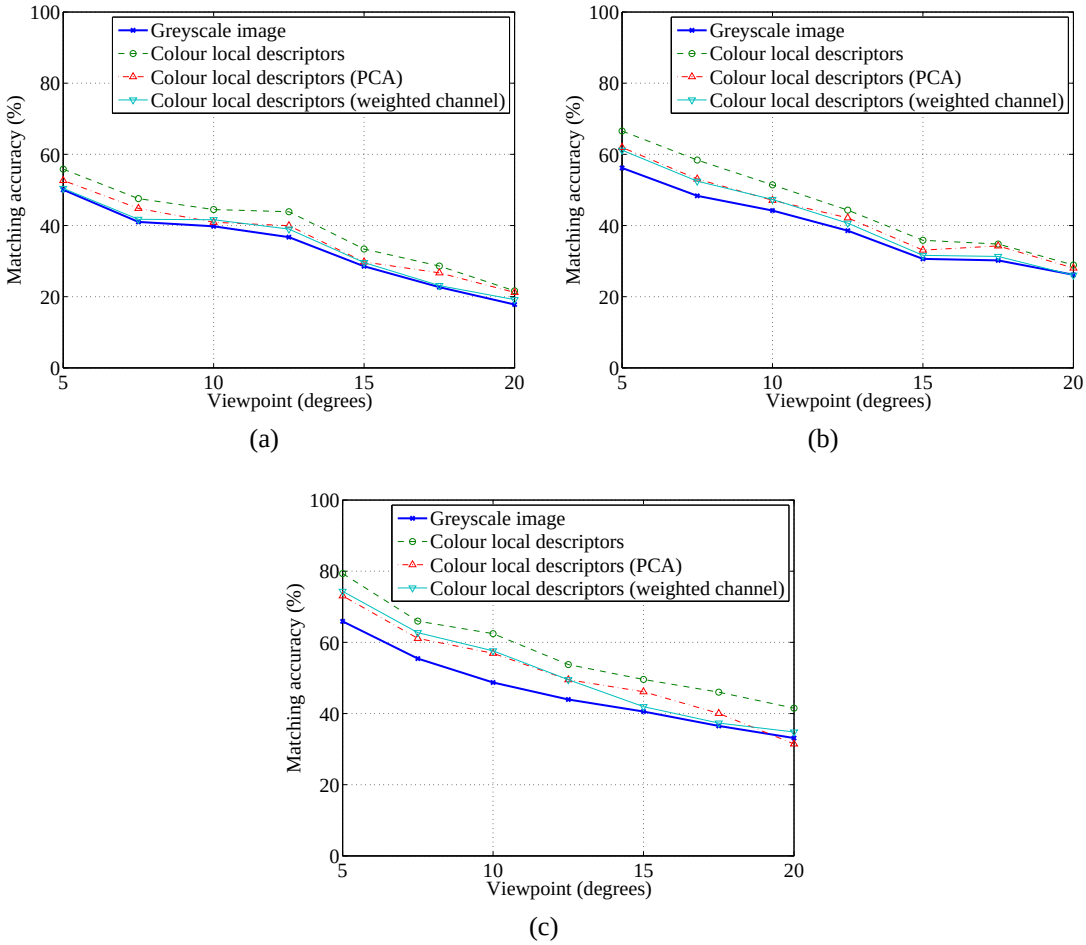
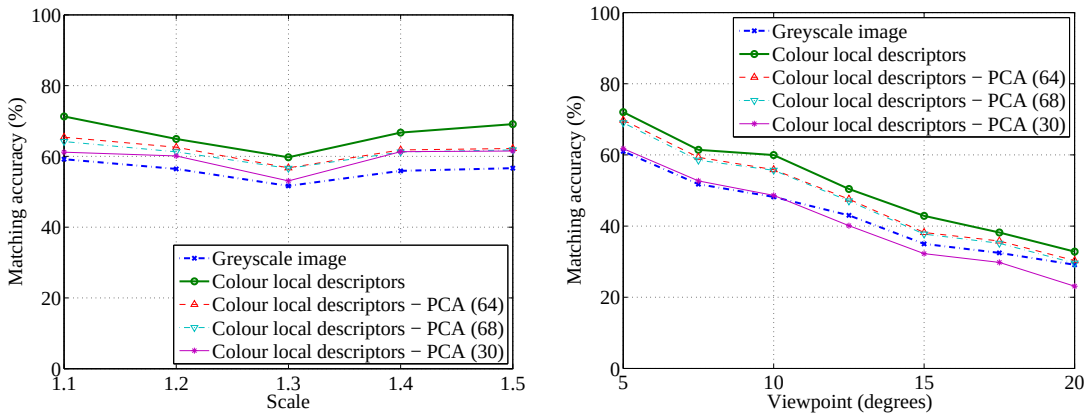


Figure 4.14: Image matching results using four different local descriptor methods for the patu, wahaika and tiki artefacts with viewpoint transformations: (a) patu; (b) wahaika; and (c) tiki.



(a) (b)

Figure 4.15: Image matching results using five different matching methods for the flute artefact with different image transformations: (a) scale; and (b) viewpoint.

As can be seen from the figure, using more PCs than recommended has limited benefits, where the improvement in matching accuracy is approximately $1-2\%$ between 64 and 68. When fewer PCs were used, the matching accuracy reduced by approximately $5-10\%$ compared to when 64 was used. This means that the performance degradded rapidly once the number of features was reduced below the number suggested by the scree graph and Kaiser's criterion. This was of no surprise, as these two methods measure how much information the vectors contain, and by not including the useful vectors as suggested by the two methods, critical information from the local descriptors were discarded.

As it has been discussed, in order to apply PCA, the covariance matrix is required. PCA has been applied to the methods in [97, 91] in a similar manner and is sufficiently adequate for the said techniques, however a problem arose for the reduction of the number of features in the colour local descriptor method developed. Due to the different colour make-ups of different types of objects, for objects of significantly different colour make-ups, the PCA approach would not work well as the covariance matrix computed for one type of object would have a different emphasise on which vectors of local descriptors are more useful in the registration of images.

An example to illustrate the potential problem of using PCA is as follows. Consider the *RGB* colour space instead of the $m_1m_2m_3$ colour model used in this research to simplify the example. If the covariance matrix computed from an image where the surface is made up primarily of different shades of red, then the covariance matrix would put strong emphasise on the local descriptors computed from the red component of the image, while placing little or no emphasise on the local descriptors computed from the green and blue components. If the covariance matrix obtained from this training image is applied to an image pair of an object where the surface is made up primarily of different shades of green, then the PCA-reduced local descriptors would have little or no distinguishing ability since the emphasis

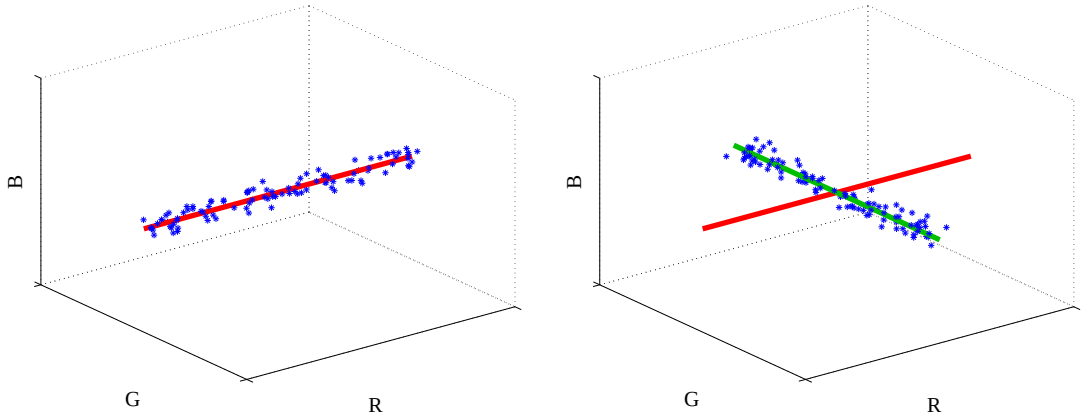


Figure 4.16: Example of the importance of choosing the right images for training and estimating the covariance matrix when using the PCA method for reducing the number of features of colour local descriptors: (a) covariance matrix estimated using an image primarily made up of different shades of red; (b) the covariance matrix computed from the (a), shown by the red line, does not suit the image since this image is primarily made up of different shades of green, and the true covariance matrix, shown by the green line, is significantly different.

is primarily placed on the red component of the image. In the case that the surface of the object is constructed primarily of different shades of green, there would be little change in the red colour channel of the *RGB* colour space and therefore the local descriptors computed based on the red colour channel would be very similar to each other, since they would all appear to be near zero. This example is shown graphically in Figure 4.16, with the three axes in each figure representing the corresponding *R*, *G* and *B* values of the pixels in an image. Figure 4.16a shows the case when an image is made up of different shades of red, therefore there is little variance in the green and blue components, and as a result, PCA is able to represent this data cloud using a single axis along the red axis. If the covariance matrix computed from this data is applied to Figure 4.16b, because the image is made up of different shades of green, the covariance matrix is of little use and in this case, the correct covariance matrix which describes the data is best represented by the green line instead of the red line in Figure 4.16b.

This is not an issue for the objects studied in this thesis, as the colour make-up of many artefacts are similar, since the majority of the Māori artefacts are carved from wood and therefore have a similar surface colour, and the problem discussed does not degrade the performance of the local descriptors enough to render them useless. This may however pose an issue for other applications, where the colour make-up of objects can vary greatly from one to another. This factor needs to be taken into account when determining the suitability of the PCA feature-reduction method for colour local descriptors for other applications.

Weighted Channels

The results for reducing the number of features of colour local descriptors are shown in Figures 4.13 and 4.14. It can be see from the figures that the performance of the local descriptors after the number of vectors has been reduced by the weighted channels method is poorer compared to the PCA-reduced colour local descriptors, where differences of $3-4\%$ were observed for different image transformations. The only exception is shown in Figure 4.14a, where the performance is only marginally better than the conventional approach in most cases for all experiments conducted. Again the size of the colour local descriptors had an impact on the performance and similar to the results from the PCA-reduced local descriptors, by having a higher number of vectors in the local descriptors, the performance was improved.

Despite the slightly lower performance, an advantage of this approach over the PCA method is that it does not require *a priori* knowledge of the colour make-up of the object. This is untrue in the case of the PCA method depending on the application involved, as the covariance matrix is required. Although it can be argued that in the application concerned in this thesis, due to the similarity of the raw material used for many of the objects, the need for this requirement is reduced, it is certainly a factor for other applications. By eliminating the need of this *a priori* knowledge, it is more versatile and can be used for a wide variety of images, as the selection process for which vectors to be included in the local descriptors is achieved by analysing the images concerned, instead of basing the dimension reduction process on a set of similar images compared to those which need to be registered. This property makes this approach more suitable for applications where colour changes are more significant, however, in the case of Māori artefacts, the PCA method is more useful as it achieved a better matching accuracy as shown in Figures 4.13 and 4.14.

It should be noted that while the two dimension reduction methods both performed favourably compared to the conventional method, the aim of this research was to robustly register images for the purpose of 3D reconstruction. As real-time performance is not expected, nor required for many real-life image registration applications, the best approach is to retain all the vectors of the colour local descriptor method to achieve the best matching accuracy possible.

4.6.4 Illumination Changes

The performance of the colour local descriptor and hybrid local descriptor methods in the presence of illumination condition changes are shown in Figure 4.17. Five different illumination colours and five intensity levels were studied. For the illumination colours, the reference was white light, and the other five colours were: (1) red; (2) green; (3) blue; (4) random colour 1; and (5) random colour 2. The two random colours were

chosen randomly by selecting random values for (R,G,B). This was done to evaluate the performance of the two developed methods under unknown illumination colours. In addition to the two developed methods, the SURF and CSIFT local descriptor methods were also used in the experiments which were used as baselines for comparing the performance of these methods.

The results from the experiments in Figure 4.17a show that the colour local descriptor method is the most robust against changes in illumination colour with approximately $10-15\%$ improvement in matching accuracy. This is followed by the hybrid local descriptor method, CSIFT and finally SURF. The figure also shows that the reduction in matching accuracy is the lowest for the colour local descriptor method, and CSIFT and SURF both have similar reduction in matching accuracy. While the hybrid local descriptor method is placed second, the matching accuracy reduced significantly for this method, and in comparison to the other three methods, the reduction in matching accuracy is around $20-25\%$. This reduction is significantly more than the colour local descriptor method's reduction of $10-15\%$ and CSIFT and SURF's reduction of $15-20\%$. The cause for this was identified as the hybrid nature of the method. Unlike the other three methods, the hybrid local descriptor method combines both area- and feature-based methods, and for the colour patches computed, even though a colour model was utilised and the cross-correlation of the colour patches were normalised, these colour patches are based on area-based image registration approaches and are more susceptable to changes in illumination colours compared to local descriptors, which are feature-based methods.

For the experiments which studied the effects of changes in illumination intensities, the standard lighting condition used for all the other experiments was used as the reference. A reduction in illumination intensity is defined as the reduction in lux with respective to the reference lighting condition. From Figure 4.17b, a similar trend to Figure 4.17a can be observed. While the hybrid local descriptor method has the highest matching accuracy initially, this quickly deteriorated as the change in intensity increased, again due to the area-based component in the hybrid approach. Similar performances were observed for CSIFT and SURF, while the colour local descriptor method has the smallest negative gradient, suggesting that the colour local descriptor method is the most robust against illumination intensity changes. The colour local descriptor method is approximately 10% more accurate compared to existing methods in all cases.

4.6.5 Hybrid Local Descriptors

The results for the hybrid local descriptor method are shown in Figure 4.18. The results for the colour local descriptor method as well as the SURF and CSIFT local descriptor methods are also shown in the figure to ease the comparison of these methods.

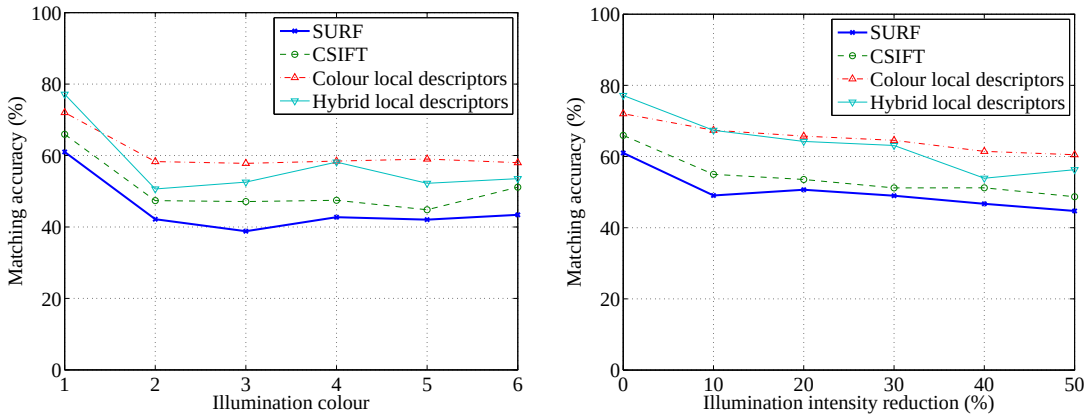


Figure 4.17: Image matching results using four different local descriptor methods for the flute artefact with different illumination changes: (a) colour; and (b) intensity.

As can be seen from the figure, the hybrid local descriptor method had mixed results compared to the SURF, CSIFT and the colour local descriptor methods. Examination of the results shows that the change in the colour patches used by the hybrid local descriptors, defined by the interest regions changed rapidly as the change in transformation increased. This is in contrast to the colour local descriptors which are more invariant to these changes. As a result of this, while the performance of the hybrid local descriptor method was superior to the colour local descriptor method in certain cases, with gains of up to $3-5\%$ in accuracy, the performance quickly deteriorated as the transformation increased. This was observed for the rotation, tilt and viewpoint changes. The accuracy was lower than the colour local descriptor method for angles below approximately $20-25^{\circ}$ for rotation changes, 10° for tilt changes and $10-15^{\circ}$ for viewpoint changes. The only image transformation that did not suffer from this is the scale changes. This is due to the colour patches being standardised to a $p \times p$ pixel region, and as a result the change in scale of the object in the images did not have a dramatic effect as it did in the other three image transformations studied.

From these results, it can be concluded that the hybrid local descriptor method is most suitable for small magnitudes of image transformations, which, based on the results, can be defined as up to approximately 10° in image transformations. In these conditions, it can out-perform the colour local descriptor method. However, in the case that larger image transformations exist, it is best to utilise the colour local descriptor method for its consistency and robustness throughout the different ranges in the four image transformations studied. It should also be noted that the computation time of this approach is relatively higher compared with the colour local descriptor method. This was due to the need to compare colour patches of regions using the modified normalised cross-correlation algorithm. This is again not a major disadvantage as real-time performance was not required. The characteristics of not being able to handle large image transformations and a relatively high computation time

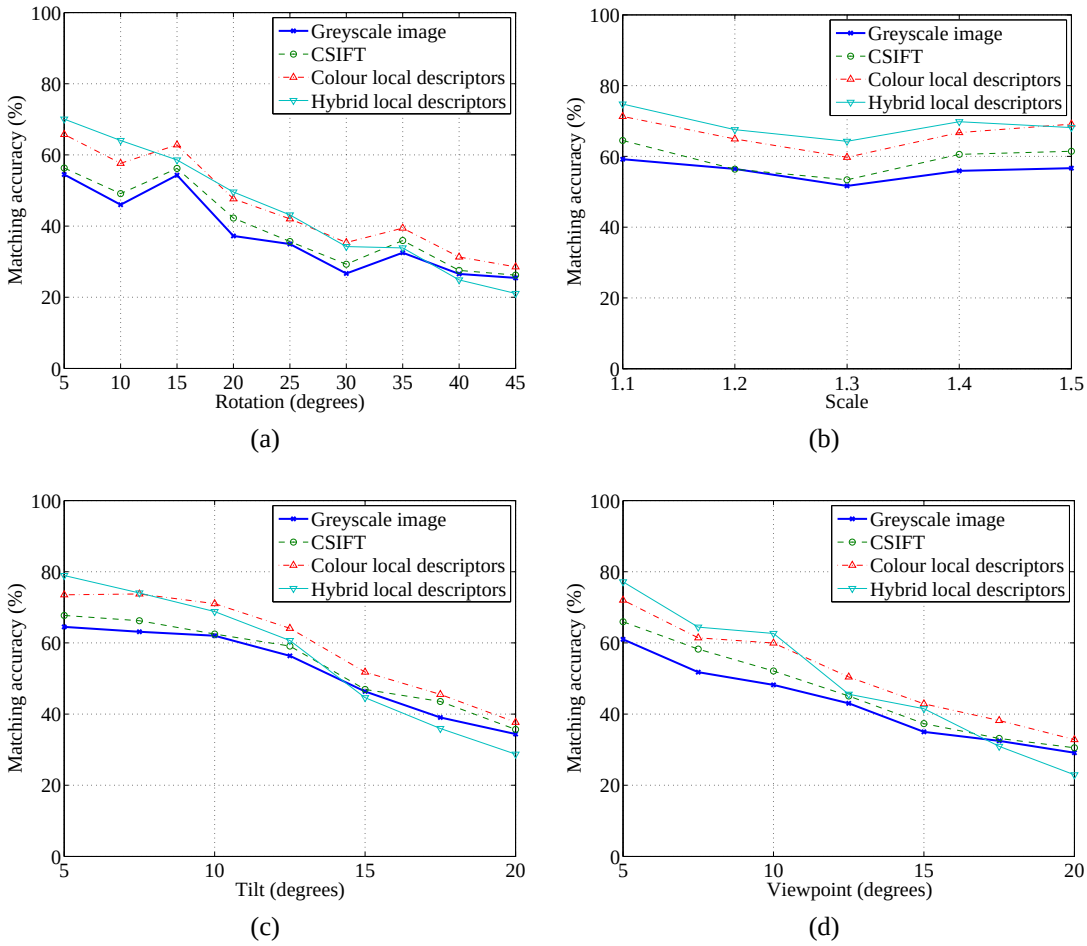


Figure 4.18: Image matching results using four different local descriptor methods for the flute artefact with different image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint.

is in line with the attributes of area-based image registration methods. This was expected considering the nature of the hybrid local descriptor method which combines feature-based with area-based methods.

4.6.6 SURF versus SIFT

The results for computing colour local descriptors using the SURF and SIFT local descriptor methods are shown in Figure 4.19. From the figures it can be seen that there are only minor differences in performance of the SURF and SIFT local descriptor methods. Overall, while small, SURF out-performed SIFT slightly by around $1-3\%$ under the different image transformations. This difference is not as significant as those observed in Chapter 3, and was therefore difficult to conclude that there was a clear advantage of one method over another. One important factor is that due to the advantage of SURF being a smaller local descriptor, where SURF is 64D versus 128D of SIFT in their original form, and 192D versus 384D for the colour local descriptor method, it was significantly faster in both forming the local

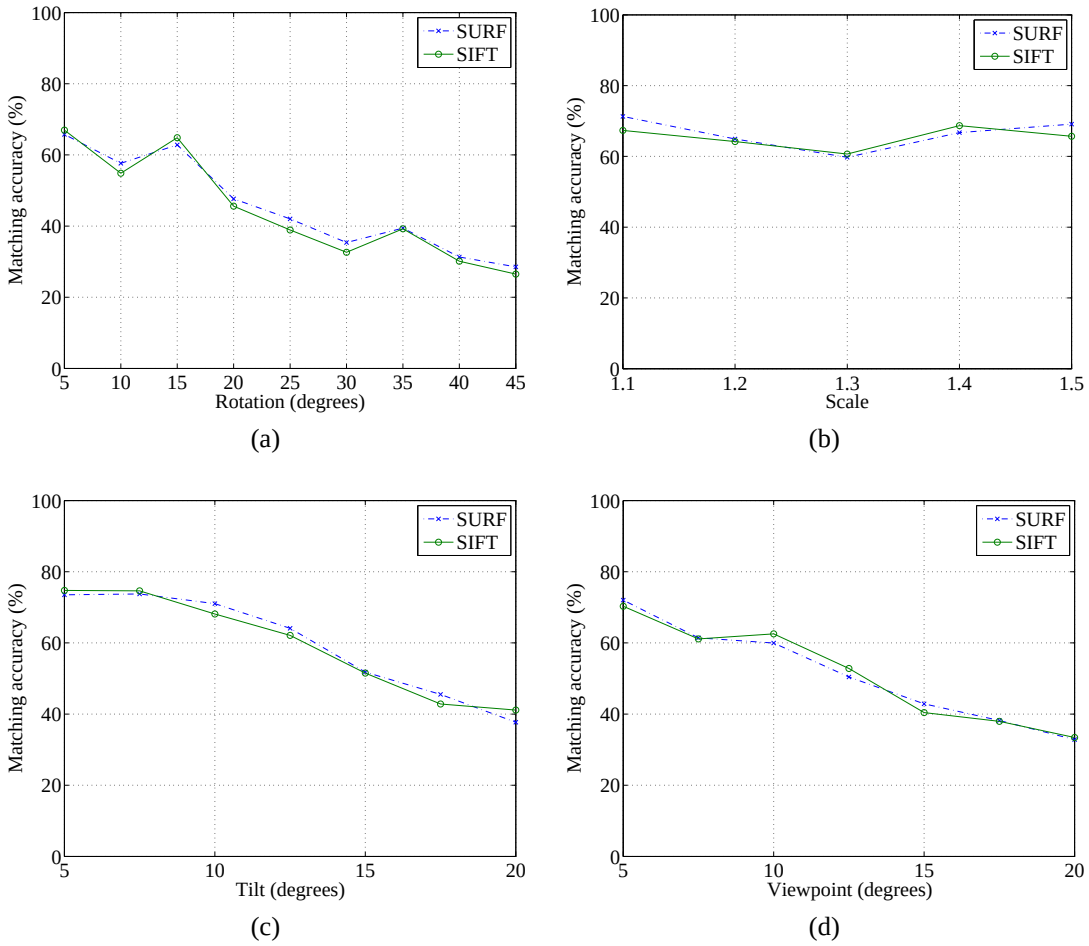


Figure 4.19: Image matching results using colour local descriptors computed from two different local descriptor methods for the flute artefact with different image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint.

descriptors and in the matching process. Based on these results, the choice of SURF was justified.

4.7 Conclusions

Two new local descriptor methods were presented in this chapter, namely the colour local descriptor and hybrid local descriptor methods. These methods have different approaches to improving existing local descriptor methods. Experiments were conducted to verify the performance of these methods, and based on the results, it was concluded that the colour local descriptor method performed better compared to existing methods for all the image transformations, as well as illumination condition changes, studied. The gain in matching accuracy was approximately 10% for this method. The results were also supported by the uniqueness test, where the colour local descriptors were the most unique. This means that it is easier to distinguish between regions which appear similar and therefore less likely for

mismatches to occur.

The hybrid local descriptor method had mixed results, where for small image transformations of up to approximately 10° for rotation, tilt and viewpoint changes, the method performed better than both existing, as well as the colour local descriptor methods. In these conditions, the matching accuracy was up to approximately 5% better than the colour local descriptor method. For larger transformations, the performance was not favourable and this was concluded to be due to the hybrid nature of the method.

Based on the results, it was concluded that the colour local descriptor method is the more versatile and robust of the two, and is capable of handling larger magnitudes of image transformations compared to the hybrid local descriptor method. As the colour local descriptor is the preferred method because of its robustness, two feature-reduction methods were developed to reduce the computation time required. Both the PCA and weighted channels methods reduced the computation time, while minimising the amount of accuracy lost. The PCA out-performed the weighted channels approach slightly by approximately $3 - 4\%$. The PCA feature-reduction method is approximately 5% less accurate compared to when PCA was not applied, however, still performed better than existing methods by approximately 5%.

These two methods are considered to be important contributions to local descriptor methods, as new approaches have been proposed which cover a variety of image transformations, including: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint changes, as well as changes in illumination colour and intensity. Based on the results from the experiments conducted, it can be concluded that significant gains in matching accuracy were observed in all the conditions listed. They are suitable not only for the artefacts used as case studies in this research, but can also be applied for other applications. This is due to the focus of development being placed on computing more unique local descriptors, and hence making them easier to match for any image the methods are applied to.

Chapter 5

Local Descriptor Matching with Support Vector Machines

Local descriptor matching is the most overlooked stage of the three stages of the local descriptor process, and this chapter proposes a new method for matching local descriptors based on support vector machines. Results from experiments show that the developed method is more robust for matching local descriptors for all image transformations considered. The method is able to be integrated with different local descriptor methods, and with different machine learning algorithms and this shows that the approach is sufficiently robust and versatile.

Local descriptor matching is the final stage of the three stages of the local descriptor process as shown in Figure 2.4, and is by far the most overlooked stage in current research in local descriptor processes. While much research had been on the robustness of local descriptor processes by improving existing or developing new region detectors and methods for computing local descriptors, the matching of these local descriptors often relied on methods based on metric distance measures. Using metric measures, the difference vectors of local descriptor pairs are reduced to scalar values, and these scalar values are then used for determining the correctness of matches. This method removes the majority of the useful information from the difference vectors, and is therefore undesirable, in particular when similar local descriptors exist, and it can easily lead to mismatches. With this understanding of the methods which are currently used for matching local descriptors and from the results acquired in the evaluation study in Chapter 3, there is a strong need for improvement to the matching of local descriptors.

This chapter describes a new method for matching local descriptors based on SVM. Instead of computing scalar values from difference vectors of local descriptor pairs, the developed local descriptor matching method utilises all the vectors of the difference vectors, and is capable of handling outliers in these difference vectors, which are previously ignored. The method aims to provide a more accurate and robust method for registering images. The

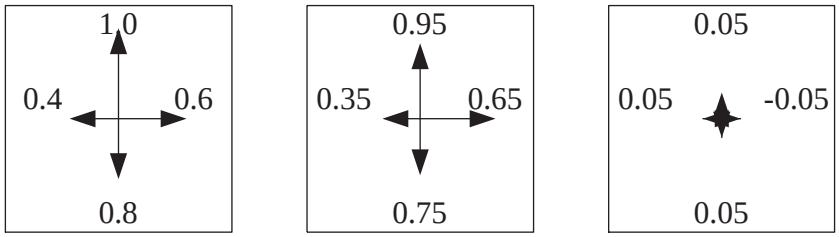


Figure 5.1: (a) Local descriptor from the reference image; (b) local descriptor from the sensed image; and (c) difference vector of the local descriptors in (a) and (b).

approach is robust and a wide variety of machine learning algorithms, not just SVM, can also be used with the developed approach.

Also, it should be pointed out that throughout this chapter, the word ‘vector’ has two different meanings, as shown in Figure 5.1. First, it refers to the individual vectors found in a local descriptor. For many local descriptor methods, in the construction stage of these local descriptors the individual values are computed and assigned a magnitude and orientation, therefore the word ‘vector’ is more suitable than simply referring to them as ‘values’. Graphically, this refers to the four arrows and their corresponding values in Figure 5.1a. The values, 1.0, 0.6, 0.8 and 0.4 are the magnitudes and the arrows are the orientations. The word ‘vector’ also refers to the difference vectors of local descriptor pairs and is similar to a row or column vector in that they are collections of values or in this case, vectors. This is shown in Figure 5.1c, and ‘vector’ in this case refers to the collection of the four arrows and their corresponding values. The values are computed from Figure 5.1a and Figure 5.1b and are the difference of the magnitudes.

This chapter is structured as follows. The currently used local descriptor matching methods based on metric distance measures and their limitations are discussed in Section 5.1. This provides an insight into the limitations of existing methods for the present research. An overview of SVM is presented in Section 5.2. The new local descriptor matching method is discussed in detail in Section 5.3. Experiments were conducted to verify the performance of the new local descriptor matching method using SVM and the experimental setup is presented in Section 5.4. The experimental results and detailed discussions on the results are presented in Section 5.5 and finally, the chapter is concluded in Section 5.6.

5.1 Limitations of Metric Distance Measures

The aim of the last stage of the local descriptor process, local descriptor matching, is to identify a set of corresponding local descriptor pairs given two sets of local descriptors, one set each from the reference and sensed images. Currently, the most commonly used method for matching local descriptors is the threshold matching method. This method compares

local descriptors by first computing the metric distance or L^p norm of the difference of local descriptor pairs:

$$L^p = \left(\Sigma_{i=1}^{n} |x_i^1 - x_i^2|^p\right)^{\frac{1}{p}} \tag{5.1}$$

The resulting distance measure is then thresholded and if this distance measure is equal to or smaller than a pre-defined threshold, then the pair is considered to be a correct match:

$$L^p \leq \tau_{\mathrm{T}} \tag{5.2}$$

Using this method, it is possible for each local descriptor in the sensed image to have multiple matches from the reference image if there are regions with similar features, and hence similar local descriptors, across the images [80]. The most popular distance measure used for the threshold method is the Euclidean distance, or the L^2 norm as defined in Equation 4.2, and has been used in various studies such as [191, 84, 13]. The other popular distance measure is the L^1 norm, also known as the taxicab geometry or rectilinear distance. The method is sometimes used as an alternative to the Euclidean distance in order to eliminate the additional computation requirement involved in computing the square and square root of values:

$$L^1 = \Sigma_{i=1}^{n} |x_i^1 - x_i^2| \tag{5.3}$$

In addition to the threshold method, two other methods for comparing local descriptors were introduced in [13]. The first is the Nearest Neighbour (NN) method, which is similar to the threshold method, except that the threshold method can have more than one match for each local descriptor from the sensed image. The NN method only considers the local descriptor pair, which has the smallest distance measure, in other words, the most similar, and if the distance measure is equal to or smaller than a pre-defined threshold, then the local descriptor pair is considered a correct match:

$$L^p \leq \tau_{\mathrm{NN}} \tag{5.4}$$

The advantage of this approach is that for each local descriptor in the sensed image, it is possible to have one match from the reference image, and does not need to deal with the multiple matches each local descriptor may have. The second method introduced in [13] is the Nearest-Neighbour Ratio (NNR) matching method. This method is similar to the NN method, except that the threshold is applied to the ratio of the two closest neighbours, in another words, the ratio of the closest and the second closest local descriptor pairs as defined by:

$$\frac{L^p_{\text{smallest}}}{L^p_{\text{second smallest}}} \leq \tau_{\text{NNR}} \tag{5.5}$$

The idea behind this approach is that in any given image pair, there should only be one correct match from the reference image for every local descriptor in the sensed image. All other local descriptors in the reference image are significantly different from the correctly matched local descriptor from the same image.

One issue with all the above-mentioned methods is that due to the large number of features of local descriptor methods, extensive search where the metric distance of all possible local descriptor pairs needs to be computed is required and often time-consuming. To rectify this problem, the best-bin-first algorithm introduced in [192] is an approximate algorithm based on the *k*-d tree search algorithm [193]. The algorithm was designed to avoid the need for the expensive search and hence improve the computation efficiency. Best-bin-first algorithm reduces the computation time by matching only a small fraction of the vectors from each local descriptor and by doing so, it is possible to gain speedups of up to 1000 times compared to when all the vectors are used for the comparison, 95% of the time.

One obvious and critical drawback of these methods for matching local descriptors is that they reduce the difference vectors of local descriptor pairs into scalar values before determining whether these local descriptor pairs are correctly matched. As such, it is impossible for these methods to compensate for factors such as outliers in the individual vectors or noise in local descriptors. As a result of this, it is possible to mismatch a local descriptor pair if, for example, one of the vectors from a local descriptor in the reference image, which is the correct match to a local descriptor from the sensed image had an abnormal value due to factors such as noise in the image. Figure 5.2 shows a simplified example of the case when the values of the individual vectors, and not the summed value of all the vectors in the difference vectors of local descriptor pairs, are used to better determine correct matches for local descriptor pairs. The L^1 norm for the difference vector of the local descriptor pair in Figures 5.2a and 5.2b is $|1.0-0.95|+|0.6-0.65|+|0.8-0.75|+|0.4-0.35|=0.2$. The L^1 norm for the difference vector for Figures 5.2a and 5.2c is $|1.0-1.0|+|0.6-0.6|+|0.8-0.6|+|0.4-0.4|=0.2$; and for Figures 5.2a and 5.2d is $|1.0-1.0|+|0.6-0.5|+|0.8-0.8|+|0.4-0.3|=0.2$. Both are the same as the first pair, resulting in the case where the matching method cannot determine the correct corresponding local descriptor for the local descriptor represented by Figure 5.2a. To overcome this issue, it is required to preserve the individual values of the difference vectors as they carry important information for determining the correctness of matches.

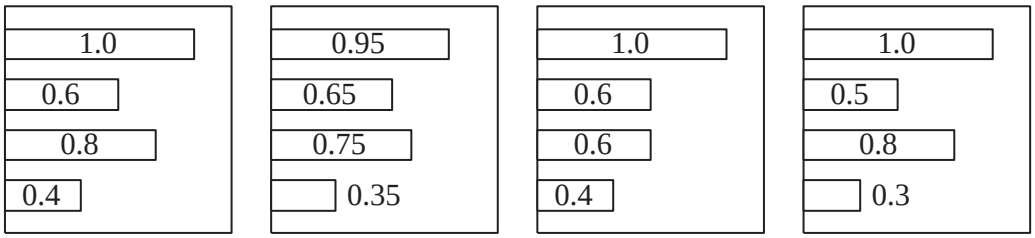


Figure 5.2: Simplified example local descriptors showing the difference of local descriptor pairs with the L^1 norm: (a) local descriptor from the sensed image; (b) correct corresponding local descriptor from the reference image; (c) and (d) possible but incorrect matches for the local descriptor in (a) from the reference image.

5.2 Overview of Support Vector Machines

To overcome the issue currently faced, an alternative approach which makes use of all the vectors in a local descriptor pair was required. Depending on the type of local descriptor method involved, the amount of data involved is potentially large. For example, the SURF local descriptor is 64D, the SIFT local descriptor is 128D and the colour local descriptor discussed in Chapter 4 is 192D. Due to this large number of vectors, machine learning methods were preferable since they can easily deal with large amounts of data. Also, because the algorithm is trained, this eliminates the problem with many manually designed systems when large amounts of data are involved. The designing of a system to handle large amounts of data in such cases is often difficult, if not impossible to achieve manually. SVM was chosen as it is one of the most widely used matching learning algorithm [194, 195] and because of its success with a large number of applications, including various image processing applications such as text recognition and face recognition [196, 197, 198, 199, 200, 201, 202]. It is a proven technique and if properly utilised, could greatly improve the robustness of the matching of local descriptors.

SVMs are a set of supervised machine learning methods used for the purpose of classification and regression in many different fields of research, in particular data-mining in statistics and pattern recognition [203, 204, 205, 195]. Because of the increase in the amount of data produced in various fields of research and applications over the years, computers are increasingly relied upon for the processing of these data for their ability to easily handle large amounts of data. Machine learning methods allow computers to easily deal with large amounts of data without the need to design complex rules or algorithms, instead the decision making criteria are learnt or trained from a set of training data, which are data with known input and output values [194, 206]. This process reduces the time and effort required to develop a system which is not only customised for the matching of local descriptors, but also very flexible. Because the system is trained and not manually designed, it can be easily changed when there is a change in the requirements of the system, for example when the

number of variables in the input is changed.

There are two types of problems in SVM, namely classification and regression problems. In classification problems, the aim is to determine the class to which the data belongs to, given an input data vector. The training data consists of values from all the available classes, or outputs. This includes the input vectors, $\mathbf{x}$ and their corresponding output values, $\mathbf{y}$. The simplest classification problem involves two classes or labels, $(+1,-1)$, and the output value $y_i \in \{1,-1\}$ represents the input vector $\mathbf{x}_i$ belonging to either class. This can be extended to a n-class problem, where each of the input vector belongs to one of the n available classes, $y_i \in \{1,\cdots,n\}$, depending on the values of the input vector. In regression problems, the aim is to try and find the value of the output y_i, given an input vector $\mathbf{x}_i$. The process for regression is similar to classification. It requires a set of training data, which identifies the underlying relationship between the input vectors and output values, $\mathbf{x} \in \Re^n$ and $y \in \Re$ [194]. For the matching of local descriptors, the aim is to determine, given the difference vector of a local descriptor pair, whether the local descriptor pair is correctly matched or not. In this research, the difference vectors of local descriptor pairs can be considered the input data vectors, and the correctness of matches the output values, and by matching the requirement of the task with the description of the two types of SVM problems, it was concluded that the matching of local descriptors is a typical classification problem.

5.3 Local Descriptor Matching with Support Vector Machines

In order to utilise SVM for local descriptor matching, a SVM algorithm called Iterative Single Data Algorithm (ISDA) [194] was implemented. ISDA is a fast and efficient algorithm, which solves quadratic programming-based problems in an iterative way. It belongs to the class of working set solvers. The process involved in developing a SVM model for matching local descriptors is shown in Figure 5.3. A set of training data, consisting of examples of both correctly and incorrectly matched local descriptor pairs is computed from a set of training images. These training data are then used to train the SVM model using a 10-fold cross-validation process, automatically tuning the parameters for the model. Once a set of parameters is determined, the SVM model is then computed and can be used to match local descriptors.

5.3.1 Training Data

The training data is a set of matched local descriptor pairs, where a combination of both correctly and incorrectly matched pairs exist. Providing the SVM with the correct training data is a critical step in ensuring that SVM can correctly match local descriptors. This

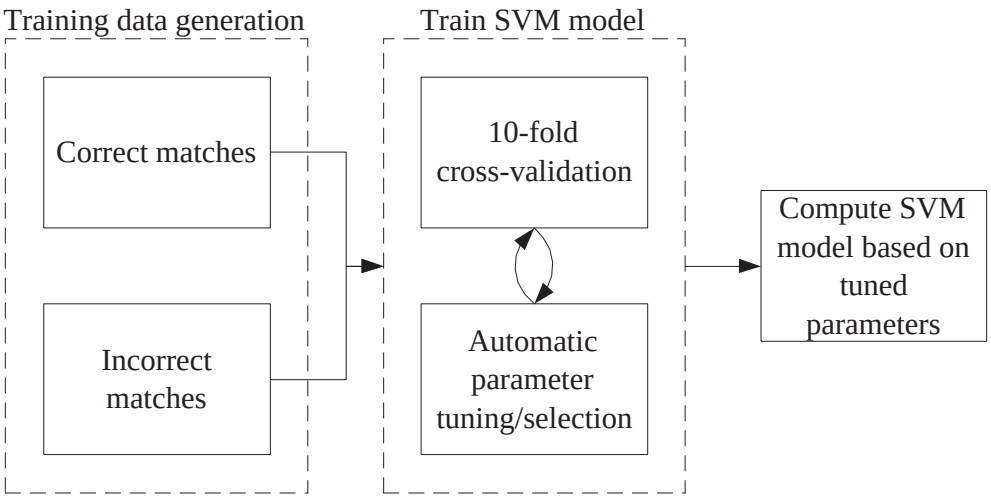


Figure 5.3: Overview of the training data used for the computation of a SVM model for the matching of local descriptors.

is achieved by using a representative SVM model that is designed for local descriptors computed from images of the objects studied. Both the correct and incorrect matches are of equal importance in the training phase. There should be sufficient examples of both types of local descriptor matches in order to generate a model that is capable of dealing with different types of both correctly and incorrectly matched local descriptor pairs. The training data consists of local descriptor pairs of images taken from different viewpoints. By considering the different image transformations involved, it ensures that the model trained is not bound to one type of image transformation and unbiased. This avoids the problem of needing to train the model multiple times when dealing with different image transformations, and the model is robust against any image transformation that may be encountered.

The procedure for generating the training data is as follows. For each image pair, in order to compute a set of both correctly and incorrectly matched local descriptor pairs, first the homography matrix of the image pair need to be computed to ensure the accuracy of the set of output values associated with the training data. Once the homography matrices for the image pairs have been computed, a set of correctly matched local descriptor pairs is then computed by projecting each local descriptor in the sensed image onto the reference image, and computing its correspondence, if one exists. Once the correctly matched local descriptor pairs are computed, the incorrectly matched pairs can be easily identified since these are all the non-correctly matched local descriptor pairs from the image pair.

It should be noted that while it might seem beneficial to use a large amount of training data to ensure that all the possible combinations are covered, this is often not practical, as the large amount of training data tends to reduce the performance of the SVM in both the training and classification phase. In addition, when using a large number of training data, the SVM model attempts to accommodate for every single training data and thus becomes unstable and

unreliable. An example of the size of the training data that can be obtained from an image pair is as follows. For an image pair with 500 local descriptors in each image, the theoretical maximum number of correct matched local descriptor pairs is 500, although this is never the case in real-life applications, since this would imply that the areas in the images which contain local descriptors overlap each other completely. For the purpose of this example, the total number of possible local descriptor pair combinations is simply the number of local descriptors in the reference image multiplied by the number of local descriptors in the sensed image:

$$n(\mathrm{LD}_{\mathrm{total}}) = n(\mathrm{LD}^1) \times n(\mathrm{LD}^2) \tag{5.6}$$

Where $n(\mathrm{LD}^1)$ and $n(\mathrm{LD}^2)$ are the number of local descriptors in the reference and sensed images, respectively and $n(\mathrm{LD}_{\mathrm{total}})$ is the total number of possible local descriptor matches. Given the maximum number of correct matches and total number of local descriptor pairs, the number of incorrect matches then becomes:

$$n(\mathrm{LD}_{\mathrm{incorrect}}) = n(\mathrm{LD}_{\mathrm{total}}) - n(\mathrm{LD}_{\mathrm{correct}}) \tag{5.7}$$

In the example presented above, the number of incorrectly matched pairs is $500 \times 500 - 500 = 249,500$. Experimental work will show that only a limited number of these incorrect matches is required in order to generate an useful SVM model.

By using the method described above, an adequate model was generated in the experimental work conducted, however, the results from the preliminary study showed that the SVM model sometimes generated incorrect local descriptor matches similar to those found when a threshold matching method was used. In order to improve the robustness and performance of the SVM model, the matching results from a matched set of local descriptor pairs using the threshold method were analysed. By using all the mismatched local descriptor pairs of the threshold method as part of the training data for the SVM model, it was found that the mismatches existed in the threshold method can be eliminated, while the correctly matched pairs retained. The training data used for developing a SVM model now consists of the following: the correctly matched pairs are obtained by projecting all local descriptors in the sensed image onto the reference image. This contains all the correctly matched pairs from the threshold method as well as local descriptor pairs that are not identified by the threshold method; the incorrectly matched pairs are obtained from two sources: (a) a set of randomly selected incorrect matches from the method discussed above; and (b) all the mismatched local descriptor pairs from the threshold method. It was found through experimental study that a set of incorrectly matched local descriptor pairs approximately three to four times bigger than the set of correctly matched local descriptor pairs gave adequate results. More training data did not guarantee higher matching accuracy, and a smaller set of training data

$$\left|\begin{array}{ccccc} y_1 & x_{11} & x_{12} & \cdots & x_{1v} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2v} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_m & x_{m1} & x_{m2} & \cdots & x_{mv} \end{array}\right|$$

Figure 5.4: Data format for ISDA, consisting of both the output values and input vectors.

was insufficient in developing an useful SVM model due to the lack of local descriptor match examples available for training the model.

Figure 5.4 shows the format of the training data used for the ISDA algorithm utilised. The first column consists of the output vector, $\mathbf{y}$, where $y_i \in \{1,2\}$, with $1 =$ correct match and $2 =$ incorrect match. Each row in the remaining columns represents an input vector $\mathbf{x}_i$ for the output of the same row. Note that the labelling system used by the particular implementation of ISDA used in this research is different compared to other software and programmes in that the labels are assigned in increasing order, $\{1,\cdots,n\}$ for an n-class classification problem, and this is true even in the case that a two-class problem is concerned. The programme is different from many other SVM software, where the labels of outputs are $\{1,-1\}$ when a two-class problem is concerned. This explains the difference in notation in this section compared to Section 5.2. The output generated by the ISDA programme used in the case of a two-class problem is however $[-1,1]$, where -1 corresponds to a label of 2 or incorrect match. Since one row is created for each new training data, if there are m training data available and each training data has a difference vector size of v, the input data for the training process is then a $(m, v+1)$ matrix.

A MATLAB programme was developed to allow quick generation of the required training data, and parameters can be defined such as how many incorrectly matched local descriptor pairs are required as a function of the total number of correctly matched local descriptors, and the labels assigned to these two types of local descriptor pairs. A screen shot of the developed programme is shown in Figure 5.5.

5.3.2 Training Support Vector Machines

To train a SVM model, two types of data are required which include a set of training data as discussed in the previous section, and a set of validation data with known labels, or correct or incorrect matches. The validation data is often a subset of the training data, and is used to validate the SVM model generated by the training data by computing the accuracy of the model. This is achieved by fitting the validation data to the model. The accuracy of the model is computed by comparing the known correct labels of the validation data with the labels generated by the SVM model. In order to utilise all the local descriptor pairs available for training the SVM model, a method known as cross-validation is utilised [207]. Cross-validation is a statistical method of partitioning the data into subsets in a way that a

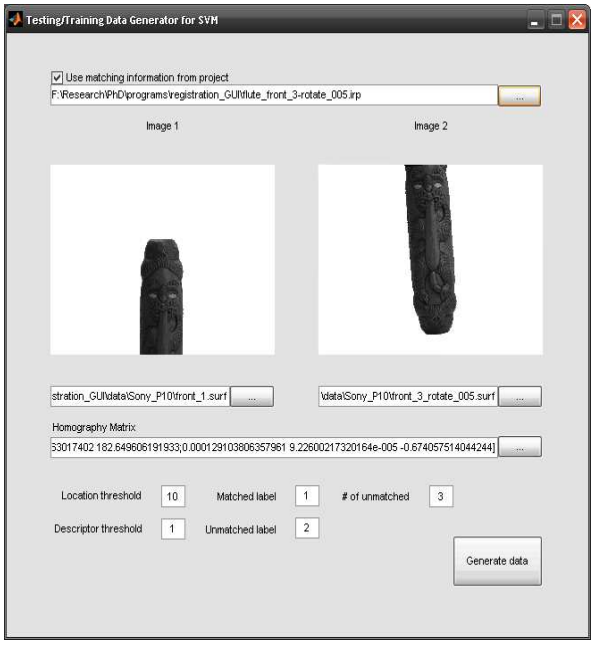


Figure 5.5: MATLAB programme interface for generating the required training data.

large proportion of the training data is used for training, and the remained data is utilised to validate the model. Specifically, the *k*-fold cross-validation method is applied, where the training data is divided into *k* subsets. The *k*-fold cross-validation process is repeated *k* times, each time using one subset of the data as the validation data, while the remaining $k-1$ subsets are used for training. At the end of the cross-validation process, the results from the *k* iterations are combined to produce an overall result.

The advantage of the *k*-fold cross-validation method is that all the training data available can be used for both the training and validation phase. Each data set is used once, eliminating bias in the results, caused when some data are not used or used multiple times. The disadvantage is that the training of the SVM model needs to be repeated *k* times and can be time-consuming. However, as the training of a SVM model for the matching of local descriptors is an offline process and would ideally only need to be performed once for the same type of objects or objects that share the same types of features. Hence, this was not considered a drawback in this research. The commonly used 10-fold cross-validation is used and this method is shown in Figure 5.6. First the data is structured as shown in Figure 5.4, the dataset is then shuffled or randomised before being used to train the SVM model using the method described above. The shuffling of these data is particularly important if the dataset is sorted according to the output label, as it is crucial to ensure that there are approximately the same number of correctly and incorrectly matched local descriptor pairs in each subset.

A special case of the *k*-fold cross-validation method is when the number of *k* is the same as the number of training data, *m*. In this case, the method is referred to as Leave-One-Out Cross-Validation (LOOCV). This method generates the most accurate model in most

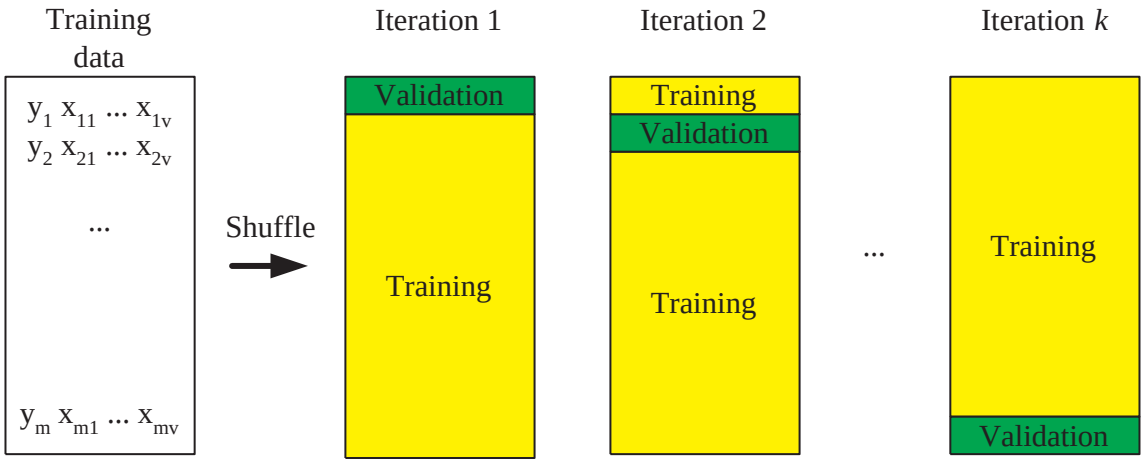


Figure 5.6: *k*-fold cross-validation method used for computing the SVM model for matching local descriptors.

circumstances [194], since every training data is used to validate the SVM model computed from all the other training data, and can provide an accurate understanding of the accuracy of the SVM model. The method is, however, very time-consuming due to the number of SVM models that need to be trained and is therefore very seldomly used for SVM. It will be shown in Section 5.4.4 that the LOOCV method was applied to a different type of machine learning algorithm known as the *k*-nearest neighbour (*k*-NN) method due to the low computation time and the nature of the approach.

A MATLAB programme was developed for performing the cross-validation task required as shown in Figure 5.7. The programme is an interface to ISDA [194], and parameters for the ISDA implementation utilised in this research can be defined. The programme includes features such as whether the data set need to be shuffled, the scaling of the data [194], and the definition of parameters such as the number of *k* and the range of the two SVM parameters, *C* and σ. These two SVM parameters are discussed in the next section. The programme can be used to automatically carry out cross-validation for training the SVM model given these parameters and is suitable for experimenting with different parameter combinations, which is a tedious task by hand given the number of possible combinations of SVM parameters.

5.3.3 Parameter Selection

The kernel used for SVM in this research is a Gaussian kernel, which is the most widely used kernel, since a true random distribution follows a Gaussian distribution [208]. There are two parameters to be tuned for the Gaussian kernel, namely the penalty parameter *C* and the threshold σ. To identify the best parameters for the model, these two parameters are automatically tuned by the following process, which is shown in Figure 5.8. First a set of values for each parameter is defined. A SVM model is then generated and its accuracy tested

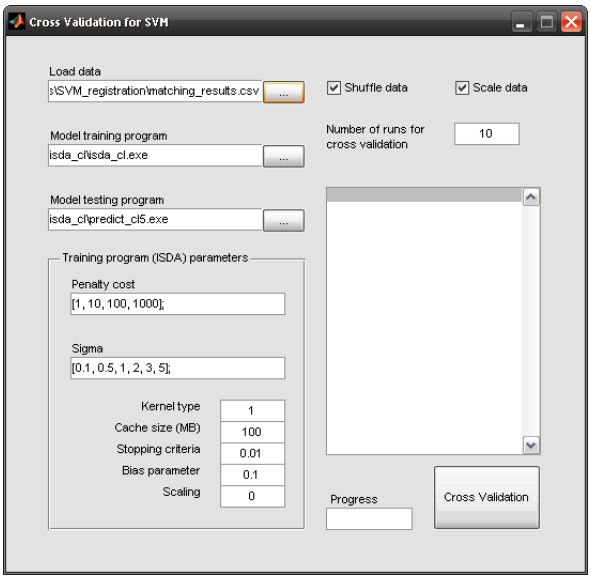


Figure 5.7: A screen shot of the developed MATLAB programme for cross-validation.

for each combination of parameters using the 10-fold cross-validation method. The total number of misclassifications is recorded for each parameter combination, and at the end the optimal solution is defined as the one that has the smallest misclassification rate, which is the number of misclassifications in the validation data over the total validation data size or the size of the training data:

$$\text{misclassification rate} = \frac{\text{number of misclassifications}}{\text{training data size}} \times 100 \tag{5.8}$$

For the automatic tuning and selection of the parameters, these values are initially set at large intervals [194] to ensure a good coverage of the possible combinations of parameters without increasing the computation time involved. After SVM models have been computed for the first set of parameters, the range of the parameters is then decreased while keeping the same number of values for each parameter, resulting in an approach similar to the quadtree method [209]. This process is repeated until the change in value of the parameters acquired is within a defined tolerance when the range is reduced. Once values for C and σ are selected, the final SVM model can then be generated using the automatically selected parameters, however without the use of cross-validation and instead, all the training data are used for training the SVM model.

In the case that a manual check on the misclassification rate is required, for example when the final SVM model computed is not sufficiently accurate, Figure 5.9 shows the plot which can be used for manually selecting the C and σ values resulted from a set of 10-fold cross-validation computations. The different values of C and σ are plotted along the x- and y-axes, while the misclassification rate is plotted along the z-axis. The C and σ combination with the smallest misclassification rate is the minima of the plot and can be easily and efficiently

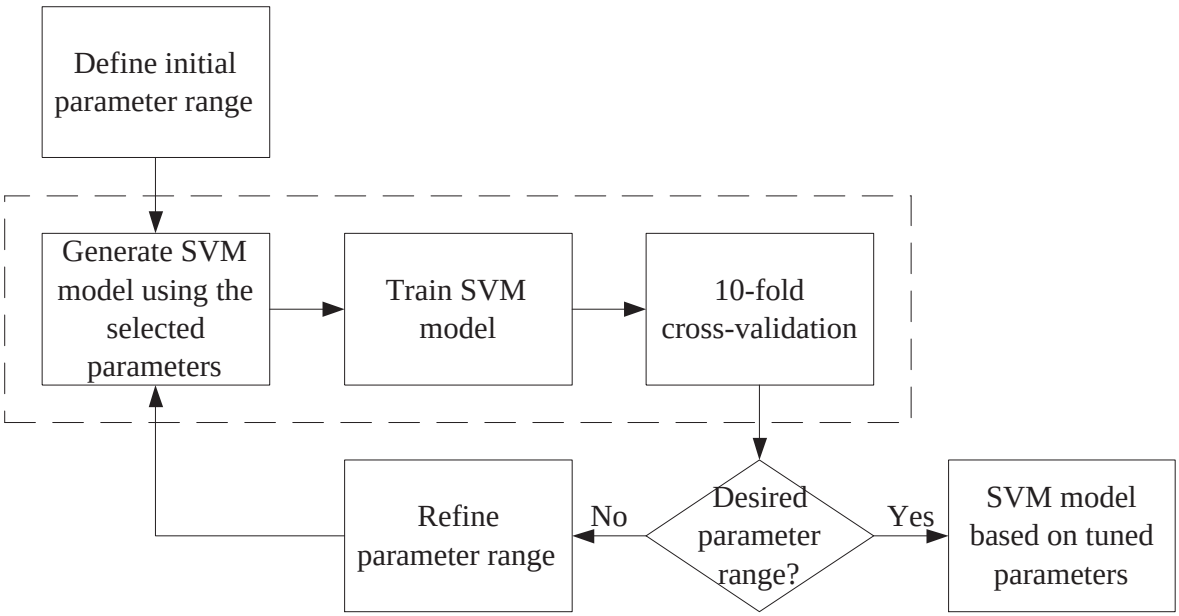


Figure 5.8: Training the SVM model and automatic tuning of the two parameters, C and σ.

identified from the figure shown. An advantage of this plot over identifying the minima from a matrix of values is that in the case where there are two or more minima, it is possible to efficiently identify these regions visually and further analysis can be performed by focusing on this set of minima from the plot.

To ensure fairness of the model design, the training data did not include any of the local descriptor pairs that needed to be registered for the experiments conducted in Section 5.4. This approach was taken as it is impossible to obtain the matching information of the local descriptor pairs for an image pair, as this implies that the homography matrix of the image pair is known, and there is no need to register the images.

The MATLAB programme developed for computing the SVM model for the matching of local descriptors is shown in Figure 5.10. This MATLAB programme was developed to simplify the process of computing the SVM model for matching local descriptors, and allows an easy and efficient way of modifying the associated parameters. The functionality of this programme is similar to that shown in Figure 5.7, due to the similarity in nature of training and generating an useful model given C and σ from the training phase.

5.3.4 Local Descriptor Matching

Given a SVM model, the aim is to check for the correctness of local descriptor pairs from an image pair. This is achieved by using every possible local descriptor pair in an image pair as inputs, and the SVM model is used to determine whether the given pair is correctly matched or not. The process for determining whether a pair of local descriptors is correctly matched or not is shown in Figure 5.11. The difference vector of a local descriptor pair is

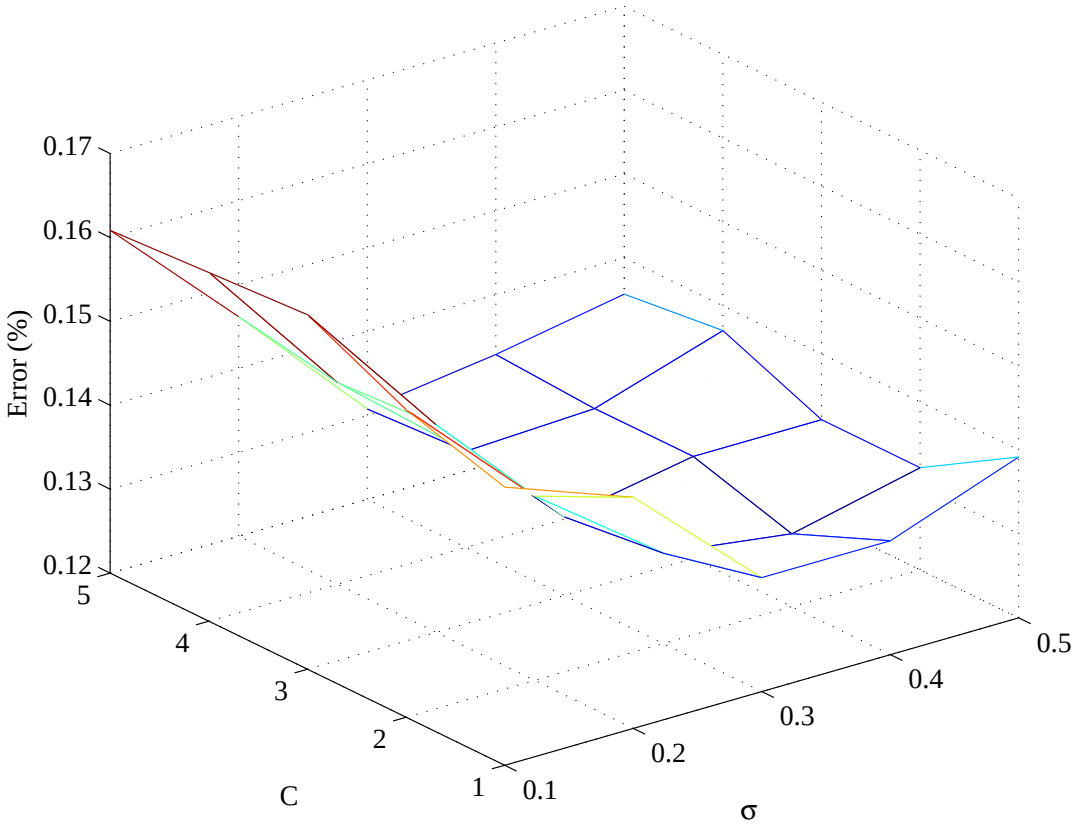


Figure 5.9: Parameter selection based on the classification error of the different combinations of C and σ for the SVM model.

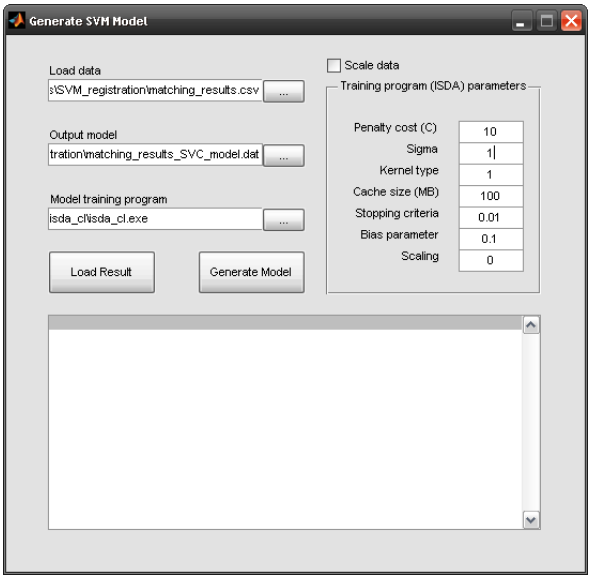


Figure 5.10: A screen shot for the developed MATLAB programme for SVM model generation.

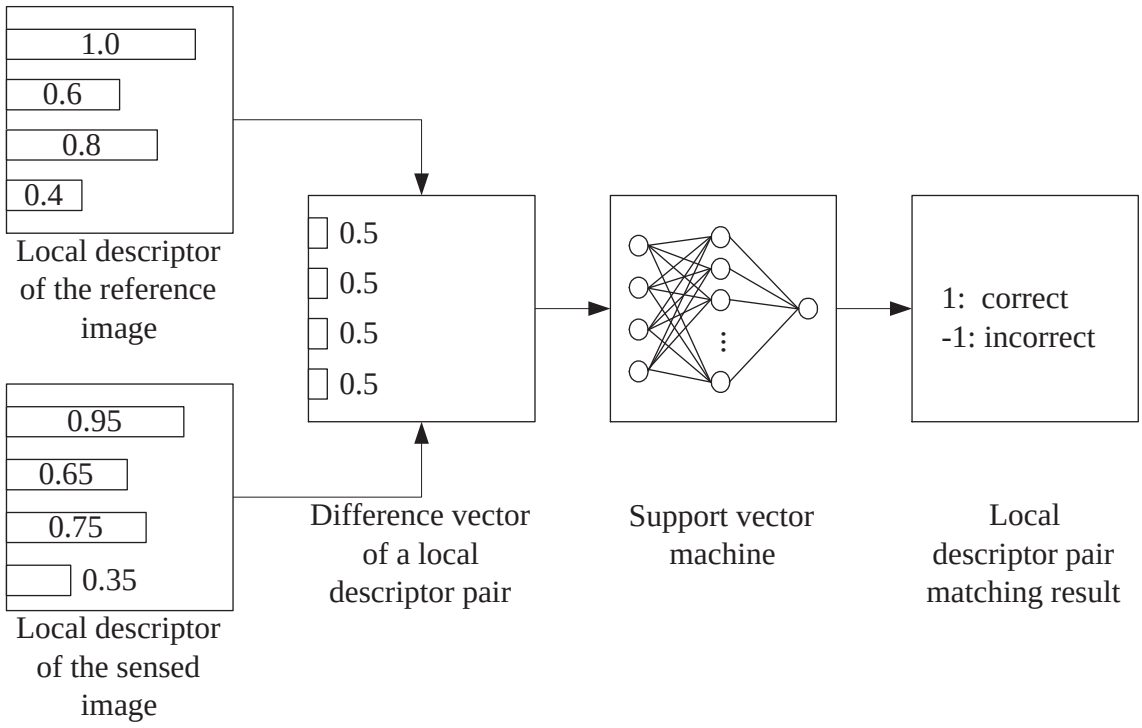


Figure 5.11: Overview of local descriptor matching using SVM.

first computed. This difference vector is then used as the input data vector into the SVM model, and the SVM model determines whether the match is correct or not by:

$$y(\mathbf{x}) = \Sigma_{i=1}^{v} \left(w_i G(\mathbf{x}, c_i)\right) + b \tag{5.9}$$

Where w_i is the weight of the i^{th} value in the difference vector, b the bias term and $G(\mathbf{x}, c_i)$ is the Gaussian kernel:

$$G(\mathbf{x}, c_i) = e^{-\frac{1}{2}\left((\mathbf{x}-c_i)^T \Sigma^{-1} (\mathbf{x}-c_i)\right)} \tag{5.10}$$

As the matching of local descriptors is a two-class problem, it is of interest to know whether the local descriptor pair is correctly matched or not. The outputs from the ISDA programme are real numbers in the range of $[-1, 1]$. If the output is closer to -1, then it is likely that the pair is incorrectly matched, and if, on the other hand, the output is closer to 1, it is likely that the pair is correctly matched. This research used the local descriptor pair with the highest output value, in other words, the pair with an output value closest to 1 is selected, and if the output from the SVM model of this local descriptor pair is equal to or above a threshold, then it is considered a correct match:

$$\max\left(y(\mathbf{x})\right) \geq \tau_{\text{SVM}} \tag{5.11}$$

Where $\tau_{\text{SVM}} \in [-1, 1]$ is the threshold placed on the output value of the SVM model.

5.3.5 Feature-Reduction

A problem with the SVM approach for matching local descriptors was the increased computation time of the method. While the computation time was not an important factor in this research since the matching of local descriptors is an offline process, it can be desirable when dealing with large amounts of data. For example, when there are a large number of local descriptors from image pairs or if many images need to be registered, as the time involved can potentially be reduced by reducing the computation time of the matching of local descriptors. In order to achieve this, the most obvious approach was to reduce the number of input data vectors into the SVM model, which in turn reduces the computation time for determining the output of the given input vector. In relation to local descriptor pairs, this means that the number of features of local descriptors should be reduced. However as it has been shown in previous studies [97, 13], reducing the number of features in local descriptors often lead to a decrease in the robustness of the method. In addition, the matching method developed in this chapter needed to be able to integrate with existing local descriptor methods and other future local descriptor methods. It was therefore desirable not to modify the local descriptors computed from these methods.

In this research, instead of reducing the number of features of local descriptors, the method reduces the number of features or vectors of the difference vectors of local descriptor pairs. By doing so, it avoids the need to modify existing local descriptor methods, as changes to these are not required. By reducing the number of vectors of the difference vectors, the useful features from the local descriptors can be retained, however, the vectors which carry little or no information about the images are discarded. It is therefore possible to reduce the number of inputs into the SVM model without affecting the robustness of the matching of these local descriptors significantly.

Two methods for reducing the number of features of the input vectors for SVM are developed. The first method makes use of PCA, which has already proven to be a powerful method for a similar task in Chapter 4. The second method is called Recursive Feature Elimination with Support Vector Machines (REF-SVMs) [210, 194]. Both methods were applied to reduce the size of the input vector for the SVM model, however they have different approaches for the task and were therefore also suitable for comparing the different types of approaches for the task. This difference in nature is described in detail in the following sections.

Principal Component Analysis

Instead of reducing the number of features of local descriptors as it was done in previous studies [97, 13], PCA is applied in a different manner and Figure 5.12 shows an overview of local descriptor matching using SVM. First, the difference vector of a given local descriptor

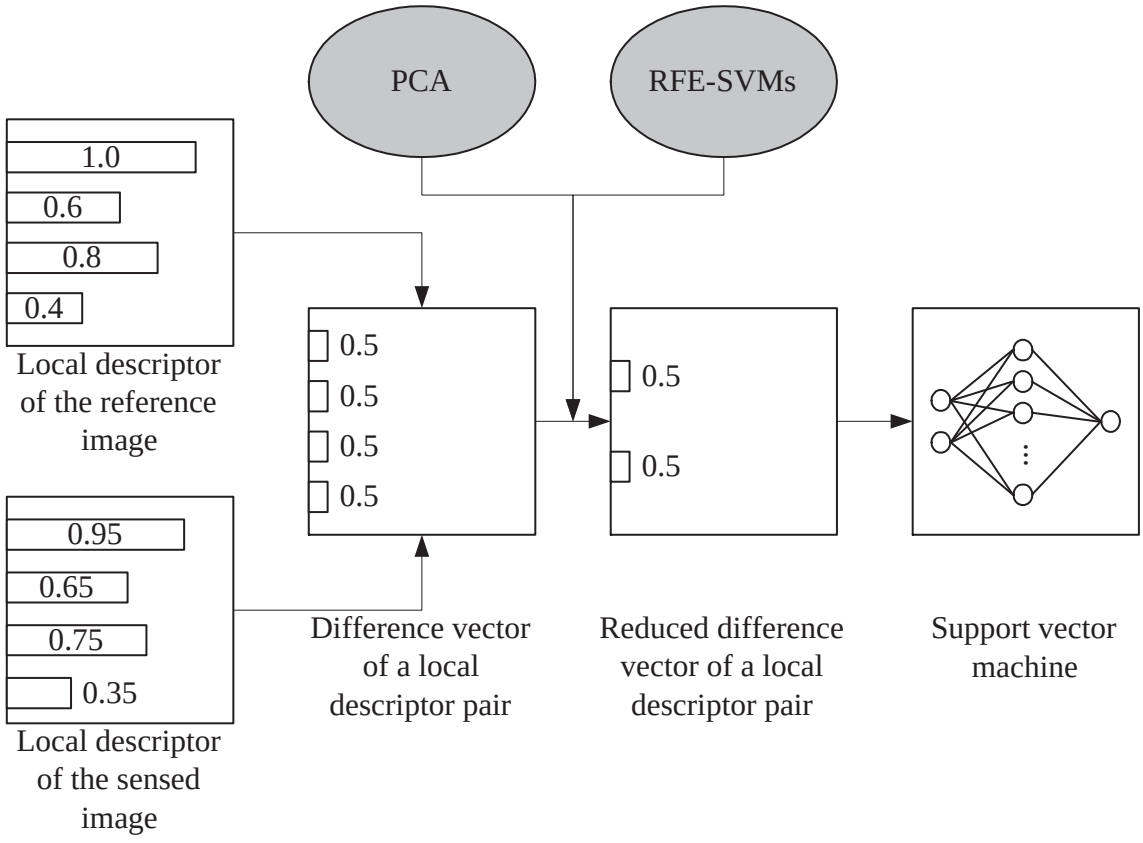


Figure 5.12: Example of how the feature-reduction methods using PCA and RFE-SVMs are integrated with the SVM matching method.

pair is computed. PCA is then applied to this difference vector, using a covariance matrix, which results in a reduced set of vectors. The reduced difference vector is then used as input to the SVM model for estimating the correctness of the match of the local descriptor pair.

The covariance matrix required for PCA can be estimated in a training phase. This is estimated using the same set of local descriptor pairs that was used to train the SVM model, and is performed before the SVM model is trained. One difference from training the SVM model is that instead of using both the correctly and incorrectly matched local descriptor pairs, only the pairs which matched correctly were used. This is due to only the important vectors of the correctly matched local descriptor pairs are of interest and of importance in improving the robustness of the method. As the size of the input data is changed when PCA is applied, the SVM model needs to be retrained and this is done after a covariance matrix is estimated from the training data and PCA applied to the input data. The Kaiser's criterion and scree graph, discussed in Section 4.3.2, were used to determine the optimal number of features which should be kept when using the PCA feature-reduction method.

Recursive Feature Elimination using Support Vector Machines

RFE-SVMs takes a different approach than PCA in performing feature-reduction of difference vectors of local descriptor pairs. Instead of transforming the data before selecting a subset of the PCs for local descriptor matching like PCA, RFE-SVMs reduces the number of features by using a subset of the original features from the difference vectors. The variables are first defined [194]: for a two-class classification problem with $(m, v+1)$, the input matrix is $\mathbf{X}^0 = [\mathbf{x}_1, \cdots, \mathbf{x}_m]^T$, the set of surviving features' indices, in other words, the features from the difference vectors yet to be removed from the subset **s**, and **r** the feature ranking list. Given these definitions, the process for RFE-SVMs is as follows: (1) **s** is set to a row vector $\mathbf{s} = [1, \cdots, m]$, and **r** set to an empty vector; (2) for each iteration, the training data $\mathbf{X}$ is a subset of the input data matrix $\mathbf{X}^0$, consisting of the features defined in **s**; (3) the SVM model is trained, and weights **w** are computed for all the features in the subset; (4) weights are ranked according to the square of their values $c_i = (w_i)^2$; (5) a new set is defined consisting of features from the subset in decremental order, and the ranking list **r** retains the list of features in order of their respective weights; (6) the p smallest features, defined by the square of their weights c_i are removed from the subset **s**. This process is repeated until all the features in **s** have been removed. The vector **r** contains the list of features from the difference vectors in order of importance, or the amount of information they contain which describes the matching of local descriptors.

In many real-life applications, instead of a single feature being removed at each iteration, often a subset of features are removed to reduce the computation time. This is, however, not the case in this research as the number of features involved in the matching of local descriptors is relatively small. An example in the case of a SURF local descriptor is that there are 64 features, compared to the original application of RFE-SVMs in gene selection in Deoxyribonucleic Acid or DNA studies, where thousands of features exist [210]. Due to this reason, in order to maximise the robustness of the method, only one feature is removed at each iteration until no features remain. Unlike the PCA method, where the Kaiser's criterion and scree graph can be used to determine the optimal number of features to retain, there are no such methods for RFE-SVMs. Because of this, different numbers of features were experimented with in the experiments conducted, to be discussed in detail in the next section, in order to identify the optimal solution, where a balance between the computation time and matching accuracy can be achieved.

5.4 Experimental Design

To compare the performance and demonstrate the matching accuracy of the presented local descriptor matching method using SVM against existing methods, experiments have been

carried out to validate these methods. In addition, to verify the robustness and versatility, three sets of experiments were conducted including: (a) training the SVM model using different datasets; (b) adapting the same approach for different machine learning algorithms to examine the versatility; and (c) integrating the developed local descriptor matching method with different local descriptor methods.

5.4.1 Euclidean Distance-Based Methods versus Support Vector Machines

The first set of experiments compared the SVM local descriptor matching method with existing methods, in particular, the threshold matching method. This was an important set of experiments as this provided a direct comparison of the methods, and an insight into the performance of the SVM method. To compare the threshold matching method with the SVM method, experiments were designed and conducted for the following four image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint changes.

5.4.2 SVM Models from Different Training Data

Initial experiments computed the SVM model using images of the object which the images need to be registered originated from. This approach was first used to verify the concept of using a machine learning method for matching local descriptors. This is however not practical in real-life applications, as in practice it is almost impossible to have a set of images from the object to be registered which can be used for training the SVM model, as this implies that the images of the object have known camera poses and therefore do not need to be registered.

To overcome this issue, experiments were carried out to validate the robustness of the SVM local descriptor matching method using different images for training the SVM model. In order to do this, instead of training the SVM model using a set of images from the same object that needs to be registered, which is not a practical solution in real-life applications, a set of images from the other available artefacts were used. In addition, by training and re-using the same SVM model for different objects, it demonstrated that the method is sufficiently robust to handle training data from different objects. The computation time required for training and using the SVM approach overall can be reduced significantly by using the same SVM model for different objects. The experiments were conducted using four sets of training data for training the different SVM models: (a) images from the flute artefact; (b) images from the patu artefact; (c) images from the wahaika artefact; and (d) images from both the patu and wahaika artefacts. For fairness of comparison, the number of training images used for all four SVM models were kept constant, and the training data were randomly selected from the available local descriptor pairs of the images. By using the

same number of training data, the conditions in which the models were computed were kept the same, and by randomly selecting the training data, this avoided bias from the data used.

5.4.3 Feature-Reduced Support Vector Machines

In order to determine the feasibility of the two feature-reduction methods discussed in Section 5.3.5, experiments were conducted which reduced the size of the input vector for the SVM model in an attempt to reduce the computation time required. The aim of this set of experiments was to reduce the number of input vectors as much as possible without compromising the robustness of the method. Different numbers of vectors were retained in the system using methods such as the Kaiser's criteria and scree graph for the PCA method, and the weights of the vectors for the RFE-SVMs method to identify the optimal solution.

5.4.4 Versatility of Machine Learning Algorithms

In addition to SVM, other machine learning algorithms were also studied. Given the vast number of machine learning algorithms developed over the past years, it is often a difficult decision to choose a suitable algorithm for the task. In this chapter, two of the most common machine learning algorithms, and a different kernel for SVM, were compared in an attempt to verify the choice of utilising SVM for the matching of local descriptors.

SVM with Polynomial Kernel

In addition to the Gaussian kernel, which is the common choice due to the true random distribution being a Gaussian distribution [208], SVM models using the polynomial kernel were also studied. SVM with a polynomial kernel is similar to Equation 5.9 and has the form:

$$y(\mathbf{x}) = \Sigma_{i=1}^{v} w_i \left(c_i^T \mathbf{x} + 1\right)^d + b \tag{5.12}$$

Where d is the order of the polynomial. An order of three was used, since a higher order often over-fits the model and is therefore not of practical use, as this would introduce instability to the system [194].

k-Nearest Neighbour

The k-NN is one of the simplest of machine learning algorithms, and is a type of instance-based learning method where the function is only approximated locally and all computation are performed in the classification phase [211]. A query is classified by a majority vote of its neighbours, where the query belongs to the class of the most common class amongst its k nearest neighbours. The distance between a query, in this case the difference vector of a local

descriptor pair, and its neighbours was determined by the Euclidean distance, although other distance measures could also be used. The parameter k was determined in the training phase and LOOCV was used for this purpose. The LOOCV instead of 10-fold cross-validation was used since LOOCV is much stricter compared to 10-fold cross-validation, and it can be expected that the model generated using LOOCV will be more accurate [194]. This is suitable for k-NN due to its relatively low computation requirement, however is not suitable in the case of SVM, as LOOCV for SVM is too time-consuming. When using k-NN for a two-class problem, it is often desirable to use an odd number for k, as this reduces the chance of tied votes, where a query might belong to two similar classes [211].

Adaptive Local Hyperplane

Adaptive Local Hyperplane (ALH) [212] is a recently proposed classification method based on the k-local hyperplane distance nearest neighbour (KHNN) method [213] which is in turn based on the k-NN method. In KHNN, the aim is to approximate the potentially unseen instances in classification by a local hyperplane. First the k-NN of a query is selected from each class as the prototypes using k-NN, then a local hyperplane is constructed to approximate the local manifold of each class, based on the prototypes. The class label is then assigned according to the distance between the query and the local hyperplane of each class.

Although KHNN has performed well in applications [214, 215], drawbacks exist in the KHNN approach. Firstly, KHNN only works well for small values of k [213], as the k-NN method for selecting the prototypes suffers from bias in high dimensions [212]. In addition, KHNN assumes that all the features are important for classification and may result in unsatisfactory performance for complex datasets. This issue is overcome in ALH where the feature weights are considered. In ALH, the prototypes are selected by the adaptive nearest neighbour method, and the feature weights are estimated by considering the ratio of the between- and within-class sums of squares. The prototypes are then assigned with the most discriminant feature dimensions. By considering the shape of the neighbourhood around the query, ALH handles high dimensionalities by fulfilling the assumption of k-NN that class conditional probabilities are constant [212].

5.4.5 SURF versus SIFT

The aim of integrating the proposed matching method with different local descriptor methods was to verify that the developed SVM method for local descriptor matching can be used for any local descriptor method. Two local descriptor methods were used in this study. In addition to the SURF local descriptor method, the integration of the method with SIFT local descriptors was also investigated. The SURF and SIFT local descriptor methods have

different number of features and were therefore ideal for demonstrating that the developed SVM method is versatile, works with any local descriptor method, and is robust against a change in the number of features of the local descriptor method involved.

5.5 Results and Discussion

5.5.1 Euclidean Distance-Based Methods versus Support Vector Machines

The results for the two different matching methods compared are shown in Figures 5.13 and 5.14. The first is the widely-used threshold method where the Euclidean distance of local descriptor pairs were used to determine the correctness of matches, and the second is the presented SVM matching method. For viewpoint changes, the results for all four artefacts are presented. For rotation, scale and tilt changes, results for the flute artefact are shown in this chapter, however similar trends were observed for the patu, wahaika and tiki artefacts. Because it was of interest to study the trend, since the methods were compared and the results are relative to each other, and as the trend of the results from the objects are similar, they are not presented in this chapter. Instead, complete results for the experiments conducted, as well as the experiments in the sections to follow, can be found in Appendix B.

As can be seen from the figures, the SVM matching method out-performed the threshold matching method by up to approximately 20%, depending on the image transformation studied. The results for scale changes for the flute artefact are shown in Figure 5.13b, and as can be seen, there are no significant changes in the matching accuracy for both types of matching methods for the different scales studied. This indicates that the local descriptors are robust against scale changes and the only time when scale changes had an effect is when the magnitude of change is large. An example is if an image is taken from far away, then the camera would not be able to capture the finer details of the object and would therefore cause a degradation in performance of the matching algorithm.

The results for rotation changes are shown in Figure 5.13a. The figures show a very slight decreasing trend in the matching performance as the rotation angle increased, however this is not significant. Ideally, a *t*-test using the test statistic or similar should be performed to determine statistically whether there is a trend in the results, however due to the data size this test statistic would not have been reliable and the test was therefore omitted.

For both the tilt and viewpoint changes, a notable decreasing trend can be observed in Figures 5.13c-5.13d and 5.14. These plots show that, for all four artefacts studied, as the change in tilt or viewpoint angle increased, the matching accuracy decreased. It can be seen that even with this decreasing trend, the SVM matching method consistently out-performed the threshold method and is capable of providing an accurate matching result.

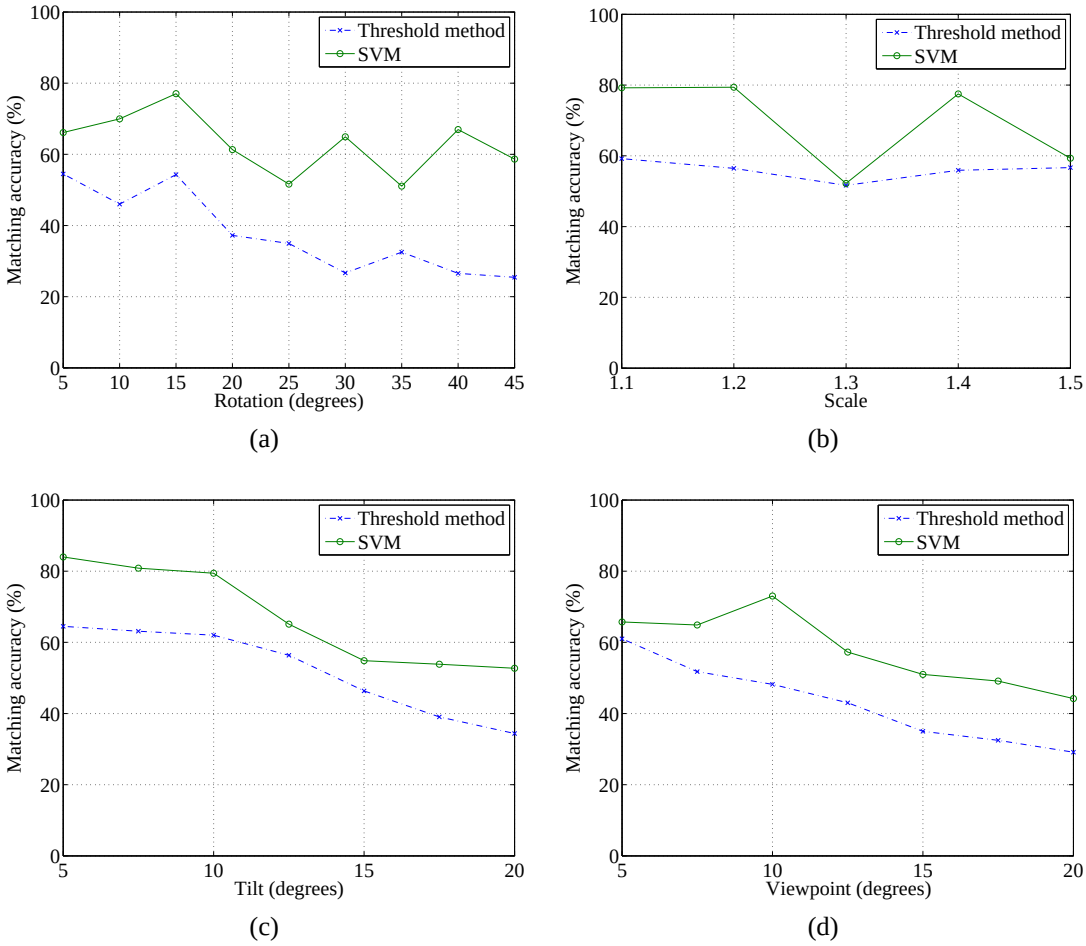


Figure 5.13: Image matching results using three different matching methods for the flute artefact with different image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint.

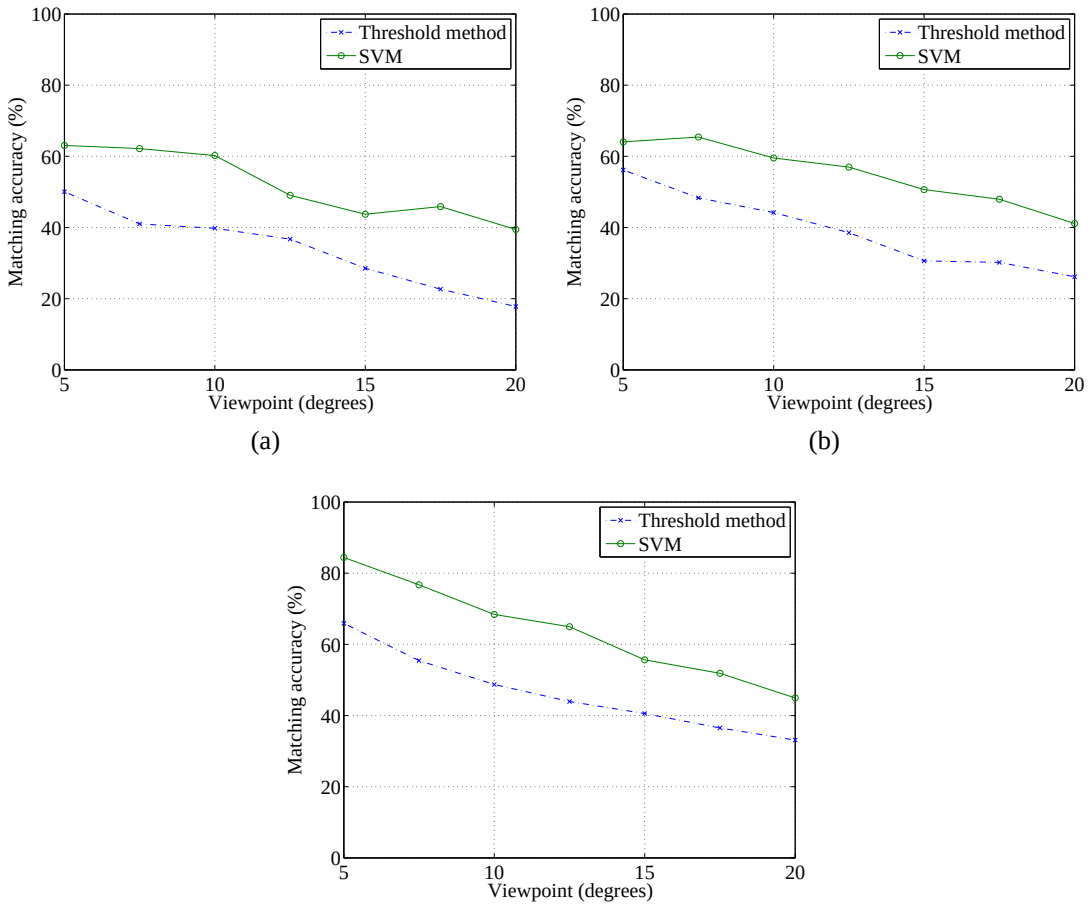


Figure 5.14: Image matching results using three different matching methods for the patu, wahaika and tiki artefacts with viewpoint transformations: (a) patu; (b) wahaika; and (c) tiki.

5.5.2 SVM Models Using Different Training Data

The results for SVM registration in Figures 5.13 and 5.14 were obtained by training the SVM model using images from different regions of the same object, that is, the images from the same object as the training images that needed to be registered were not included in the training data used for generating the SVM model. To check the robustness of the SVM model and ensure that the same model is usable for different objects, experiments were conducted to test the SVM model generated using images of the patu and wahaika artefacts. Figure 5.15 shows the matching results using the SVM models using the different training data from these objects.

For both results shown in Figure 5.15, four different training datasets were used when computing the SVM model. All four training datasets generated models, which were used on images of the flute artefact for matching local descriptors. For both the rotation and viewpoint changes shown in Figure 5.15, when using the different SVM models, noticeable trends were observed. For rotation changes, there is a slight decreasing trend in the matching accuracy, similar to the results discussed in Section 5.5.1. Figure 5.15a shows that the accuracy of the matches using the SVM model generated from the patu artefact is significantly worse than the other three models by approximately $15-20\%$, whereas the other three models show similar performance. By comparing the three artefacts, it was discovered that the patu artefact is significantly different from the flute and wahaika artefacts in terms of the types of features found on these artefacts. The patu artefact consists of mostly patterns from the wood grains and the pattern drawn on the object itself, whereas the flute and wahaika artefacts consists of highly detailed carvings, and these appear similar visually. This difference in the artefacts used contributed towards the decline in accuracy observed.

The results for viewpoint changes shown in Figure 5.15b show a similar pattern and reduction in accuracy to that found in rotation changes, where the matching accuracy for the SVM model computed using the patu artefact is noticeably lower than the other three models. A different pattern in this plot is that the results of the SVM model computed using the wahaika artefact is slightly lower than the other two models. The model generated using both the patu and wahaika artefacts again showed a similar performance to the model generated using images of the flute artefact itself. From these two figures, it can be seen that while some artefacts are more suitable for computing SVM models than others, it is possible to use images from different objects for training SVM models and then using this model for matching a different object, where there is no *a priori* knowledge, for example, the type of features and geometry about the new object. When more objects were used for training the model, the accuracy of the match was increased, and from this, two conclusions can be drawn on the type of objects that should be used for training the SVM model. First, if there is *a priori* knowledge of the object which need to be registered, then other objects which are similar to that object should be used for training the model. In the case that no *a priori*

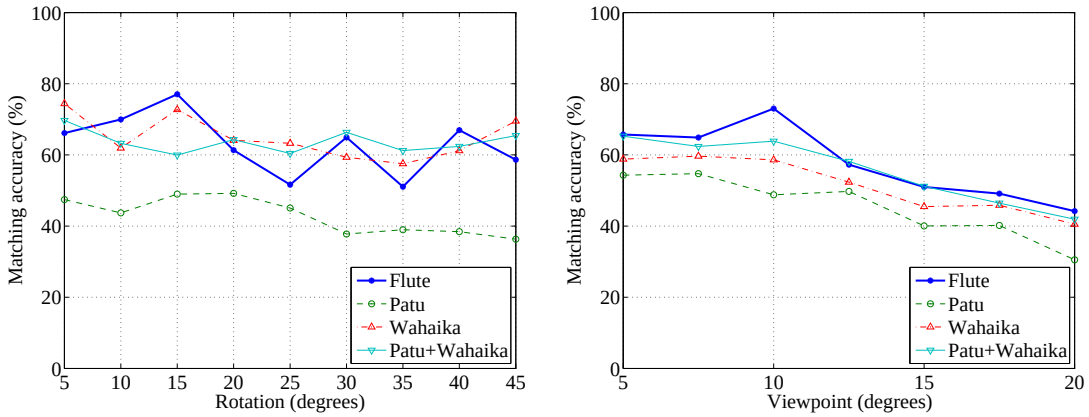


Figure 5.15: Image matching results by using SVM models generated from different artefacts for the flute artefact with different image transformations: (a) rotation (flute); and (b) viewpoint (flute).

knowledge is available, more objects should be used, as results from experiments show that this produced better matching results.

5.5.3 Feature-Reduced Support Vector Machines

Principal Component Analysis

The results for feature-reduction using PCA are presented in Figure 5.16b. The scree graph, as well as the Kaiser's criterion used for selecting the number of PC scores for the result presented are shown in Figure 5.16a.

The aim when selecting the number of features, or PCs in the case PCA is used, is to use as few features while retaining as much information as possible. From the Kaiser's criterion, as well as the 'elbow' of the curve in the scree graph, it can be seen that 11 PCs are suitable for reducing the number of features required, while retaining the robustness of the match. To verify this, in addition to the 11 PCs suggested by the two selection methods, two other values, nine and 13, were also used to compare the effects of using different number of PC scores.

Figure 5.16b shows the results for feature-reduction using PCA. As a reference, the results for SVM without feature-reduction and the threshold matching method are also shown. As can be seen, by using 11 PCs as suggested by the selection criterion, the matching accuracy is approximately 5 – 10% worse than the original SVM model. Using 13 PCs did not have a significant impact on the accuracy over using 11 PCs, as the matching accuracy only increased by 2 – 3%. By using nine components, the matching accuracy dropped by up to approximately 20% and is consistently worse than the threshold matching method. From the results, it is clear that by using PCA, it is possible to use a subset of the original data

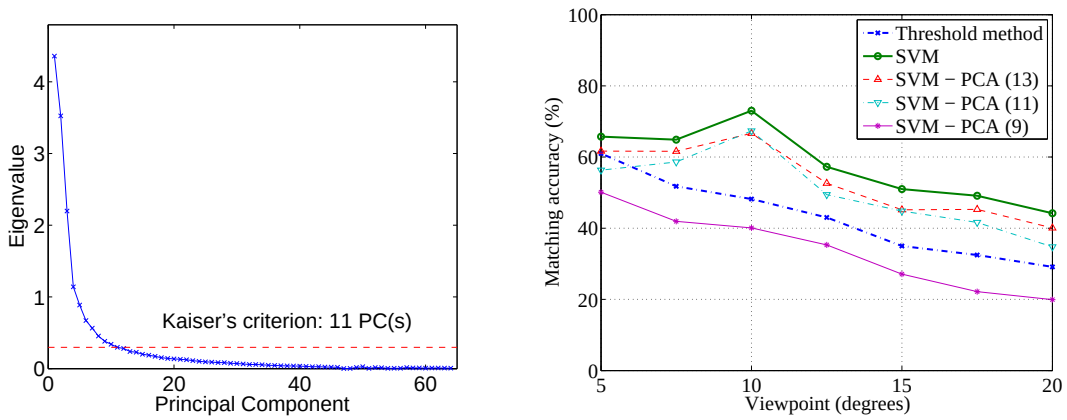


Figure 5.16: Feature-reduction: (a) scree diagram for selecting the number of principal components to use; and (b) image matching results using five different matching methods for the flute artefact with viewpoint transformations.

for matching and the reduction in matching accuracy is minor in the experiments conducted. One important factor to take into account is that since many image registration applications are offline processes, feature-reduction methods are often not required, as these methods sacrifice matching accuracy for computation time. The PCA method proposed and verified is best suited for when the computation time is of importance.

Recursive Feature Elimination using Support Vector Machines

Unlike PCA, no useful results were obtained by using RFE-SVMs to reduce the number of features of the difference vectors of local descriptor pairs. This was observed and confirmed when the weights of the individual features when computing the ranking criteria for RFE-SVMs were analysed. While the algorithm was able to rank the features according to the weights assigned, poor results were obtained. Upon further study, it was found that the weights were similar for many features, meaning that many features had similar importance and therefore by reducing the number of features, the performance degraded rapidly and therefore the results were of no use for local descriptor matching.

The difference in the results obtained for PCA and RFE-SVMs is due to the nature of the two methods. PCA first transforms the difference vectors of local descriptor pairs before feature-reduction takes place, whereas in the case of RFE-SVMs, feature-reduction is applied directly to the original data. By transforming the original difference vectors to a new set of axes, PCA was able to use this new set of axes to retain the critical information, represented by the new axes, while discarding the less significant ones, as shown in Figure 4.9a. RFE-SVMs on the other hand used the information from the difference vectors directly, and many vectors had similar importance and it was therefore difficult to determine which vectors were the more significant, and was therefore unable to be successful in reducing the number of

vectors or features required, while maintaining the accuracy of the method in the experiments conducted.

5.5.4 Versatility of Machine Learning Algorithms

While SVM is one of the most common machine learning algorithms today, it is beneficial to experiment with different machine learning algorithms to compare the performance of these approaches. This also demonstrates that the developed approach can be easily adapted for different machine learning algorithms. Two machine learning algorithms, *k*-NN and ALH, as well as SVM with a polynomial kernel, were studied. The results for this section are shown in Figure 5.17.

As can be seen, the four algorithms have similar performance, with ALH having the highest matching accuracy overall, which is approximately $1-2\%$ higher than SVM with a Gaussian kernel. SVM with a polynomial kernel performed similarly with *k*-NN which is just slightly worse compared to SVM with a Gaussian kernel by approximately $3-4\%$. These results agree with the results presented in [212], where ALH out-performed the other machine learning algorithms. One thing to keep in mind however is that due to the way queries are classified in ALH, the computational time is significantly higher than SVM. While the computation time is not a deciding factor in many registration applications, in the experiments conducted, the computation time required for ALH was approximately $10-20$ times of SVM. The computation time is dependent on the implementation of the algorithms, which were not optimised in this research as the computation time was of no important concern. However, given the similarity in performance of ALH and SVM, careful consideration should be taken when determining a suitable algorithm for the application. Based on the proven performance of SVM over the years [194], and the similar performance observed in the experiments, it was concluded that SVM is the most suitable approach for matching local descriptors in this research.

5.5.5 SURF versus SIFT

The last set of experiments conducted compared two local descriptor methods and aimed to demonstrate that the SVM method for matching local descriptors is adaptable for different local descriptor methods, consisting of different number of features. The results are shown in Figure 5.18. The SURF local descriptor method used is 64D while SIFT is 128D. From a previous study [80] and Chapter 3, it was shown that SURF out-performed SIFT in all cases when registering images of the objects studied in this research. This trend was not observed in the experiments conducted, where SIFT consistently out-performed SURF by approximately $1-3\%$. A logical explanation for this is due to the number of features of the two local descriptor methods. When the threshold matching method is used, the difference

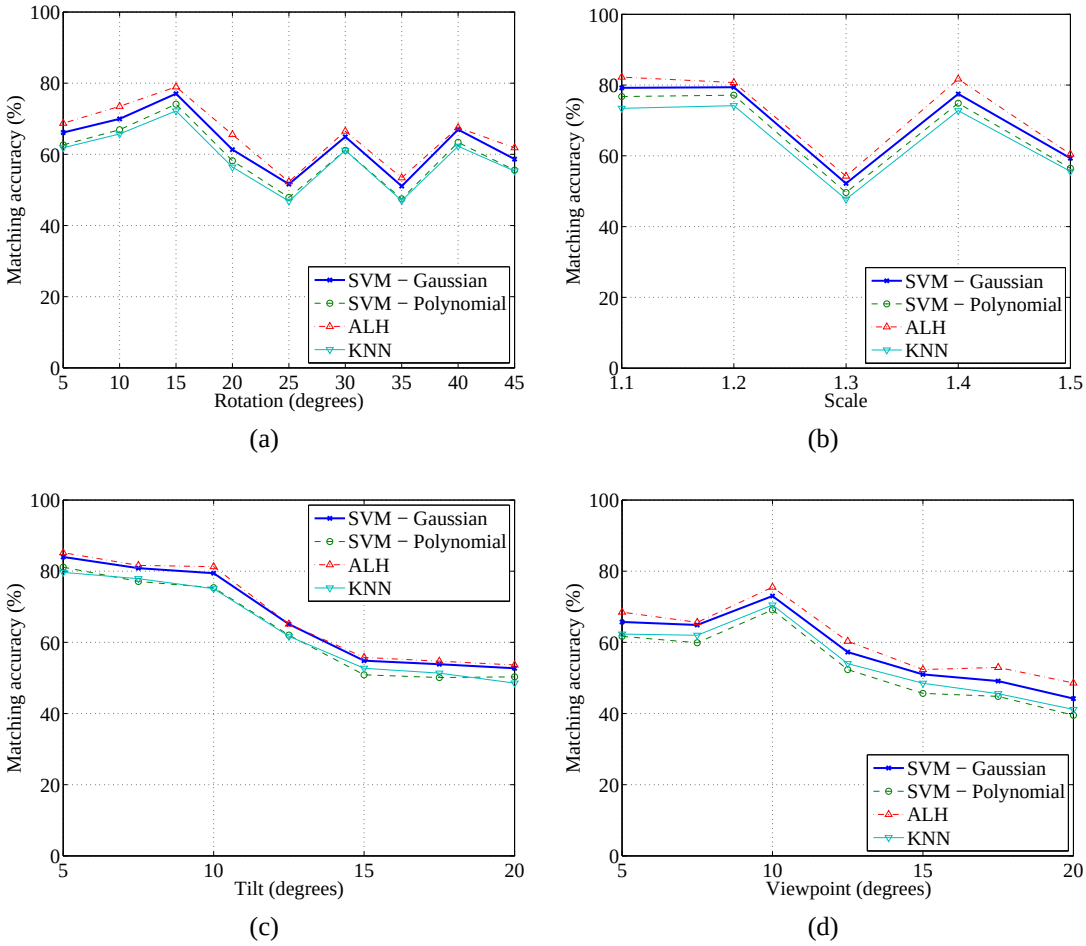


Figure 5.17: Image matching results using four different machine learning algorithms for the flute artefact with different image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint.

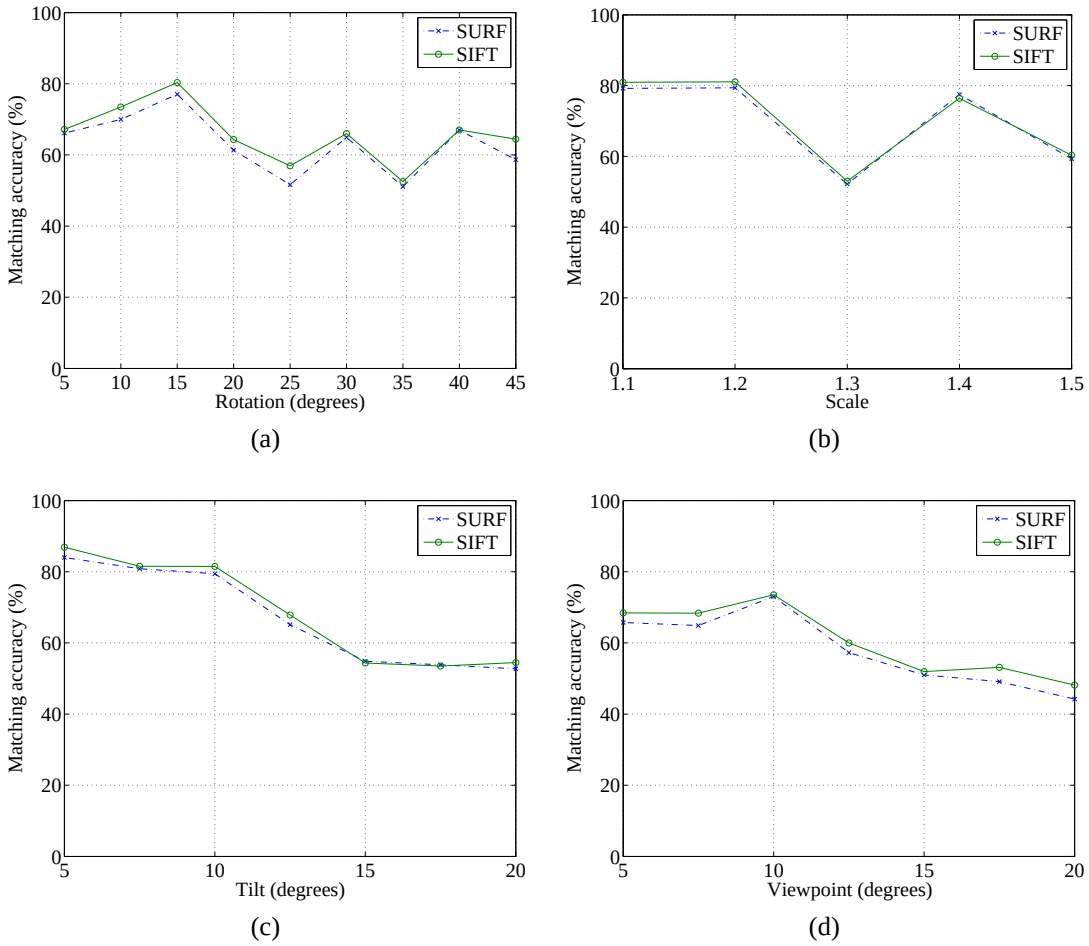


Figure 5.18: Image matching results using two different local descriptor methods combined with the SVM matching method for the flute artefact with different image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint.

vectors of local descriptor pairs are reduced down to scalar values before decisions are made on the correctness of the matches. In the case of SVM, all the individual values are used for decision making, and therefore the larger-sized SIFT appears to have an advantage over SURF.

From the results, it can be seen that the SVM matching method was integrated successfully with different local descriptor methods. A disadvantage is that due to the different nature and dimension of the two methods used in the experiments, the SVM model needed to be trained and developed for each local descriptor method even when the same image pair was concerned. This is, however, not considered a drawback as the training of a suitable SVM model is an one-time process for each local descriptor method.

5.6 Conclusions

This chapter presented a new approach based on SVM for matching local descriptors, and included processes for formulating the approaches for input data selection, model training, automatic parameter selection and classification of local descriptor matches. To validate the performance of this method, five sets of experiments were conducted. The first set of experiments compared the proposed method and the threshold matching method for different image transformations, and results show that the new method had higher matching accuracies of approximately $10-20\%$. From the results, it can be concluded that the SVM matching method is more robust for these image transformations. The second set of experiments aimed to identify the feasibility of using different objects for training the SVM model, and results show that no *a priori* knowledge is required in order to compute a suitable model.

A drawback of the SVM method was identified as its higher computation time, and to overcome this, two feature-reduction methods were proposed. It was concluded that while the RFE-SVMs method is not suitable for the task, the PCA method was able to reduce the computation time, and at the same time, keep the reduction in matching accuracy to approximately $5-10\%$. Despite this reduction in accuracy, it still performed better than the threshold method. The fourth set of experiments aimed to investigate the use of different machine learning algorithms. Results show that the ALH and SVM performed similarly, however ALH had a much higher computation time. Because SVM has a proven performance and is significantly more efficient, it was concluded that SVM is the most suitable approach for matching local descriptors in this research. The last set of experiments validated the robustness of the SVM method in handling local descriptors of different sizes, and it was shown that this can be achieved easily regardless of the type of local descriptor used.

This research which has never before been published demonstrated how SVM can be integrated with local descriptors. It is considered an important contribution in the field as this method can be integrated with different types of local descriptor methods, and can therefore be integrated with not only existing, but also future local descriptor methods. The matching accuracy of the integration of the SVM method with these methods is significantly improved as shown in the results presented. To enhance the method's usability for applications which require a faster computation time, feature-reduction methods were proposed. In addition, research into different machine learning algorithms provided a good foundation for further development in the future.

Chapter 6

Integration of Local Descriptor Methods and Assisted Image Registration

This chapter describes the integration of the local descriptor formation and matching methods. Experiments were carried out to compare the performance of the integration of these two sets of methods with existing methods. An assisted image registration programme was also developed, which combines the strengths of both the manual and computer-based image registration methods. Results from the experiments using this programme showed that it is capable of registering images in a more robust manner compared to pure image registration methods.

This chapter presents the research carried out in two distinct areas. Firstly, the integration of colour and hybrid local descriptor methods with the SVM matching method. Secondly, the development of an user-assisted registration tool for the purpose of improving the matching accuracy of images.

The first part of this chapter presents the experiments conducted which studied various combinations of the local descriptor formation and local descriptor matching methods, and their corresponding feature-reduction methods. The aim was to not only investigate how the matching accuracy can be improved when the two sets of methods are combined, but also how the feature-reduction methods can be utilised to reduce the computation time associated with the methods. Note that the computation time was regarded as a secondary objective, as the main objective of this research was to develop methods which can robustly register images. Changes in the illumination conditions were studied, since these conditions significantly affect the registration performance.

The second part of this chapter presents an assisted image registration approach to improve the performance of image registration. This was developed as an efficient tool to improve registration, as image registration methods have limited success when dealing with large viewpoint changes [9]. As the registration of images is often an offline process, and the

Table 6.1: The different combination of methods tested in order to identify the optimum combination of the two stages of local descriptor processes: (a) local descriptor formation; and (b) local descriptor matching. The values in the table refer to the order in which they are discussed in the thesis.

	SVM	Feature-reduction (PCA)
Colour local descriptor/hybrid local descriptor	(1)	(2)
Feature-reduction (PCA/weighted channel)	(3)	(4)
Illumination (colour/intensity)	(5)	

focus is on the robustness of the matches instead of a completely automated approach, the purpose of the development of this tool was to study approaches which require a minimal amount of user input and by doing so, improve the matching accuracy significantly. The work discussed in this section is of a more practical nature and looked at how the methods presented in the first part of this chapter can be further improved by combining a manual and computer-based approach.

This chapter is structured as follows. Experiments were conducted which aimed to investigate the different combinations of the colour local descriptor and hybrid local descriptor methods and the SVM local descriptor matching method, as well as their corresponding feature-reduction methods. The results and findings are presented in Section 6.1. Section 6.2 reviews existing software packages which have been developed for matching images, followed by a description of the developed user-assisted image registration programme in Section 6.3. This section also discusses the experimental setup and results achieved using the assisted image registration programme. The chapter is concluded in Section 6.4.

6.1 Integration of Local Descriptor Formation and Matching Methods

6.1.1 Experimental Design

Four image transformations, namely: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint, were experimented with for the different combinations of the two sets of methods. By combining the local descriptor formation methods with the SVM matching method, and taking into account the effects of the feature-reduction methods for both sets of methods, it was possible to study the different combinations of these methods separately. This allowed for a better understanding of which method contributes the most towards improving the overall accuracy of the local descriptor process. Four combinations of the two sets of methods were studied: (1) local descriptor formation methods with the SVM matching method; (2) local descriptor

formation methods with the SVM matching method and the PCA feature-reduction method; (3) colour local descriptor method with the PCA and weighted channel feature-reduction methods and the SVM matching method; and (4) colour local descriptor method with with the PCA and weighted channel feature-reduction methods and the SVM matching method with the PCA feature-reduction method. This is shown in Table 6.1, where the four combinations are numbered one to four in the table, respectively. This table is presented for ease of comparison between the different combinations of methods, and corresponds to the order in which the combinations are presented. For the PCA feature-reduction method utilised for both the colour local descriptor and SVM matching methods, the number of PCs were determined using the Kaiser's criterion [185] and scree graph [186]. In the case of the colour local descriptor method, the number of PCs was independent of the number of features found in existing methods. This is different than the approach taken in Chapter 4, where the number of PCs matched the number of features found in existing methods like the SURF method, in order to provide a direct comparison between these methods. This was utilised as the aim was to achieve the best balance of matching accuracy and computation time, as suggested by the Kaiser's criterion and scree graph.

In addition to these four sets of experiments, the effects of combining the local descriptor formation methods with the SVM matching method were also studied for the two illumination condition changes, namely colour and intensity. This is numbered five in Table 6.1.

Note that in order to simplify the discussion of results, when only one of the proposed methods was utilised in the experiments, an existing method for the other stage was utilised. For example, the statement "the use of the SVM matching method alone" means that an existing method for computing local descriptors was used. For computing local descriptors, the SURF local descriptor method was utilised, and for matching local descriptors, the threshold method was utilised.

Test 1: Local Descriptor Formation Methods Combined with SVM Matching Method

The first set of experiments combined the colour local descriptor and hybrid local descriptor methods with the SVM matching method. The aim was to determine the matching accuracy when these two sets of methods are integrated. The results were compared with existing approaches, including the SURF and CSIFT local descriptor methods, as well as the performance of the two sets of methods when they are not integrated. To conduct the experiments, first the colour local descriptors and hybrid local descriptors were computed for each of the images used in the experiments. The SVM models were then trained and computed on two sets of training data, one for each of the colour local descriptor and hybrid local descriptor methods. These two sets of methods were then combined to evaluate the performance of the combination of these methods. For the hybrid local descriptor method,

the SVM matching method replaced the threshold matching method, and the colour patches in the hybrid local descriptors were matched using the modified normalised cross-correlation algorithm [189].

Test 2: Local Descriptor Formation Methods Combined with Feature-Reduced SVM Matching Method

The experiments conducted in the previous section aimed to maximise the matching accuracy of the methods by combining the colour local descriptor and hybrid local descriptor methods with the SVM matching method. A downfall to the tests is that due to the increased number of features found in the local descriptors computed, the computation time was increased. While the robustness was of more concern than the computation time, it was also of interest to study the effects of applying the PCA feature-reduction method to the SVM matching method. The aim was to determine whether it was possible to reduce the computation time while maintaining the performance of the methods. In order to apply PCA to the SVM matching method, the number of PCs were determined by the Kaiser's criterion and the scree graph.

Test 3: Feature-Reduced Local Descriptor Formation Methods Combined with SVM Matching Method

This set of experiments applied the PCA and weighted channel feature-reduction methods to the colour local descriptors, and these were integrated with the SVM matching method. Similar to the first set of experiments, the feature-reduced colour local descriptors were first computed, the SVM models were then trained based on these local descriptors. The hybrid local descriptor method was not evaluated in this test, as no feature-reduction methods were developed for this method.

Test 4: Feature-Reduced Local Descriptor Formation Methods Combined with Feature-Reduced SVM Matching Method

The fourth set of experiments studied the effects of utilising feature-reduction methods on both the colour local descriptor and SVM matching methods. By utilising the PCA and weighted channel methods for the colour local descriptor method, and the PCA method for the SVM matching method, it is possible to significantly reduce the computation time. However, the most important factor in many image registration applications is to robustly register the images. As shown in previous experiments, feature-reduction methods sacrifice accuracy for computation time. Due to this drawback, there was a need to investigate the amount of reduction in performance caused by the feature-reduction methods.

Test 5: Local Descriptor Formation Methods Combined with SVM Matching Method for Illumination Changes

In addition to the four sets of experiments that combined the two sets of methods in different forms, a fifth set of experiments studied the robustness of these methods against illumination changes. The colour local descriptor and hybrid local descriptor methods were combined with the SVM matching method to register images taken under changes in illumination colour and intensity.

6.1.2 Results and Discussion

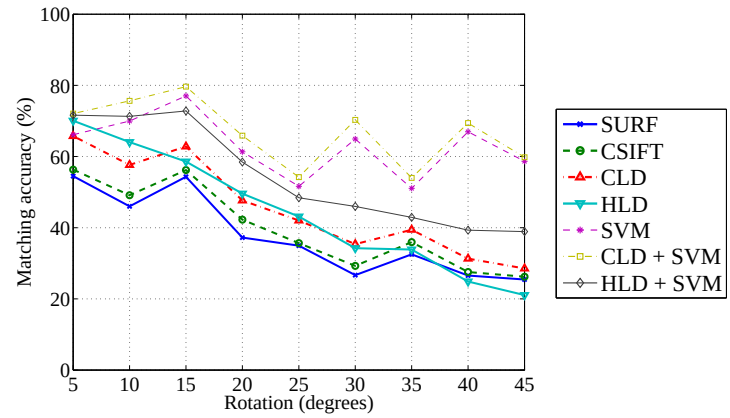
Before discussing the results in detail, the abbreviations used in the result graphs need to be discussed. Due to space constraints, in order to ensure that the legends for the figures are legible and that the names of the methods evaluated are presented in the figures, abbreviations were adapted. These abbreviations are only present in the figures and are not used in the discussions.

The colour local descriptor method is abbreviated as 'CLD' and the hybrid local descriptor method is abbreviated as 'HLD' in the figures. In the case that a feature-reduction method was applied, the relevant method is followed by the word 'with'. For example, 'CLD with PCA' denotes that the PCA feature-reduction method was applied to the colour local descriptor method. When a local descriptor formation method was combined with the SVM matching method, the $+$ sign was used. For example, 'CLD $+$ SVM' denotes that the colour local descriptor method was combined with the SVM matching method.

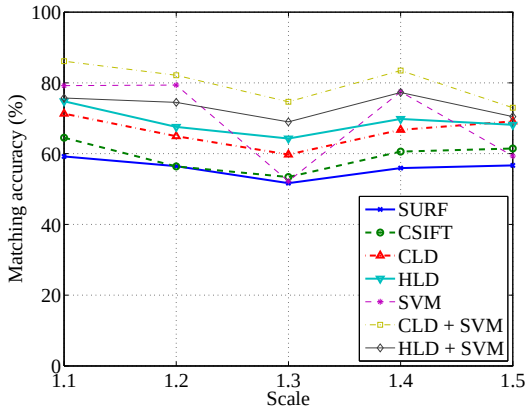
Test 1: Local Descriptor Formation Methods Combined with SVM Matching Method

The results for the different combinations of the two local descriptor methods and the SVM matching method are shown in Figures 6.1 and 6.2. For the viewpoint changes the results for all four artefacts are presented. For the rotation, scale and tilt changes, results for the flute artefact are shown in this chapter. Because it was of interest to study the trend, since the methods were compared and the results are relative to each other, and as the trend of the results from the objects are similar, they are not presented in this chapter. Instead, complete results for the experiments conducted, as well as the experiments in the sections to follow, can be found in Appendix B.

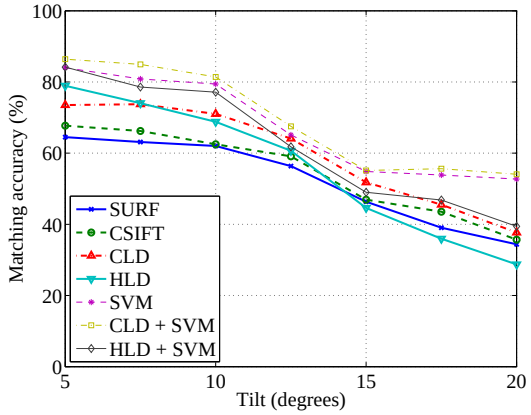
When the colour local descriptor method was combined with the SVM matching method, improvements were observed for all image transformations, with gains of approximately $5-10\%$ in matching accuracy over the use of the SVM matching method alone. This results in improvements of approximately $20-25\%$ over the use of the SURF method. The trend of the results of the combination of these two methods follows closely of that of the SVM



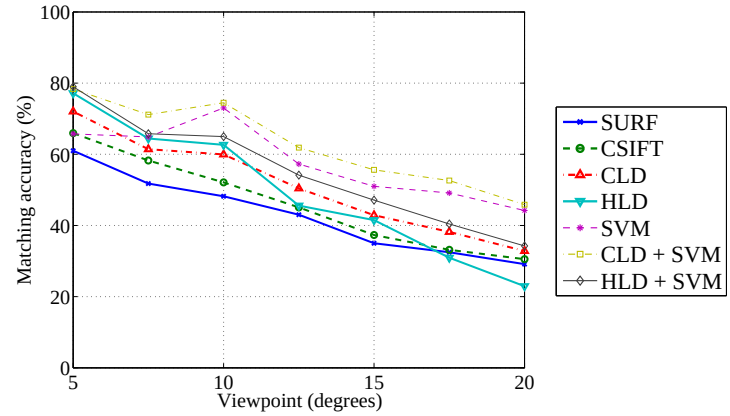
(a)



(b)



(c)



(d)

Figure 6.1: Image matching results using the colour local descriptor and hybrid local descriptor methods combined with the SVM matching method for the flute artefact with different image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint.

matching method alone. These results indicate that the conventional threshold matching method has significant drawbacks for matching local descriptors.

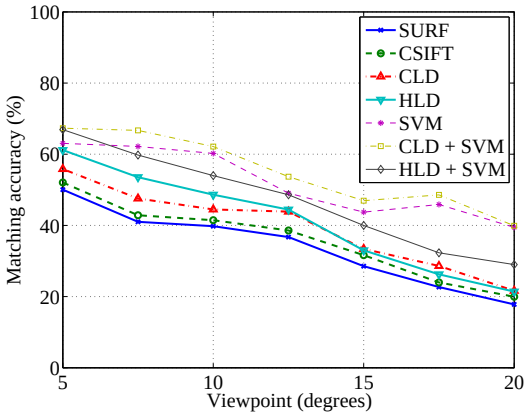
For the hybrid local descriptor method, improvements were less significant compared to the colour local descriptor method when combined with the SVM matching method. The matching accuracy is higher compared to the use of the hybrid local descriptor method alone, with gains of up to approximately $5-15\%$ over the hybrid local descriptor method. However, depending on the image transformation and scale of transformation concerned, the overall matching accuracy is lower than that of the SVM matching method alone. This was due to the nature of the hybrid local descriptor method. Unlike the colour local descriptor method, the local descriptors computed with the hybrid local descriptor method are combinations of feature-based local descriptors and area-based colour patches. As the area-based colour patches were matched using the modified cross-correlation method [189], the SVM matching method was only applied to the feature-based part of the method. Because of this, the performance gains of the SVM method observed in other experiments, for example when the SVM matching method was integrated with the colour local descriptor method, was not as significant.

From the results shown in this section, two conclusions can be drawn. First, when the colour local descriptor method was integrated with the SVM matching method, the best performance was observed. This was consistent for the range of all the image transformations studied. Second, when the hybrid local descriptor method was integrated with the SVM matching method, a similar trend to when the hybrid local descriptor method was used was observed. The accuracy reduced faster than the other methods evaluated for the rotation, tilt and viewpoint changes. The only exception to this were the results for scale changes. Because area-based methods are less affected by scale changes [9], this combination was not affected by changes in this transformation. This supports the claim in Chapter 4, where it was concluded that the colour local descriptor method is a more robust method across a wide range of image transformations compared to the hybrid local descriptor method.

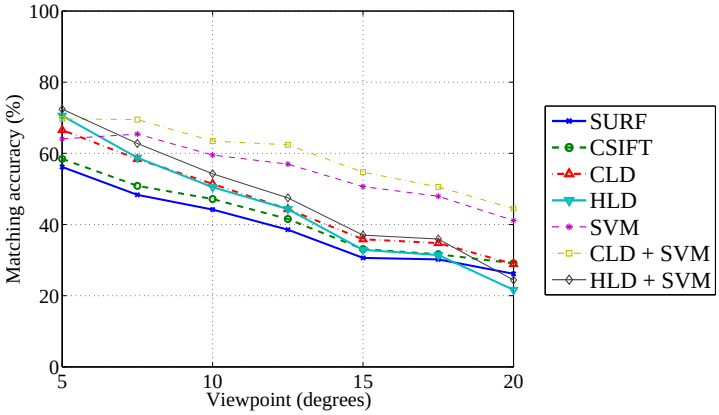
Test 2: Local Descriptor Formation Methods Combined with Feature-Reduced SVM Matching Method

The results for the different combinations of the two local descriptor formation methods and the SVM matching method with the PCA feature-reduction method are shown in Figure 6.3. For all the results presented in this section, it can be seen that when PCA was applied to the SVM matching method, the performance degraded.

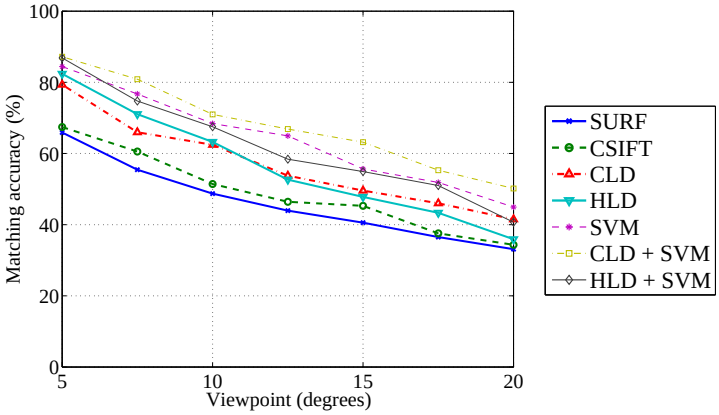
For the colour local descriptor method, the reduction in accuracy is approximately $5-10\%$ compared to the combination of the colour local descriptor and SVM matching methods as shown in test 1. The trend of the results follows closely of that of the colour local descriptor and SVM matching methods combination. This trend is similar to those presented



(a)



(b)



(c)

Figure 6.2: Image matching results using the colour local descriptor and hybrid local descriptor methods combined with the SVM matching method with viewpoint transformations for the: (a) patu; (b) wahaika; and (c) tiki.

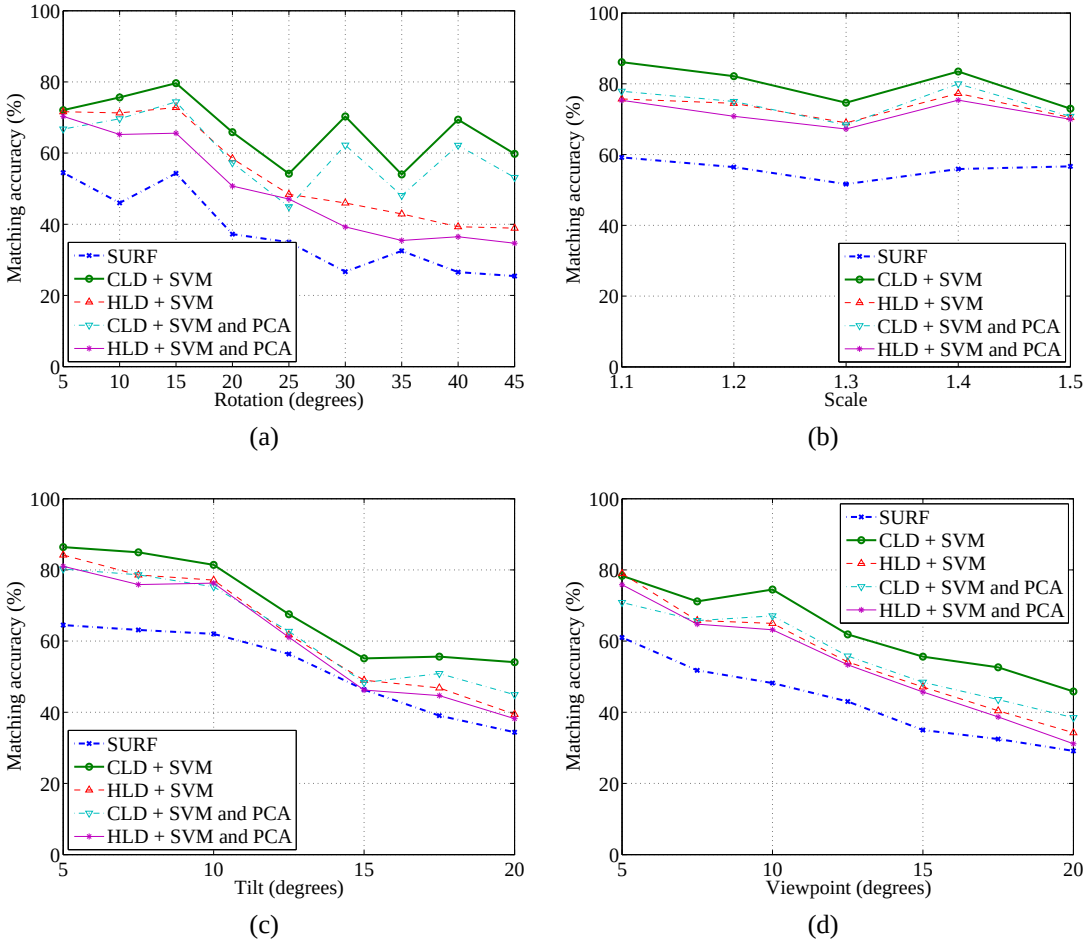


Figure 6.3: Image matching results using the colour local descriptor and hybrid local descriptor methods combined with the SVM matching method with the PCA feature-reduction method for the flute artefact with different image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint.

in Section 5.5, when PCA feature-reduction was applied to the SVM matching method. This indicates that the SVM matching method and the PCA feature-reduction method works consistently regardless of the type of local descriptor method utilised.

For the hybrid local descriptor method, the reduction in performance is not as significant compared to the colour local descriptor method. The matching accuracy is approximately 5% lower compared to the hybrid local descriptor method combined with the SVM matching method for all the image transformations. This is consistent with the results presented in test 1, where due to the combination of area- and feature-based methods, the hybrid local descriptor method was not as significantly affected by the SVM matching method. As a result of this, when feature-reduction was applied to the SVM matching method, the reduction in matching accuracy was not as significant for the hybrid local descriptor method compared to the colour local descriptor method.

From the results, it can be seen that for small changes in rotation, tilt and viewpoint changes, the combination involving the hybrid local descriptor method performed better than the colour local descriptor method one. For all three transformations, this gain in accuracy quickly reduced as the magnitude of transformation increased over $5-10^{\circ}$ as can be seen from the results. This was again due to the nature of the hybrid local descriptor method as discussed in test 1. The only exception was for scale changes for the same reason.

Test 3: Feature-Reduced Local Descriptor Formation Methods Combined with SVM Matching Method

The results for the different combinations of the colour local descriptor method with its two feature-reduction methods and the SVM matching method are shown in Figure 6.4. The hybrid local descriptor method was not considered in this section as no feature-reduction methods were developed and applied to the hybrid local descriptor method.

For both the PCA and weighted channel feature-reduction methods, by combining these methods with the SVM matching method, the matching accuracy improved by approximately $5-10\%$ compared to the colour local descriptor method with either of the two feature-reduction methods alone. The improvements observed in this test are lower than the performance gain achieved in test 2, when the PCA method was applied to the SVM matching method. This was due to the way the SVM matching method works. Instead of representing the difference vectors of local descriptor pairs using scalar values, the SVM matching method utilises all the available values from the difference vectors, and therefore by reducing the number of features in the local descriptors, the number of inputs for the SVM matching method was reduced which led to a reduction in the matching accuracy gain. As a result, unlike the previous section, the difference and drop in matching accuracy in the results presented in this section was found to be higher. Results from Figures 6.3 and 6.4 show that, on average, when either the PCA or weighted channel methods were applied to

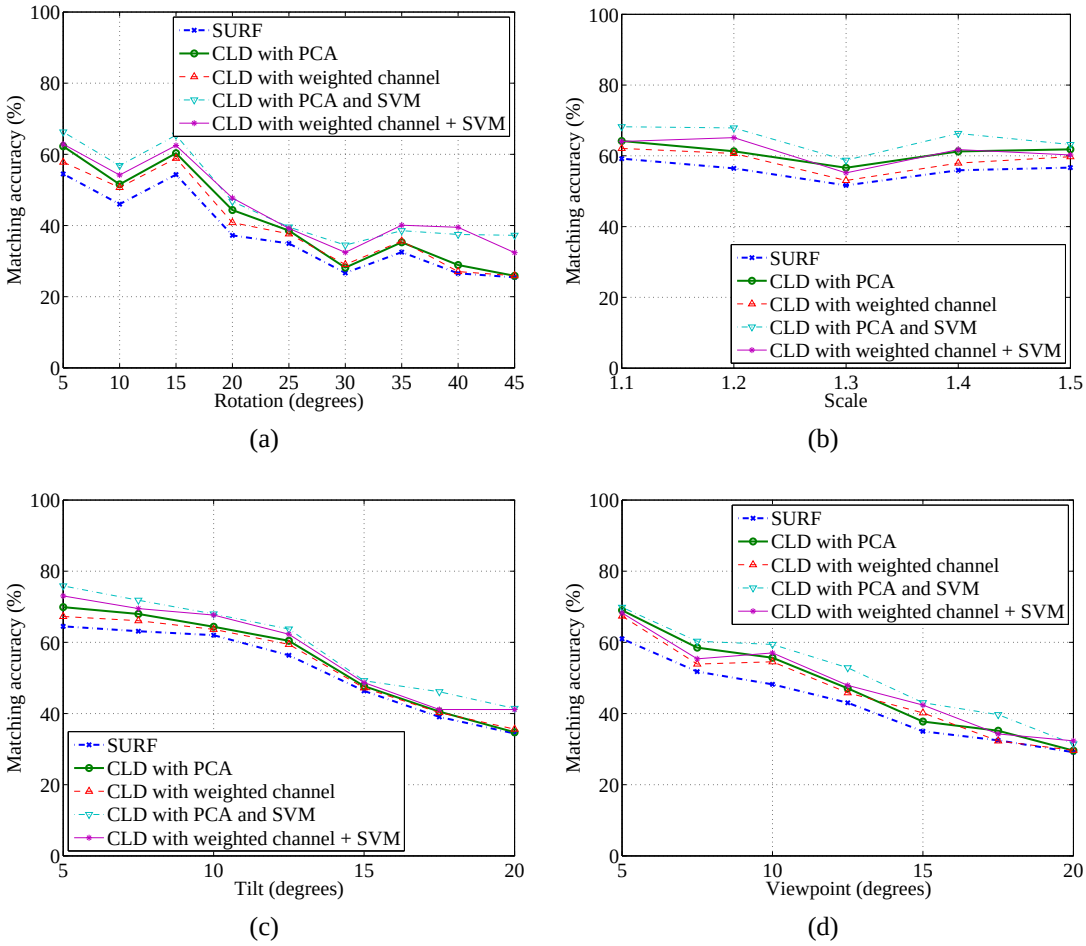


Figure 6.4: Image matching results using the colour local descriptor method with the two feature-reduction methods combined with the SVM matching method for the flute artefact with different image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint.

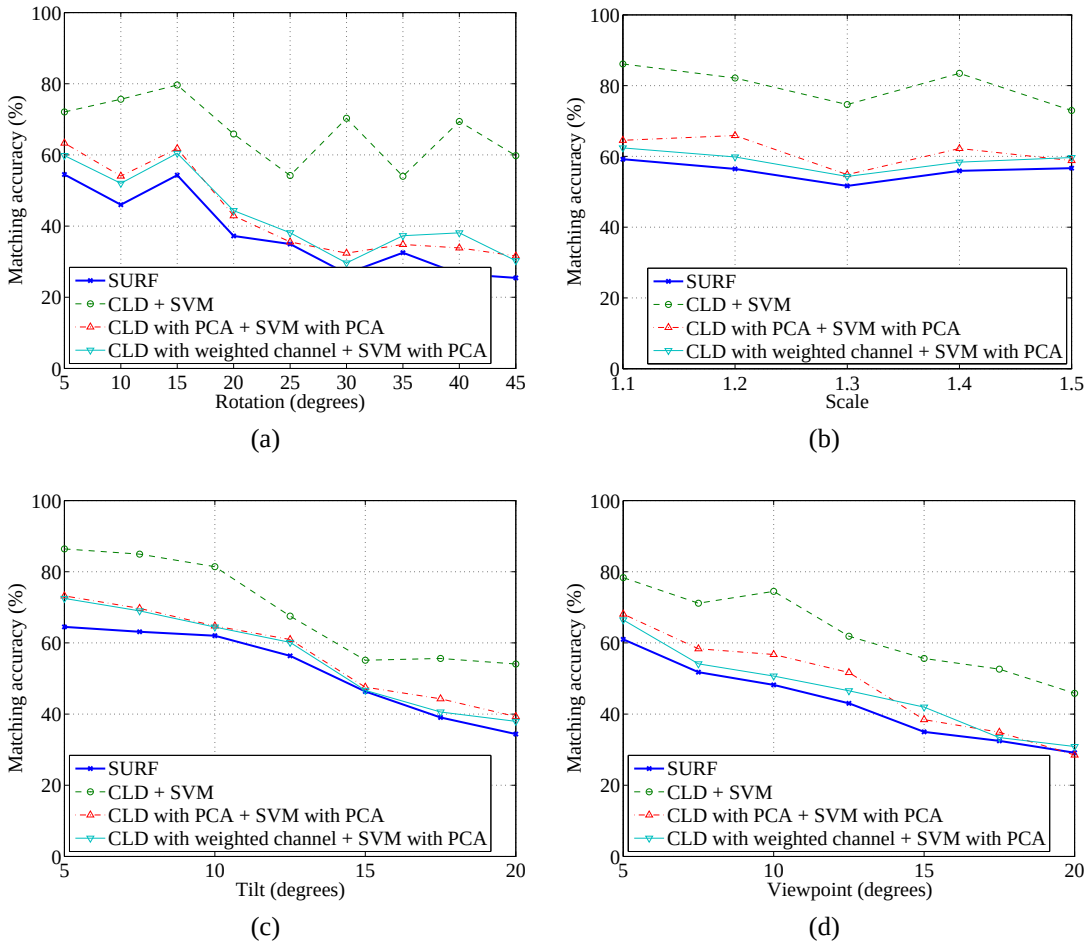


Figure 6.5: Image matching results using the colour local descriptor with the two feature-reduction methods combined with the SVM matching method with the PCA feature-reduction method for the flute artefact with different image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint.

the colour local descriptor method, the accuracy was approximately 5% lower than when the PCA method was applied to the SVM matching method.

Test 4: Feature-Reduced Local Descriptor Formation Methods Combined with Feature-Reduced SVM Matching Method

The results for this section are shown in Figure 6.5. As with test 3, because no feature-reduction methods were developed for the hybrid local descriptor method, no tests were carried out for this method.

As can be seen in the figure, the reduction in performance from the combination of the colour local descriptor method with the SVM matching method are significant for both the PCA and weighted channel feature-reduction methods integrated with the PCA feature-reduction method for the SVM matching method. The reduction in matching accuracy ranges from approximately $15 - 25\%$ for all four image transformations studied. In most cases,

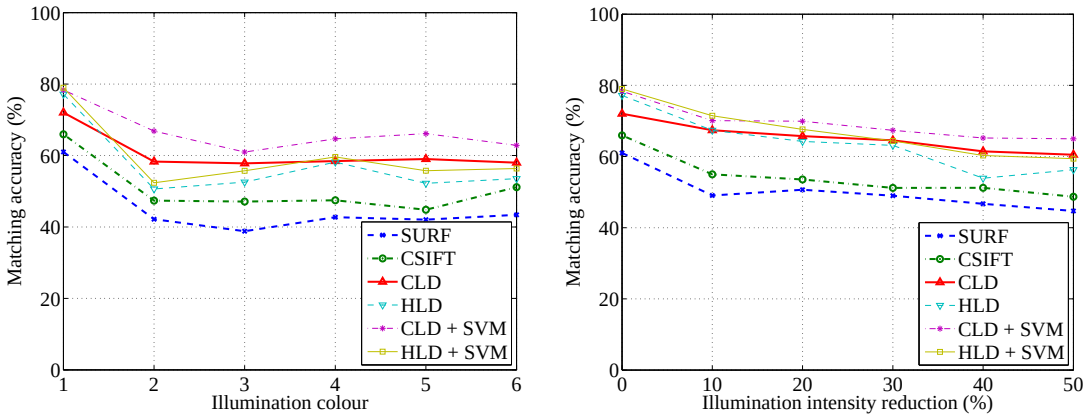


Figure 6.6: Image matching results using the colour local descriptor method combined with the SVM matching method for the flute artefact with different illumination changes: (a) colour; and (b) intensity.

when the PCA method was applied to the colour local descriptor method, better results were obtained compared with the weighted channel method. This difference in accuracy ranged from approximately $0-5\%$. The only exceptions were due to natural fluctuations in the results.

Despite this reduction in matching accuracy for both combinations, they were consistently higher than the conventional approach which utilised the SURF local descriptor method and the threshold matching method. From the result graphs, the gain in accuracy ranged from approximately $0-10\%$. This reduction, however, meant that no significant advantages were gained by using the integration of these methods, and is therefore not suitable for real-life applications due to the low accuracies observed.

Test 5: Local Descriptor Formation Methods Combined with SVM Matching Method for Illumination Changes

The fifth set of experiments was conducted to study the performance of the colour and hybrid local descriptor methods integrated with the SVM matching method in dealing with illumination condition changes. This section presents the matching accuracy of the SURF and CSIFT local descriptor methods, as well as the colour local descriptor and hybrid local descriptor methods and their integration with the SVM matching method for changes in illumination colour and intensity. The results are shown in Figure 6.6.

By combining the colour local descriptor and hybrid local descriptor methods with the SVM matching method, it can be seen that for the colour local descriptor method, the matching accuracy is approximately 5% higher for both changes in illumination colour and intensity compared to using the colour local descriptor method alone. This is consistently the best performing combination for the range of colour and intensity changes studied in this

test.

For the hybrid local descriptor method, the improvements were not consistent throughout the two illumination condition changes, and ranged from approximately $0-5\%$ compared to using the hybrid local descriptor method alone. This is in contrast to the colour local descriptor method, where consistent improvements for the different transformations were observed. This is similar to the results in the first two sets of experiments, and the reason is attributed with the nature of the hybrid local descriptor method, and the fact that the SVM matching method only improved the feature-based component of the method. Regardless, the matching accuracies observed were consistently higher than both the CSIFT and SURF methods as shown in the figure.

For both combinations and both illumination condition changes, it can be seen from the figure that for intensity changes, both combinations described above had more gains in accuracy over the CSIFT and SURF methods. This ranged from approximately $10-20\%$ for intensity changes. This is slightly higher than the experiments for colour changes, where the gains ranged from approximately $5-20\%$. This is due to the colour model utilised which can handle changes in intensity better than colour [180].

6.2 Recent Work in Assisted Image Registration

In recent years, software packages have been developed for registering and aligning images, in particular for stitching multiple images to form panoramic photographs [216]. These software packages range from full graphical editing programmes to dedicated software, which are designed solely for the purpose of matching images. One of the most well-known graphical editing software for such purposes is Photoshop by Adobe [217]. However, full graphical editing software packages are often excessive for tasks such as matching images and instead, more popular software packages include PTGui [218] and hugin [219], both derived from the Panorama Tools or PanoTools developed by Dersch [220, 221, 222, 223].

Another software package for this purpose which is more similar in nature to the methods developed in this research is Autostitch [224, 12]. The software is based on the SIFT local descriptor and RANSAC methods. The main difference between this and other software packages is that Autostitch is capable of matching images, to a certain extent, that have been misaligned or have undergone scale changes without user input due to the capabilities of SIFT. The software packages discussed above utilise a similar approach to matching or stitching images. First, a set of images are taken from the same location, these images are then used as inputs into the software. For each image pair, a set of corresponding features are manually selected and based on this selection, the transformation of the image pairs are computed and the images are then transformed and aligned.

The main drawback of the reviewed software packages which limit their use in registering

images for 3D reconstruction purposes is that they require the images to be taken from the same location [216, 224]. While this is not an issue for panoramic images, for 3D reconstruction a good coverage of the object is required and thus the images need to be taken from different camera poses.

A research area which is closely related to assisted image registration is the use of landmarks or control points for registering various types of images. Control points are points that are used to guide image registration algorithms, and there are two types of control points, intrinsic and extrinsic [25]. Intrinsic control points are markers in the images which are not part of the object itself, but are rather used for aiding the registration process due to their easily identifiable features. Control points are often used in medical applications, where they are referred to as fiducial markers [225]. An example of this is in [10, 11, 226, 227] where fiducial markers were used in assisting in the registration of fluoroscopic images for a robot-assisted long bone alignment system. Extrinsic control points are control points determined from the images itself and do not require special markers when images of the object of interest are taken. This type of control points can be extracted either manually or automatically. They are used in many image registration algorithms. Control points should be rigid, stationary and easily identifiable in all images to be registered. Intrinsic control points, while significantly simplifying the task of identifying corresponding features from image pairs, are intrusive and not always practical in real-life applications. In the case of medical applications discussed previously, sometimes it is required to insert the fiducial markers into a patient's body.

In the case of Māori artefacts studied, due to the way many of these objects are stored and their high historical values, it would be desirable to avoid moving the artefacts in order to place landmarks. In order to generate more accurate 3D models, the object would ideally occupy the majority of the images in order to allow for a higher quality of 3D reconstruction, therefore placing landmarks away from the object in order to be non-intrusive is also not a viable option. Extrinsic control points suit the need of this research and as they are non-intrusive, no physical landmarks are required to be placed around the objects. Due to the limitations of existing software packages which mainly deal with panoramic images and require the images to be taken from the same location, it was determined that the development of a simple and yet effective programme for registering these images utilising extrinsic control points would be of great assistance to the methods presented in Chapters 4 and 5.

6.3 Assisted Image Registration

6.3.1 Manual Selection of Corresponding Point Pairs

Despite advancements in image registration techniques over the past decades, the human eyes are still unparallel in identifying and matching images in many cases, due to the eyes' ability to recognise and match objects from a set of images in a much more efficient manner. A major drawback with the human eyes, however, is that it is very difficult to obtain an accurate match of images. Image registration techniques on the other hand excel in aligning images to the pixel or sub-pixel resolution if correct matches can be easily identified [58, 82, 228, 229]. To take advantage of both the human eyes and image registration methods, a programme was developed which integrates image registration techniques with the manual selection of points from images. The programme first lets end-users manually select one or more corresponding point pairs from image pairs. Based on these point pairs, image registration methods discussed in Chapters 4 and 5 are utilised to accurately match the images.

For each image pair, the aim is to obtain an accurate homography matrix which describes the spatial relationship of the images, and as such, accurate matching of the corresponding points is required. As discussed in Chapter 3, at least four pairs are required to compute the homography matrix. Due to imperfections of the location of the points, in order to compensate for these errors, more than four point pairs are almost always used, resulting in an overdetermined system. This overdetermined system is then solved using the RANSAC algorithm in combination with a least squares fit approach.

In addition to using an overdetermined system for improving the accuracy of the homography matrix computed, an edge detector can be utilised for the manually selected points. Since it is very difficult to manually select the exact location of points for the human eyes and in order to improve the accuracy of these manually selected points, edge detectors can be used to achieve sub-pixel accuracy of the location of these points. The use of an edge detector, in this case a Harris detector [44] refined to provide sub-pixel accuracy [82], is optional because the manually selected points are not always used for computing the homography matrix. Often, these points only serve as starting points for searching for a set of corresponding local descriptors. By reducing the search space by using these manually selected points as starting points, there are two advantages. As there are less local descriptors from the reference image which need to be compared for each local descriptor in the sensed image, the computation time is reduced significantly. Also, with less potential matches for each local descriptor, the chance of mismatches is reduced, which means that the matching accuracy can be increased and there is more room for error between a matching local descriptor pair, as the threshold for matching can often be increased while maintaining the desired matching accuracy, therefore increasing the robustness of the method. Two methods for reducing the size of the search space are utilised in the programme

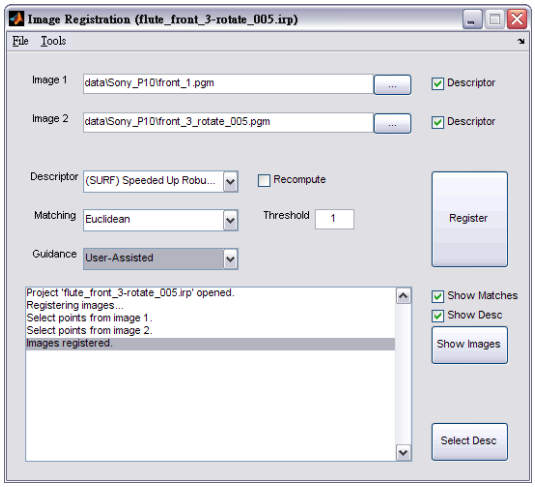


Figure 6.7: A screen shot of the developed MATLAB programme for automatic or assisted image registration.

developed, and these will be discussed in the sections to follow.

Once a set of matched local descriptor pairs has been computed, the homography matrix can be computed. To ensure that the correct homography matrix has been computed, two methods are available for checking the correctness. The first is a visual approach, which allows end-users to visually check the match of the images by projecting and overlaying the sensed image onto the reference image, while the second is a numerical approach where each local descriptor is checked for consistency. This is done by projecting each local descriptor in the region of the sensed image, which overlaps the reference image onto the reference image, then comparing each projected local descriptor from the sensed image with the closest local descriptor in the reference image. In the case that a correct homography matrix has been computed, a high percentage of the local descriptors would be correctly matched. Ideally, all the local descriptors would be correctly matched, however in practice it is not possible to rule out issues such as noise in the images and as such, only a pre-defined percentage of local descriptors needs to be correctly matched in order for the algorithm to determine that the homography matrix computed is correct.

Figure 6.7 shows the user interface of the MATLAB programme developed for the user-assisted image registration approach. The programme allows for easy selection of the type of local descriptor formation, matching and what kind of assistance the user can provide for registering images, as well as various parameters such as the threshold for matching local descriptors.

6.3.2 Reduced Search Space

Search Region

The first approach utilised for reducing the search space for identifying local descriptor correspondences is to define a search region for each pair of points manually selected. The process is shown in Figure 6.8.

First, a set of points are manually selected from the two images, with each pair of points corresponding to the same point in 3D space. Once a set of point pairs has been selected, a search region is defined for each point in the images. Two important notes are: first, in order to have a wide coverage of the images, the points selected should be as diverse and spread out through the images as possible. Second, at least two pairs of points should be defined. In theory one pair would be enough. While some experiments showed that this is also practically achievable, sometimes not enough coverage is provided by just one pair, and hence for performance reasons at least two pairs should be defined. The third step involves comparing the local descriptors in each search region in the reference and sensed images, and a set of corresponding local descriptors can be derived more accurately based on the search regions, as the number of potential matches are reduced by utilising the search region. Lastly, once all the local descriptors in all the search regions have been matched, the homography matrix is computed.

Note that in Figure 6.8, only local descriptors from one search region in each image are shown, these are denoted by the cross mark in each of the images in the second to fourth steps. This is only for ease of representing the local descriptors in the diagram, and all the local descriptors from all regions are used in computing the homography matrix.

Search Direction

The second approach for reducing the search space defines a search direction based on two point pairs. Instead of defining search regions, a search direction is defined based on the orientation and monotonicity of the location of local descriptors in the images. This process is shown in Figure 6.9.

First, two point pairs are selected for each image. Based on these two point pairs, two lines are defined, one for each image. The lines intersect the two points and divide the images into two regions. The third step is to search for correspondences, and this is achieved by searching for correspondence based on these two regions. For ease of discussion, the regions in the images are referred to as the left and right regions as shown in the figure. For each local descriptor in the left region of the reference image, a search for correspondence is made only in the left region of the sensed image. The same holds for local descriptors from the right region. This reduces the number of comparisons required for each local descriptor by approximately half. The last step is the same as the search region approach, where the

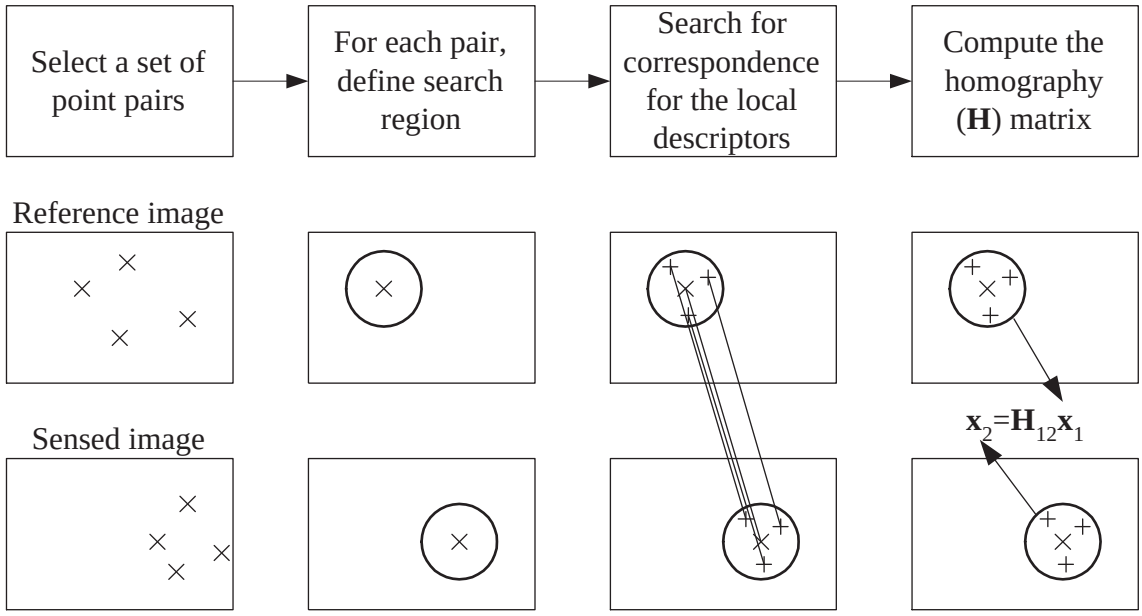


Figure 6.8: Flowchart of the search region method and figures showing the reference and sensed images.

homography matrix is computed based on the matched local descriptors.

6.3.3 Experimental Design

In order to determine the usefulness of having an user-assisted approach for registering images with large magnitudes of image transformations, experiments were conducted which studied the following four image transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint. An issue with the experiments conducted for this section was that the results in this section were highly dependent on the points selected by the end-user. Because of this, it was difficult to conduct bias-free experiments without human errors affecting the results. To minimise this effect, the number of points manually selected in each image was restricted to three. While it is possible that the matching accuracy would have been higher by using a higher number of points, as the number of points manually selected increased, the amount of time required for the operator is increased. This was avoided to keep with the aim of minimising user-intervention where possible. When only one point was used, there was insufficient coverage of the images. When two points were used, mixed results were obtained and was therefore discarded as no useful conclusion could be drawn from the results. This was due to the lack of spread of the points in the images.

In the case of the search direction approach which required two point pairs, two points which divided the images as evenly as possible were selected. To further ensure the fairness of the experiments, where possible, the same points in 3D space were selected for the different images. For all the experiments conducted in this section, the local descriptors

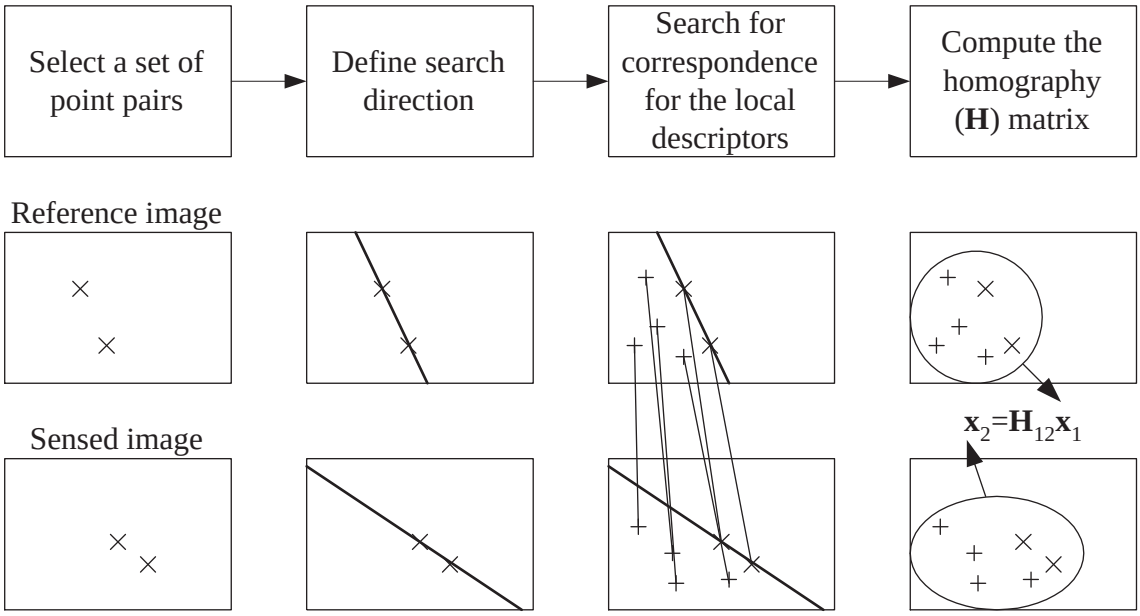


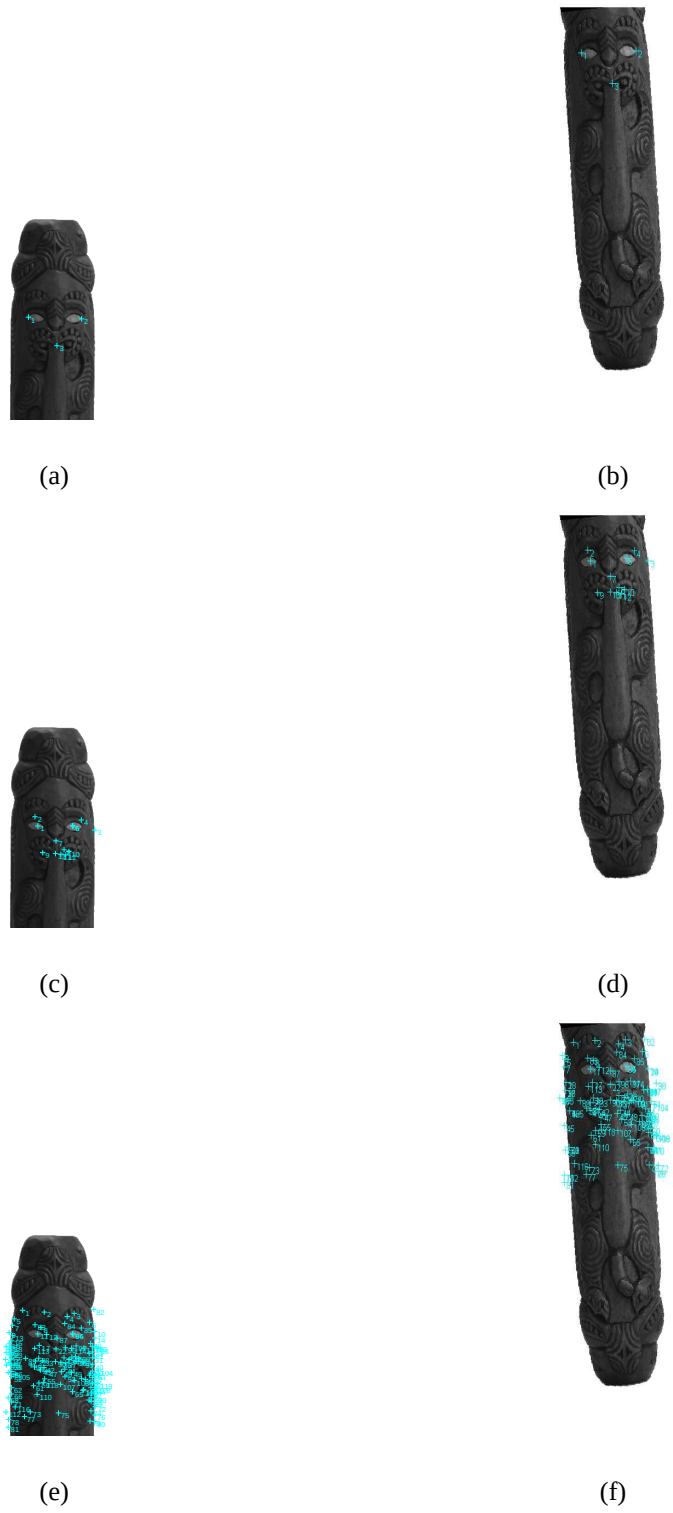
Figure 6.9: Flowchart of the search direction method and figures showing the two images involved.

were computed using the colour local descriptor method and the matching of these local descriptors were done using the SVM local descriptor method method.

6.3.4 Results and Discussion

Figure 6.10 shows two images of the flute artefact, and the steps involved in matching the images using the programme developed. The images in the first row are the reference and sensed images. As can be seen in the figure, three pairs of points were manually selected. These point pairs refer to the same locations on the flute, however the location of the points were not refined and thus are not perfect. This was ignored as these points only serve as starting points for defining search regions for searching for correspondences, and are not used in the actual matching of images. The second row of images shows the local descriptors around these manually identified points, defined by the search regions, and their corresponding local descriptors in the other image. The last row contains the two images which show all the matched local descriptors.

Figure 6.11 shows the matching of the local descriptors from the two images in Figures 6.10e and 6.10f. Lines are drawn in the figure to show the matched local descriptors. The MATLAB programme developed allows end-users to select a local descriptor from either the reference or sensed image at the top half of the figure, and based on this selection, the line which relates the two local descriptors is highlighted. When selecting a local descriptor, the nearest local descriptor is automatically selected. This can be helpful when trying to identify a pair of mismatched local descriptors. Note that in Figures 6.10 and 6.11, the images shown



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

Figure 6.10: Process of manually selecting three corresponding point pairs from an image pair: (a) and (b) show the three point pairs manually selected from the image pair; (c) and (d) show the local descriptor pairs matched from the regions around the selected point pairs; and (e) and (f) show all the matched local descriptors.

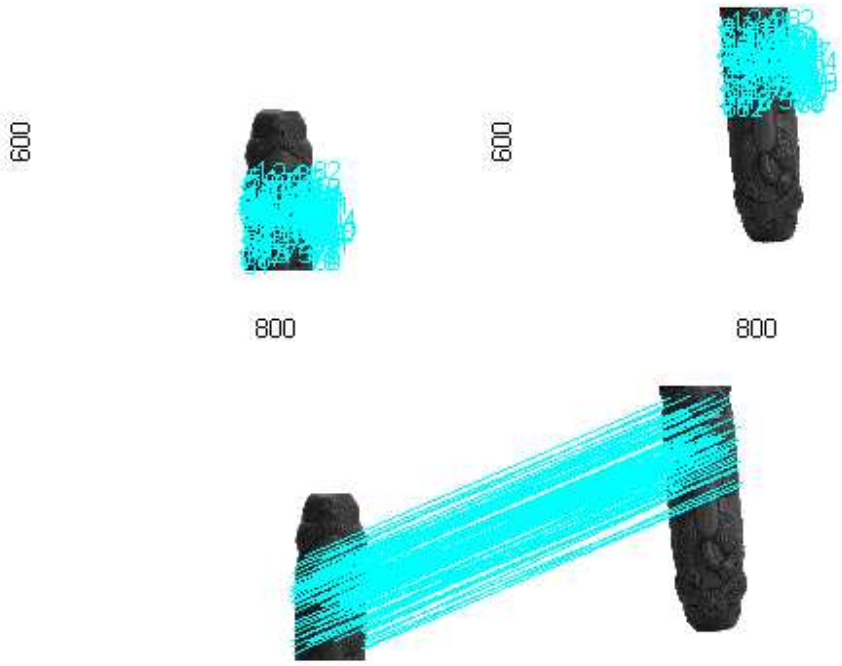


Figure 6.11: The two figures show the matched local descriptor pairs from the image pair and the bottom figure shows the matching of the local descriptor pairs.

are greyscale images. These images are only shown in greyscale as it is easier to mark the selected points on the resulting images. The images used for matching were coloured ones as the colour local descriptor method was used for computing the local descriptors.

Search Region

The matching results using the search region method to reduce the search space for the flute artefact for rotation, scale, tilt and viewpoint changes are shown in Figure 6.12. The figures include image matching results using the SURF local descriptor method and the colour local descriptor method combined with the SVM local descriptor matching method. Results for the other three artefacts are presented in Appendix B.

From these results, it is clear that by manually selecting a few points from both images, the matching accuracy can be greatly improved. For rotation and scale changes, the results show that the user-assisted approach is very robust and the performance is not significantly affected when the magnitude of image transformation was increased. The consistent performance of the method for rotation changes is due to there being no distortions of the object in the images. By reducing the search space and relaxing the threshold criteria, it is possible to robustly register the images regardless of the rotation changes unlike the other methods evaluated. For scale changes the results were expected, where consistent matching

results were observed with all the different methods compared. For both tilt and viewpoint changes, the matching accuracy decreased as the angle of transformations increased. The rate of reduction was, however, much lower compared to the methods compared in the figure, showing that the user-assisted approach is more robust against image transformations. Depending on the type of image transformation and the magnitude of the transformation, the improvement in matching accuracy ranged from approximately 25% to 50% compared to the SURF local descriptor method.

For all image transformations, a decrease in computation time was also observed, since the number of local descriptor pairs which needed to be checked was significantly reduced. The computation time for these experiments are not presented as this was not the primary aim. Furthermore, because the implementation of the various algorithms were not optimised, and due to variations in the hardware used for the experiments, it is impossible to provide an accurate and fair comparison of the computation time of all the experiments conducted throughout the course of this research.

The results from this section demonstrate the importance of having such a tool for registering images. Experiments demonstrated that by selecting three pairs of points from an image pair, the programme was capable of significantly improving the matching accuracy in all cases studied. These advantages out-weight the need to manually select these points in many applications, where real-time performance is not required.

Search Direction

For the search direction method, no results are presented for the following reasons. First, due to the way the images are segmented using the search direction method, the search space is still large compared to the search region method after images are divided. For example, each region may be one-tenth of the size of the image when the search region method is used, depending on the size of the region defined. This is in contrast to the search direction method, where the size of the region is approximately half the size of the image. This meant that there was plenty of room for mismatch compared to the search region method, resulting in poor results compared to the search region approach. Secondly, it was difficult to properly segment the images, as this is dependent on the two point pairs manually selected. In many cases, one image was segmented in a way that the two regions covered approximately the same amount of area while for the other image, the regions were of significantly different sizes.

Due to the fact that much consideration was required to ensure that the regions in both images were of similar size, this increased the difficulty and amount of user interaction required. This was impractical in many situations, especially when compared to the search region method, where it is much simpler to select a set of points which have a good coverage of the entire images. Even in the case shown in Figure 6.10, where the points selected are

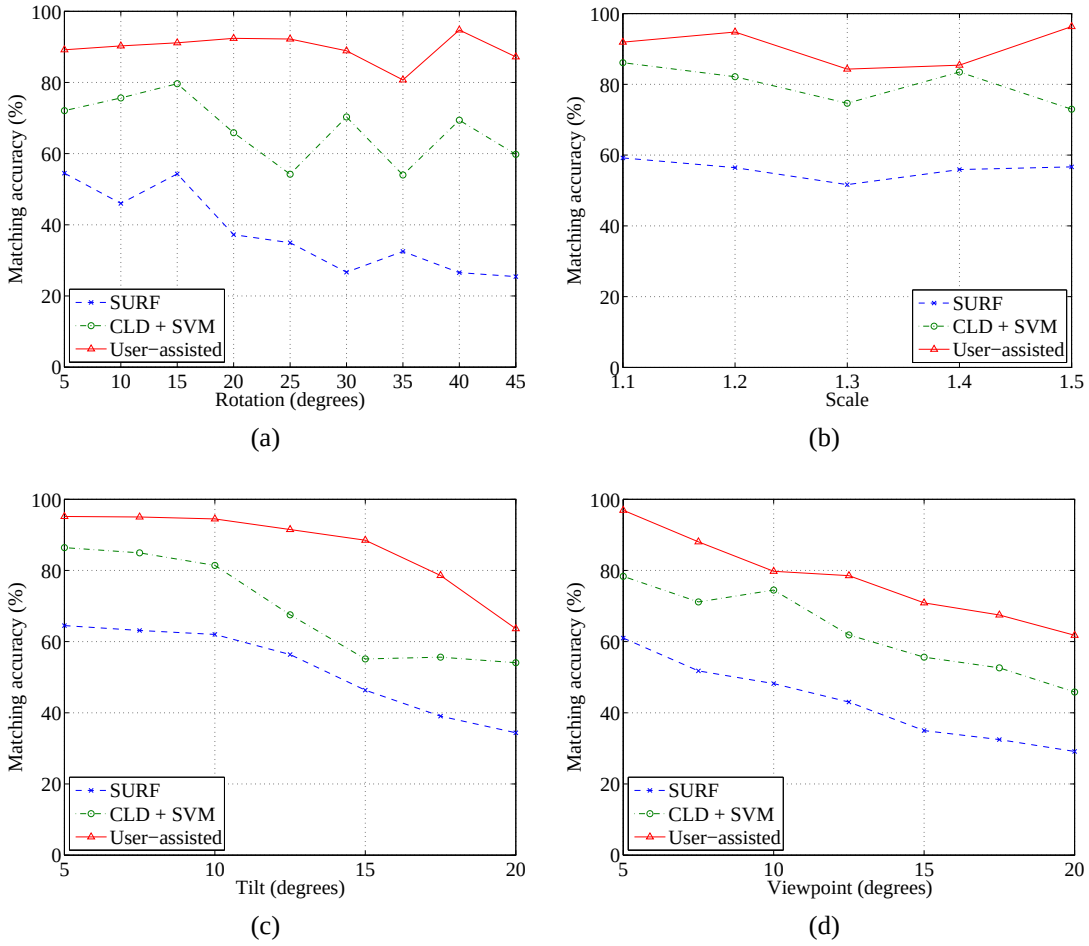


Figure 6.12: Image matching results using three different matching methods for the flute artefact with different transformations: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint.

relatively close to each other with respective to the size of the images, good matching results were still obtained using the search region method. Due to these drawbacks, the search direction method was considered to be inferior to the search region method in every way and thus was not pursued further in this research.

6.4 Conclusions

This chapter presented two pieces of work. First, the colour and hybrid local descriptor methods and the SVM matching method were integrated. Secondly, an assisted image registration programme was developed.

For the integration of the methods, five sets of experiments were conducted which studied the performance of different combinations of these methods under different imaging conditions. For the four image transformations studied, the best performance was observed when the colour and hybrid local descriptor methods were combined with the SVM matching method, and no feature-reduction was applied to both methods. In this case, the matching accuracy was up to approximately $20-25\%$ better than the use of SURF local descriptors and threshold matching method. For changes in illumination colour and intensity, improvements of approximately $10-20\%$ were noted for the same combinations. In all the experiments conducted, the integration of the colour local descriptor method with the SVM matching method performed better than the integration of the hybrid local descriptor method with the SVM matching method over the range of image transformations studied. This was due to the hybrid nature of the hybrid local descriptor method, and therefore the performance gain observed with the use of the SVM method was reduced compared to the colour local descriptor method.

For the assisted image registration programme developed, two search methods were developed in order to reduce the search space, and therefore improve the robustness and at the same time, reduce the computation time. Experiments conducted showed that the search region method was more versatile and practical. Results from the experiments showed that improvements of up to approximately 50% were possible. Furthermore, the rate at which the matching accuracy reduced as the magnitude of image transformation increased was reduced. This means that it is possible to cover a wider range of image transformations with this programme compared to relying on computer vision algorithms alone for registering images.

There are two contributions from this chapter. The first integrated the two sets of methods for computing and matching local descriptors. These methods were first investigated separately in previous chapters in order to conduct controlled experiments for each stage, and focus on the issues faced in these stages. The integration of the methods was necessary because they are part of the complete local descriptor process and are used together in image

registration applications. This integration provides a complete local descriptor process for registering images, and results from experiments show that they are significant improvements over existing local descriptor methods. The second was the assisted image registration programme developed, which is an useful tool for registering images, because it is capable of registering images in a much more robust manner compared to pure computer vision approaches. This is supported by the results from experiments, and highlights the importance of having this programme, where the advantages out-weigh the need to manually select point pairs from images.

Chapter 7

Conclusions and Future Work

The concluding remarks are presented in this chapter, providing a summary of the methods, algorithms and tools developed, and the results obtained from the experimental work to verify the performance of these algorithms. The contributions made throughout the course of the research are then highlighted. A list of future work is also outlined, suggesting potential improvements to the work presented in this thesis.

7.1 Conclusions

This thesis presented two sets of new local descriptor methods which tackled two different stages of the local descriptor process. Three sets of controlled experiments were conducted which first analysed and evaluated the performance of the developed methods while keeping the other stages constant, then combined the two sets of methods from the two stages to evaluate the overall performance of the methods developed. In addition, another set of experiments was conducted for the assisted image registration programme which integrated the developed local descriptor methods with inputs from end-users in order to further improve the robustness of image matches. In all the experiments conducted, improvements were observed.

In Chapter 1, the motivation for the research was presented. It was found through a comprehensive review of current literature that current image registration methods still struggle to register images in the presence of large magnitudes of image transformations. This is defined as approximately 22.5° in this thesis by taking into consideration the capabilities of the state of the art in 3D reconstruction. Because of this struggle, it was of interest to further develop methods in this field. In particular, the registration of images in preparation for 3D reconstruction was studied. Four Māori artefacts were used as case studies in this research, due to the added benefit of being able to contribute towards the reconstruction of these artefacts, which are important to New Zealand. The requirements

for this application were presented along with the scope and contributions resulting from the research.

Chapter 2 reviewed algorithms in the related fields. First, an overview of various image registration methods proposed over the years was discussed. This was followed by a detailed discussion on local descriptor processes. These methods are a subset of feature-based methods which showed promising performance for registering images when large magnitudes of image transformations exist. A review of 3D reconstruction methods in computer vision and a comparison of the performance of the state-of-the-art techniques in this field were also presented. As the particular issue in image registration studied in this research was to register images for 3D reconstruction, a detailed discussion on how the required parameters for 3D reconstruction can be obtained by image registration methods was presented. Lastly, projects which aimed at reconstructing artefacts or cultural and historical sites around the world were discussed. In addition, the challenges the EPICS team at the University of Auckland were faced with in dealing with the reconstruction of Māori artefacts from the Auckland War Memorial Museum by using laser scanners were discussed at the end of the chapter.

An evaluation study was carried out and presented in Chapter 3, which compared the performance of various existing local descriptor processes. The aim of this evaluation study was to identify the best performing method for the objects studied. Four types of image transformations were studied, namely rotation, scale, tilt and viewpoint changes. Detailed discussions were presented on the experimental setup for these four image transformations and the pre-processing steps required in order to prepare the images to conduct controlled experiments. Based on the evaluation study, it was concluded that the SURF local descriptor method performs best for the objects used as case studies and was subsequently used as the basis for further development.

Chapter 4 presented the first major contribution from this thesis. The two local descriptor formation methods developed are referred to as colour local descriptor and hybrid local descriptor methods, which utilises colour models instead of greyscale images for computing local descriptors. The two methods developed take different approaches in utilising colour information. The colour local descriptor method computes local descriptors directly from the colour models, and the hybrid local descriptor method computes local descriptors from greyscale images, and these local descriptors are then combined with colour patches of the same regions from colour models, resulting in a hybrid method consisting of both area- and feature-based methods. Experimental work showed that the colour local descriptor method is more robust over a wide range of image transformations, with gains in matching accuracy of up to 10%. The hybrid local descriptor method on the other hand is more robust for small magnitudes of image transformations, but does not handle large magnitudes of image transformations well due to the hybrid nature of the method.

In addition to the four image transformations, experiments were also conducted to validate the methods against changes in illumination colour and intensity. It was found that the colour local descriptor method was consistently the best performing method, where improvements of up to 15% were observed. The hybrid local descriptor method had mixed results and did not handle the changes as well, with improvements of up to 10% observed.

An increase in computation time was observed for both the colour local descriptor and hybrid local descriptor methods due to the use of colour models. Because the colour local descriptor method performed better overall, further work was carried out and two feature-reduction methods for the colour local descriptor method were presented. The first method is based on PCA and the second method considers the distribution of the intensity of the pixels in the individual colour channels of the colour model. Both methods managed to reduce the number of vectors from the colour local descriptors and hence reducing the computation time, while minimising the reduction in accuracy to approximately 5%. However, a reduction in matching accuracy was noted and therefore it was concluded that unless the computation time is of great importance, no feature-reduction methods should be applied.

Chapter 5 presented the second piece of major contribution from this thesis, which is the local descriptor matching method utilising SVM. Instead of computing scalar values from the difference vectors of local descriptor pairs, then using these scalar values for determining the correctness of local descriptor matches, the SVM matching method considers all the values of the difference vectors of local descriptor pairs. Experimental work conducted demonstrate that the developed method is more robust compared to the threshold matching method, with gains of up to 20% in matching accuracy. To further demonstrate the robustness and versatility of the method, experiments were conducted, including using different machine learning algorithms in place of SVM and integrating the SVM matching method with different local descriptor methods successfully.

Similar to the local descriptor formation methods, two feature-reduction methods were presented in an attempt to reduce the computation time. The first method utilises PCA, and experimental work showed that it successfully reduced the computation time, with a reduction in accuracy of approximately $5 - 10\%$. The other feature-reduction method presented is the RFE-SVMs method, which was not able to maintain the matching accuracy when the number of features was reduced. This was attributed to the nature of the approach compared to PCA. It was concluded that feature-reduction should not be applied unless the computation time is of great importance due to the reduction in accuracy observed in experiments.

Chapter 6 presented two pieces of work. The first integrated the local descriptor formation methods with the local descriptor matching method. Different degrees of improvement were observed for the different combinations experimented with, which included the two sets of methods and their respective feature-reduction methods. It was

concluded that, for the best matching accuracy, no feature-reduction should be applied, with improvements of up to 25% over existing methods observed. If a balance between matching accuracy and computation time is desired, then feature-reduction should be applied to the SVM matching method, where improvements of approximately 20% were observed. No computation time was presented, as the focus was on the robustness of image matches and as such, the implementations of the methods were not optimised. This meant that it was impossible to provide a fair comparison between the methods.

In addition to the various local descriptor methods developed, an assisted image registration programme was also developed. This programme allows end-users to manually select a set of point pairs from images pairs, the programme then automatically uses these points as starting points to aid in the search for local descriptor correspondences. Two methods, search region and search direction, were discussed which can reduce the search space and, by doing so, reduce the computation time and at the same time the chance of mismatches. It was found through experiments that the search region approach is a more robust and practical approach. Gains in matching accuracy of up to 50% were observed using the search region method, and it was found that the improvement over pure computer vision algorithms improved as the magnitude increased.

In conclusion, this thesis has presented two sets of new methods for different stages of the local descriptor process: local descriptor formation and local descriptor matching. In addition, the assisted image registration programme showed how the robustness of matching images can be improved significantly by utilising the strength of the human eyes. Results from experiments showed that these methods perform better than existing methods for the objects studied in this thesis. It is believed that these methods are of great asset to the registration of images for the purpose of 3D reconstruction, not only for the Māori artefacts used as case studies in this research, but also other applications.

7.2 Contributions

Below is a summary of the contributions made throughout the course of this research. This, in conjunction with the summary of the results from experiments in the previous section, show the importance of the research conducted.

Performance Evaluation of Local Descriptor Methods An evaluation study was carried out, which compared the performance of four local descriptor methods: (a) SIFT; (b) PCA-SIFT; (c) GLOH; and (d) SURF. Four image transformations were studied: (a) rotation; (b) scale; (c) tilt; and (d) viewpoint changes. Experimental results show that SURF performed best for all the transformations studied. This study was important as it provided a detailed understanding of the capabilities of existing algorithms. Without this understanding,

it is difficult to improve on existing algorithms, since the issues that exist cannot be fully understood. Because of the importance of this study for the objects studied, it is considered an important contribution towards improving local descriptor methods.

Colour and Hybrid Local Descriptor Methods Two new local descriptor methods, colour local descriptor and hybrid local descriptor methods, were proposed. These two methods compute local descriptors from colour models instead of greyscale images in order to improve the uniqueness of the local descriptors, and to provide invariance against illumination condition changes. Results from experiments show that improvements of up to 10% over existing methods were observed. Because the focus of development for these two methods were placed on computing more unique local descriptors, they are not limited to the artefacts used as case studies, and can be applied to a wide range of images and applications. These methods are important contributions, as they are significant improvements over existing local descriptor methods based on the results from experiments. They provide good foundations for further work in utilising colour images or models for image registration algorithms.

Local Descriptor Matching with Support Vector Machines A new local descriptor method based on SVM was proposed. By using SVM, it was possible to take into account all the vectors in difference vectors of local descriptor pairs, therefore overcoming the issue in existing methods where metric distance measures are used. These methods, for example the threshold matching method, compute scalar values from the difference vectors, and therefore a large proportion of the useful data are lost. By taking into account all the individual vectors, a better method for matching local descriptors resulted. Results from experiments show that improvements of up to 20% over existing methods were observed. This work has demonstrated, for the first time in literature, how SVM can be integrated with local descriptors. Experimental results show this is a major improvement over the use of methods based on metric distance measures, for example the threshold matching method. This is an important contribution as this covered the least studied stage of the local descriptor process in literature, and is a robust method that can be integrated with all existing, as well as future local descriptor formation methods.

Assisted Image Registration An assisted image registration programme was developed, which allows for a semi-automatic approach for the task. The programme requires the manual selection of three or more point pairs from image pairs, and uses these point pairs as starting points to automatically register the images. The colour and hybrid local descriptor methods, and the SVM matching method developed in the course of this research are used for matching the images. Two methods were developed to reduce the search space

when matching local descriptors, using the selected point pairs as starting points. These methods allow for a reduced computation time, and improved accuracy due to the reduced number of potential matches for each local descriptor. Results from experiments show that improvements of up to 50% over pure computer vision methods were observed. This tool is easy to use, and provides a very effective method of registering images when automatic registration of images is not required. The advantages of a faster computation time and more robust image matches out-weigh the need to manually select point pairs from images.

Māori Artefacts Four Māori artefacts were used as case studies in this research for two important reasons. First, many artefacts are difficult to register through a combination of feature-rich regions and regions that lack distinct features, and the repetitiveness of features found on many artefacts. This posed additional challenges and meant that more robust methods and algorithms needed to be developed in order to register images of these objects. Second, there is currently a need to construct 3D models of these artefacts. Limitations in the current approach meant that an alternative approach was required. This is an important contribution because this research has provided a method for registering images of these artefacts, and the registered images can be used for constructing 3D models in future research.

7.3 Future Work

This thesis has demonstrated two sets of methods for improving the robustness of the registration of images for the purpose of 3D reconstruction, using four Māori artefacts as case studies. While effort has been put into developing, improving and verifying the methods presented, like any research, the work presented does not mark the end of development of local descriptor processes. Based on the results and discussion from previous chapters, a list of future work has been identified and is discussed in detail.

7.3.1 Accuracy of Image Registration

This thesis discussed the accuracy of the matching of local descriptors, however the accuracy of the matching of images, using these local descriptors, was not discussed. While the accuracy of the matching of images is largely dependent on the accuracy of the local descriptors, the method used for computing the accuracy of the matching of images also plays an important role. The RANSAC algorithm was discussed in this thesis, however due to the vast number of other methods available [230], a comprehensive study which compares these methods should be carried out to determine the most appropriate method. As the comparison of the different matching methods is a research area in itself, this was

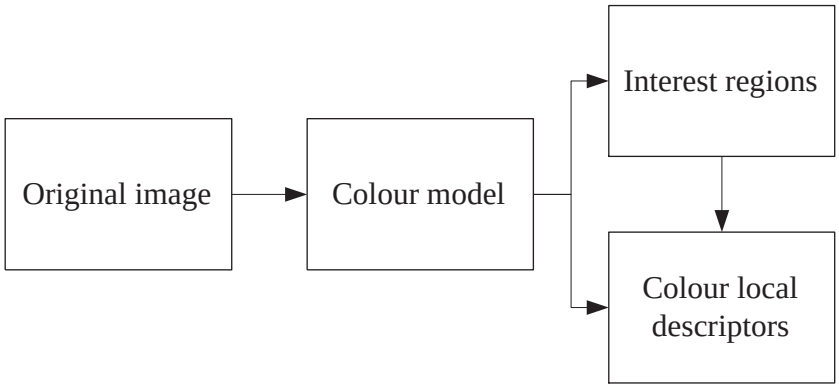


Figure 7.1: Overview of the colour local descriptor method which computes the interest regions using the same colour model used to construct the local descriptors.

considered to be outside of the scope of this thesis.

Even though this work is outside of the scope of this thesis, results from a recent study by Choi *et al.* [230] show that similar accuracies for the matching of images could be obtained for matching accuracies ranging from 30% – 90% for local descriptors. The only difference was the number of iterations, and therefore the computation time, required, with more iterations required as the matching accuracy of local descriptors reduced. These results support the results presented in this thesis, as the majority of the results had matching accuracies above 30%. This means that, in theory, the matched local descriptors presented in this thesis are sufficient in successfully matching the images used.

7.3.2 Future Development of Algorithms

Significant contributions have been made to two stages of the local descriptor process in this research and improvements in matching accuracy were observed in the results presented in Chapters 4-6. This, however, does not mean that the development of local descriptors processes stops here, and based on the work presented in this thesis, future work for both local descriptor formation and local descriptor matching methods have been identified. For the colour local descriptor and hybrid local descriptor methods, one of the drawbacks is that the interest regions are computed from greyscale images. While using the Harris-Laplace detector meant that the comparison between conventional methods and the methods presented in this thesis was simplified as it allowed for controlled experiments to be conducted by keeping all the variables constant, it also means that the presented methods are not truly based on colour images, since the computation of the interest regions relies on region detectors detecting a set of interest regions from greyscale images. It is therefore necessary to find out how the use of colour images or colour models would complement the two local descriptor methods presented in Chapter 4, particularly if the interest regions are computed from the same colour model which the local descriptors are computed from.

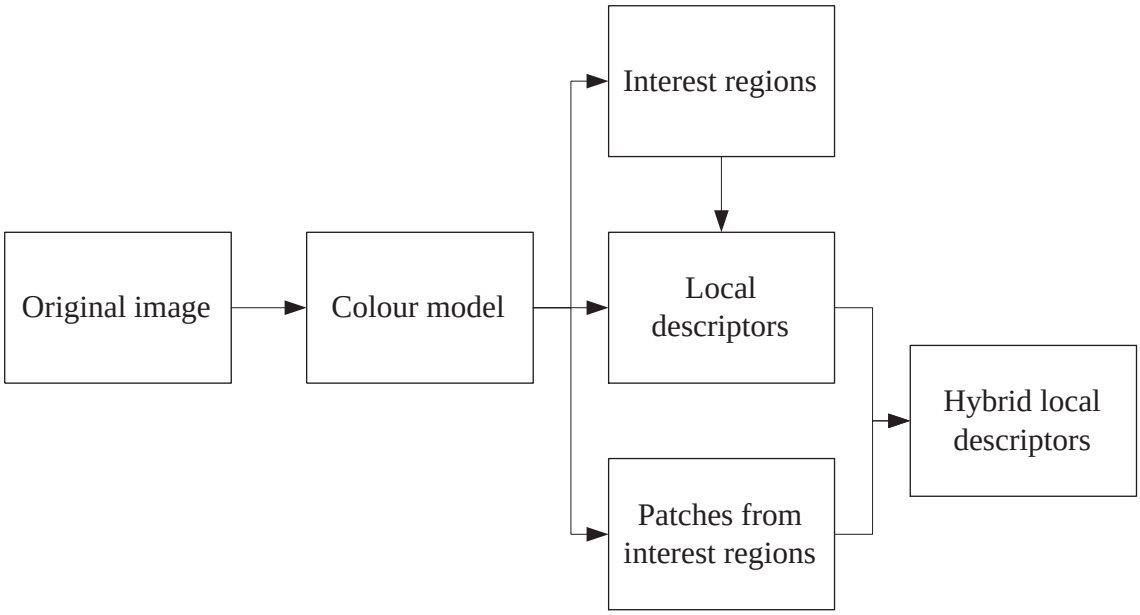


Figure 7.2: Overview of the hybrid local descriptor method which computes the interest regions using the same colour model used to construct the local descriptors and colour patches.

Overviews of the colour local descriptor and hybrid local descriptor methods using colour models to compute the interest regions are shown in Figures 7.1 and 7.2.

Another drawback for both the colour local descriptor and hybrid local descriptor methods, as well as the SVM local descriptor matching method, is the increase in computation time. This was of no concern for this thesis due to the nature of the application being an offline process, however, this property makes these methods unsuitable for real-time and embedded applications. While feature-reduction methods were presented which can be used to reduce the computation time required, the matching accuracy is reduced when these methods are applied. Even though it is widely accepted that it is difficult to have both good computation time and robustness, depending on the application involved, it would be beneficial to further study if the computation time of the methods presented can be reduced. Also, for both sets of methods, no *a priori* information is used. This approach was taken since by not considering *a priori* information, it was possible to develop methods which are generic and can be applied to a wide variety of applications. A downside to this approach is that over the years, it has been proven that while application-specific methods restricted their use in other areas, the performance can often be further improved as the algorithms are designed specifically to deal with the task involved.

Based on this understanding, a study should be carried out to examine how *a priori* knowledge can be integrated with the methods presented, whether it is for the application discussed in this thesis or other applications which may involves images of a completely different nature than those used for experimental work in this thesis.

7.3.3 Implementation

The second part of future work involves the implementation of the algorithms. The local descriptor formation and local descriptor matching, as well as the user-assisted image registration work have all been implemented in MATLAB for its ability to implement algorithms in a quick and efficient manner. However, a downside to MATLAB is its runtime speed [231]. As MATLAB is a computing environment and does not compile the codes developed and instead, the codes are interpreted each time they are executed, the time to carry out the same task is significantly longer than third-generation programming languages such as C++. Another drawback is that the codes developed in MATLAB can only be executed on machines with MATLAB installed or with MATLAB Component Runtime. Implementing the algorithms in MATLAB was ample for this thesis as the aim was to verify that the developed methods are indeed more robust and not to compare the computation time of various methods. From an application point of view however, ideally the algorithms should be implemented in third-generation programming languages to maximise the performance as well as portability of the programmes. This task was not carried out as the implementation of these algorithms in third-generation programming languages is very time-consuming.

7.3.4 Experimental Work on Other Objects

Extensive experimental work was carried out to verify the performance of the methods presented, and the results suggest that the methods developed perform well for the objects studied. Further experimental work should, however, be carried out to validate the performance of these methods for other objects. The aim of this is to determine the performance of these methods for different applications. The experimental setup presented in Chapter 3 provides a solid foundation for future evaluation study using the local descriptor methods.

7.3.5 Integration of Image Registration with 3D Reconstruction

This thesis focused on the task of registering images. The next step is to construct 3D models of the artefacts studied. In order to fulfill this requirement, the results from this thesis need to be integrated with 3D reconstruction algorithms to determine how well state-of-the-art 3D reconstruction algorithms perform for these objects. This may seem like a simple task at first, as both the intrinsic and extrinsic parameters required for 3D reconstruction algorithms are available for all the images used for experimental work in this research. The difficulty involved, however, lies in the integration and implementation of the registration and reconstruction algorithms.

Appendix A

Images for Experimental Work

This appendix presents some of the images used for the experimental work conducted for this thesis. This is by no means a complete set of images used and instead, is only a representative subset of the images used.

(a)

(b)

Figure A.1: Images of the flute artefact used for the experimental work for this thesis with different types of image transformations: (a) tilt; and (b) viewpoint changes.

Appendix B

Additional Results

This appendix presents the results from Chapters 3, 4, 5 and 6 which do not fit in the main text. The format of this appendix follows closely of the chapters listed for ease of comparison.

B.1 Colour and Hybrid Local Descriptor Methods

Results for Chapter 4: Colour and Hybrid Local Descriptors Methods.

B.1.1 Local Descriptors Based on Greyscale versus Colour Images

B.1.2 Feature-Reduced Colour Local Descriptors

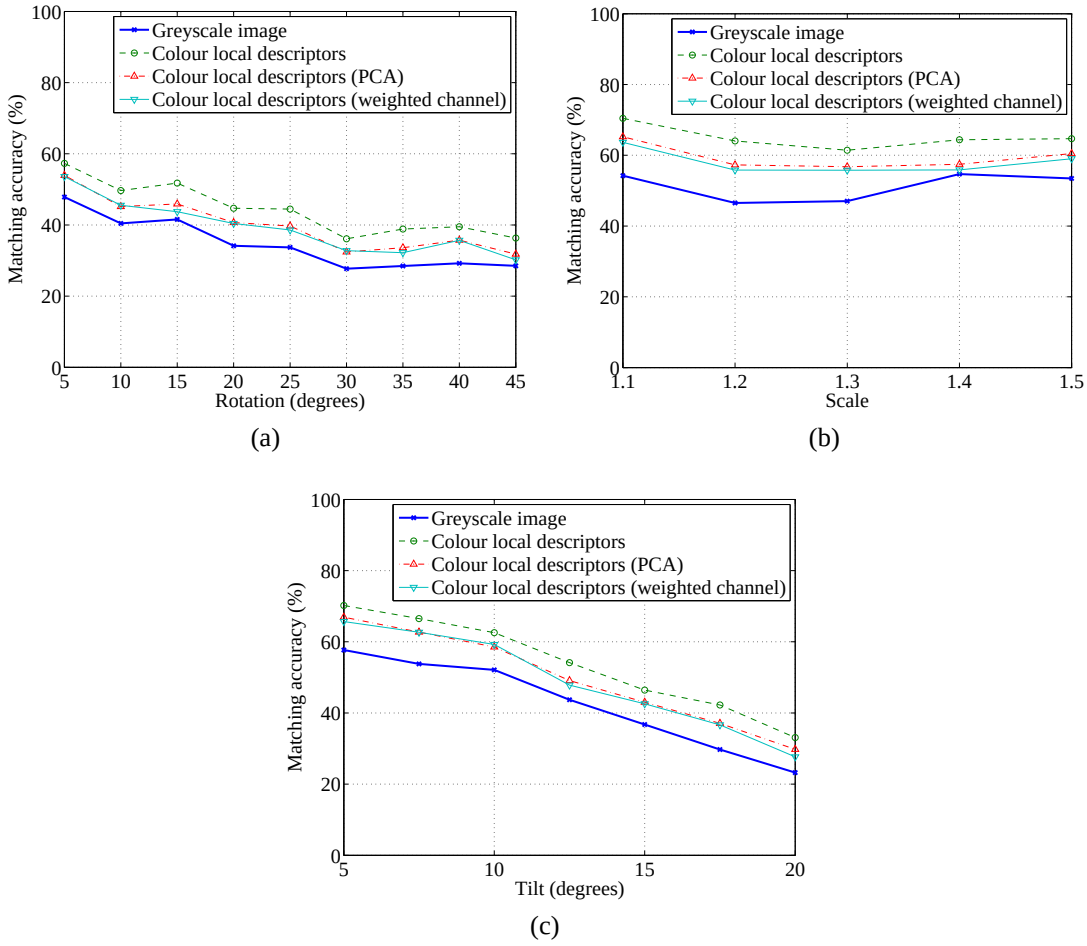


Figure B.1: Image matching results using four different local descriptor methods computed from different types of images for the patu and wahaika artefacts with different image transformations: (a) rotation; (b) scale; and (c) tilt.

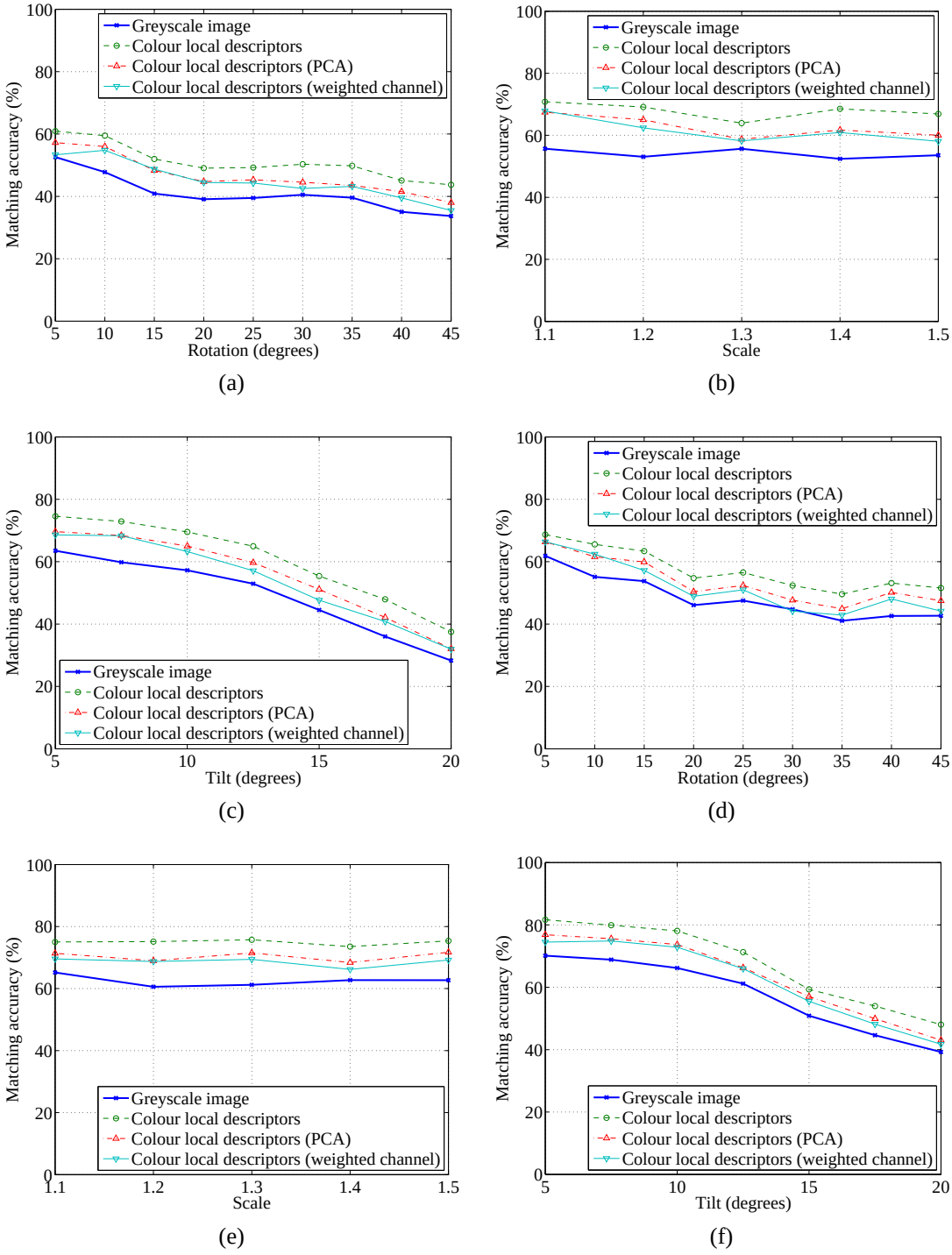


Figure B.2: Image matching results using four different local descriptor methods computed from different types of images for the wahaika and tiki artefacts with different image transformations: (a) rotation (wahaika); (b) scale (wahaika); (c) tilt (wahaika); (d) rotation (tiki); (e) scale (tiki); and (f) tilt (tiki).

B.1.3 Illumination Changes

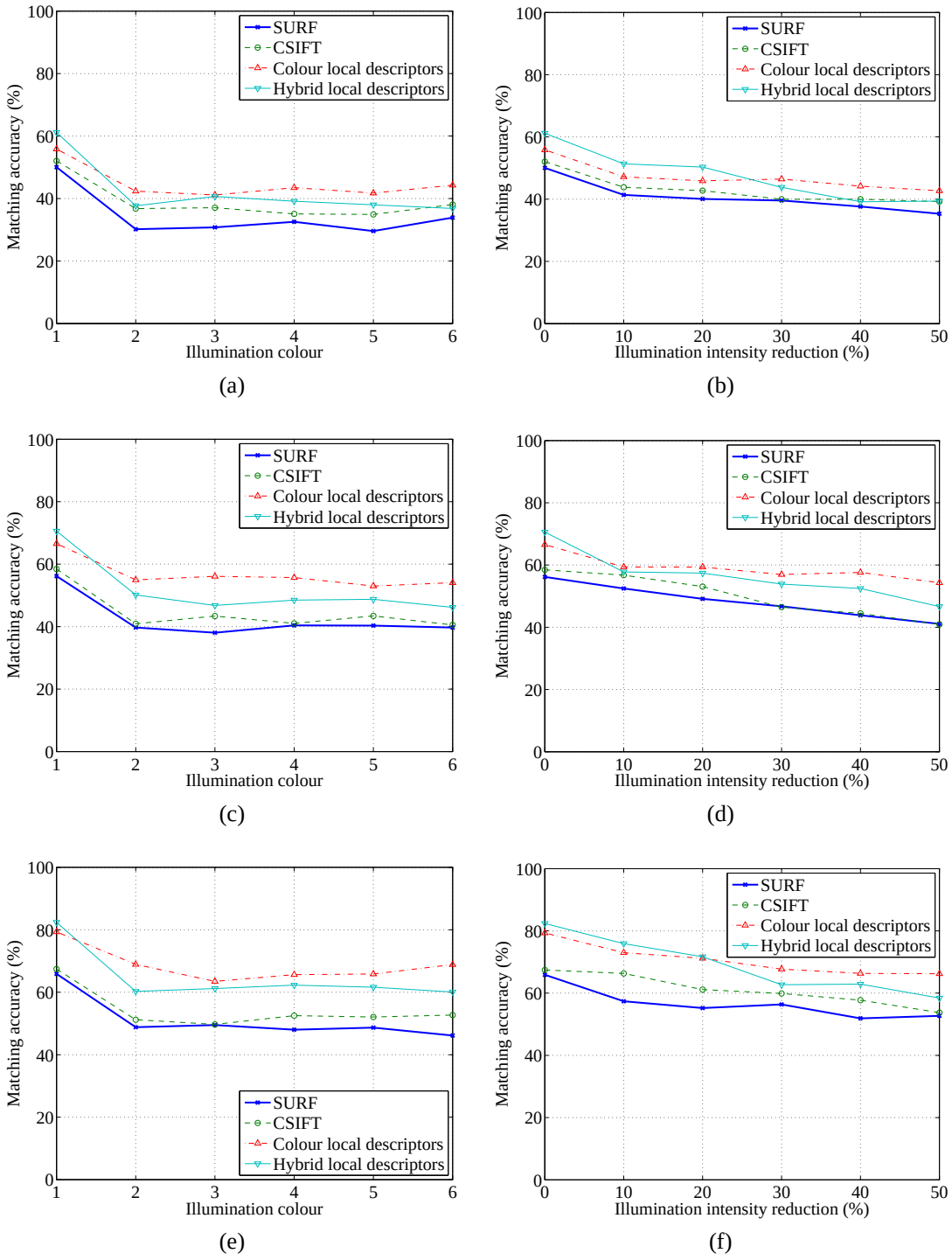


Figure B.3: Image matching results using four different local descriptor methods for the patu, wahaika and tiki artefacts with different illumination changes: (a) colour (patu); (b) intensity (patu); (c) colour (wahaika); (d) intensity (wahaika); (e) colour (tiki); and (f) intensity (tiki).

B.1.4 Hybrid Local Descriptors

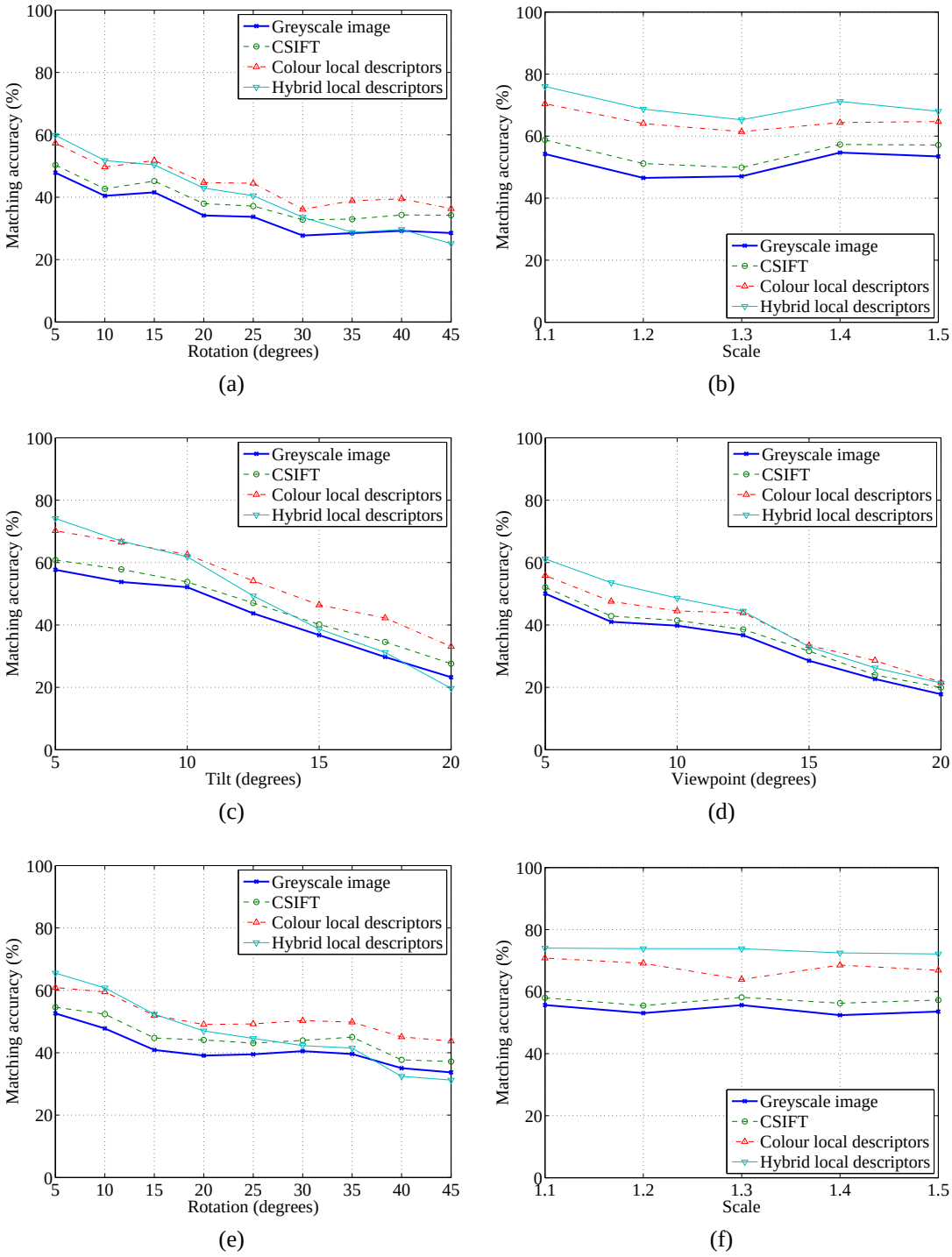


Figure B.4: Image matching results using four different local descriptor methods for the patu and wahaika artefacts with different image transformations: (a) rotation (patu); (b) scale (patu); (c) tilt (patu); (d) viewpoint (patu); (e) rotation (wahaika); and (f) scale (wahaika).

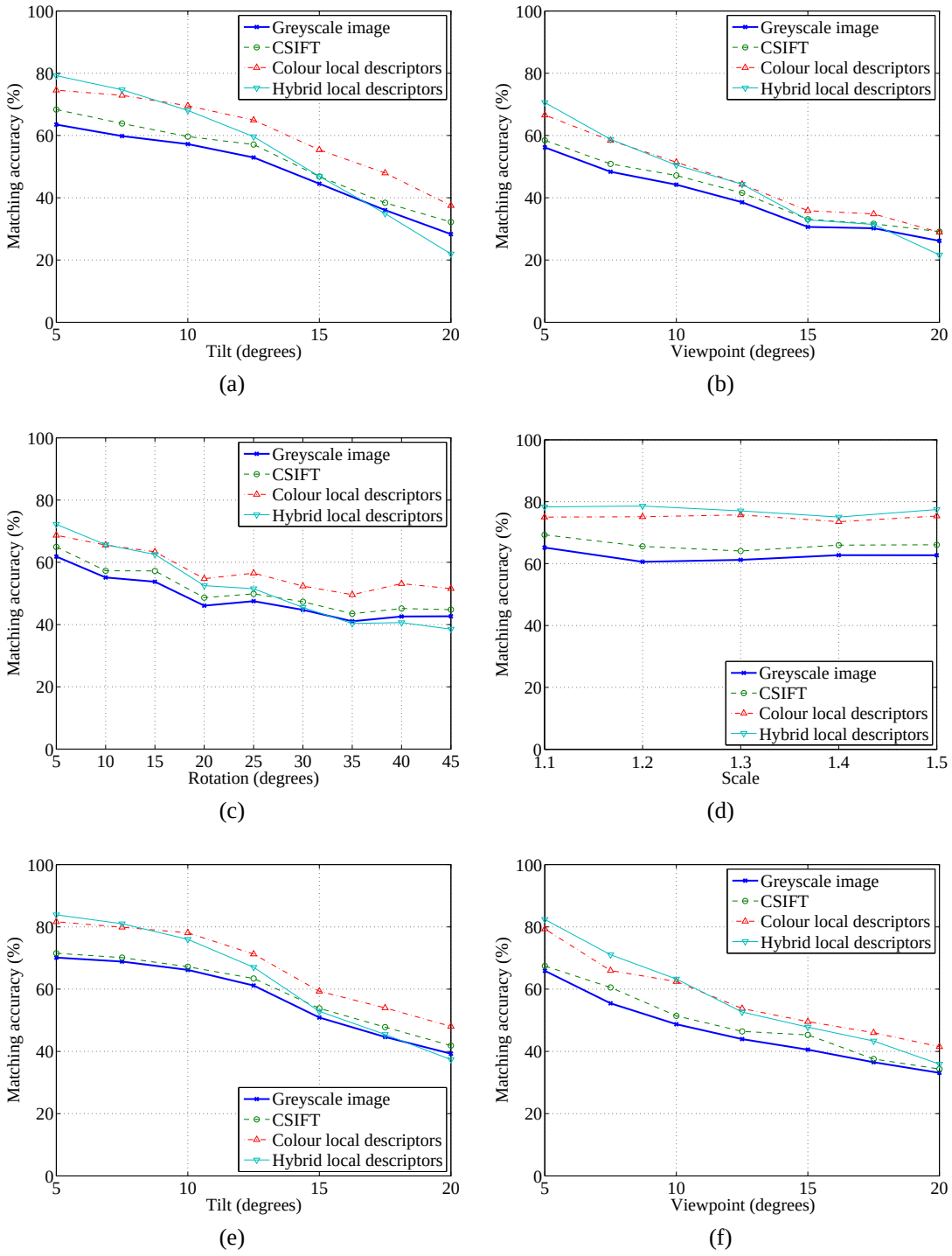


Figure B.5: Image matching results using four different local descriptor methods for the wahaika and tiki artefacts with different image transformations: (a) tilt (wahaika); (b) viewpoint (wahaika); (c) rotation (tiki); (d) scale (tiki); (e) tilt (tiki); and viewpoint (tiki).

B.1.5 SURF versus SIFT

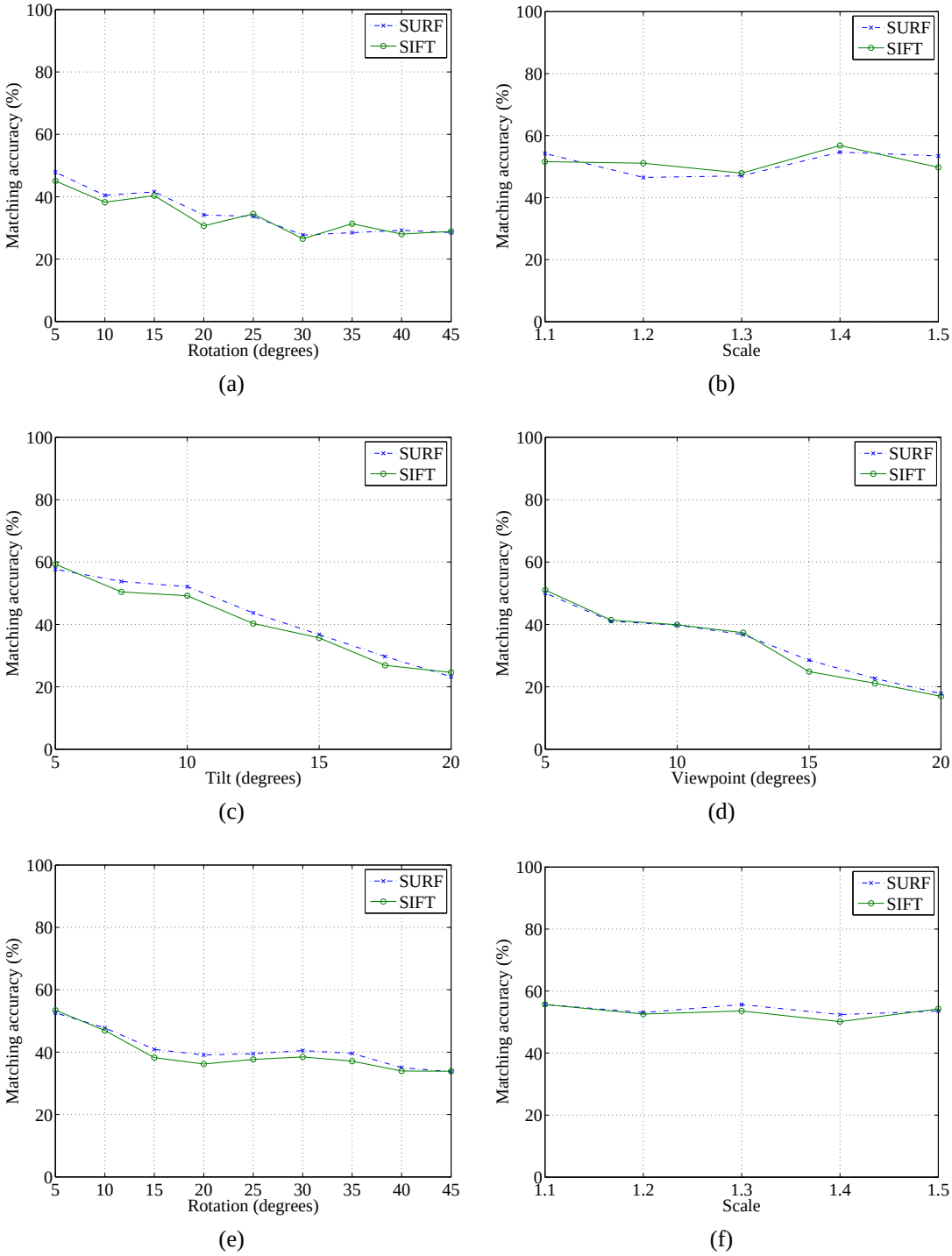


Figure B.6: Image matching results using colour local descriptors computed from two different local descriptor methods for the patu and wahaika artefacts with different image transformations: (a) rotation (patu); (b) scale (patu); (c) tilt (patu); (d) viewpoint (patu); (e) rotation (wahaika); and (f) scale (wahaika).

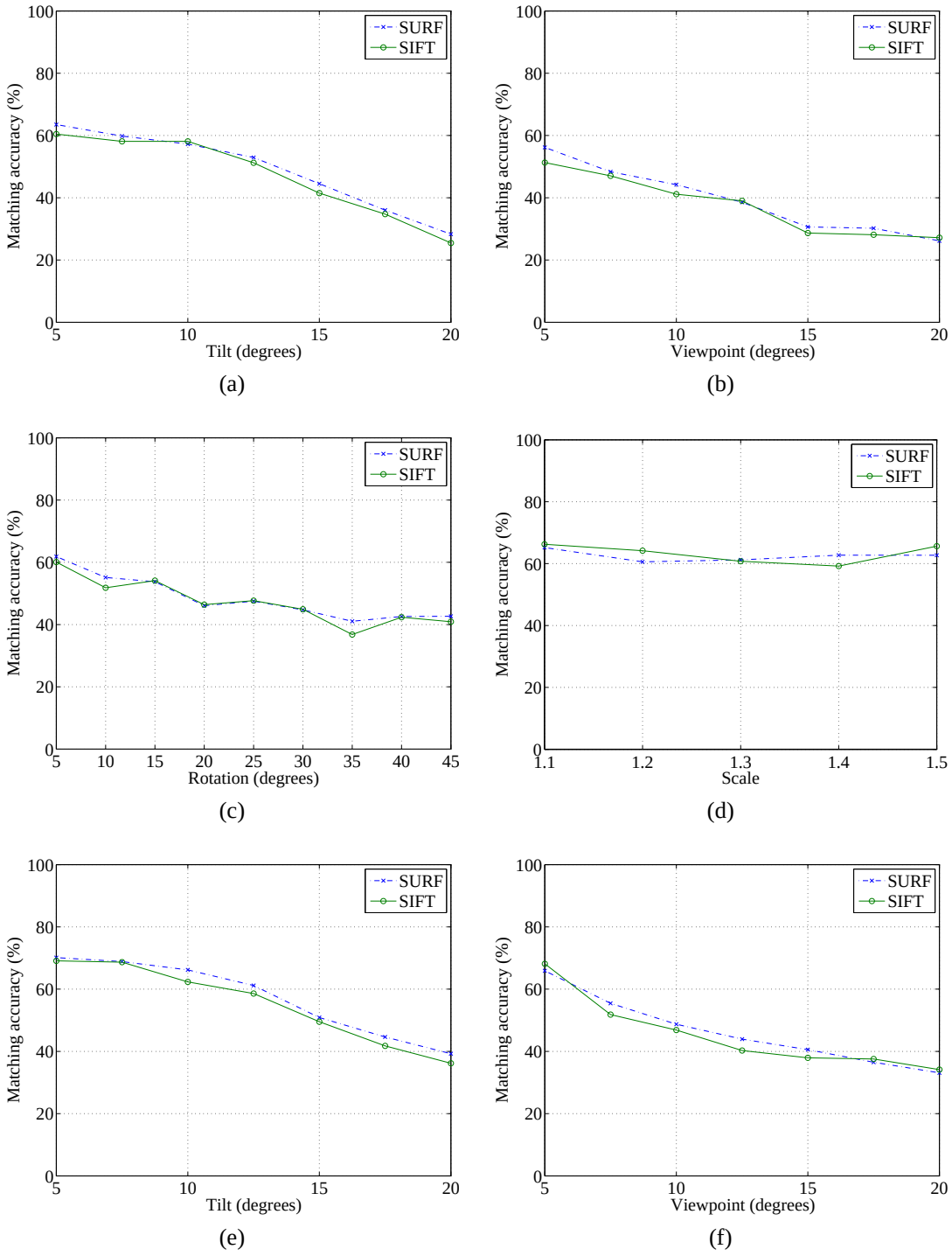
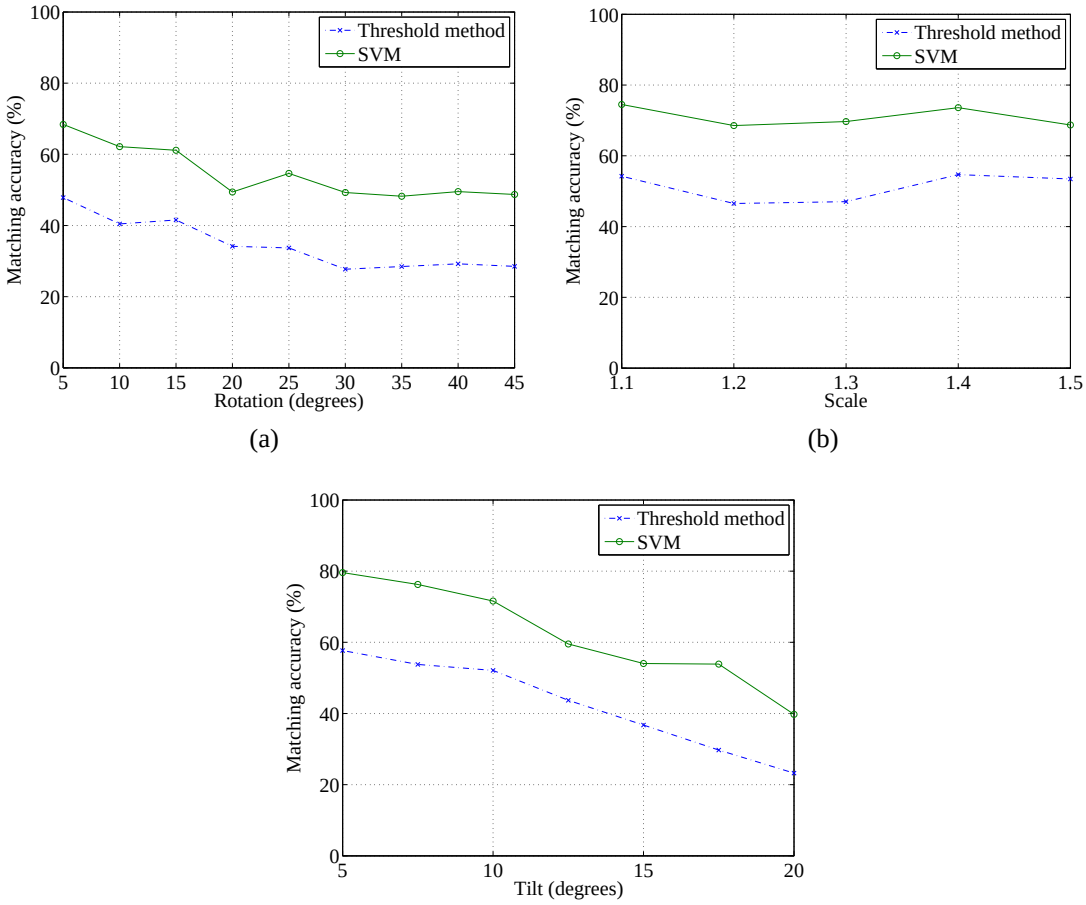


Figure B.7: Image matching results using colour local descriptors computed from two different local descriptor methods for the wahaika and tiki artefacts with different image transformations: (a) tilt (wahaika); (b) viewpoint (wahaika); (c) tilt (tiki); (d) viewpoint (tiki); (e) rotation (tiki); and (f) scale (tiki).

B.2 Local Descriptor Matching with Support Vector Machines

Results for Chapter 5: Local Descriptor Matching with Support Vector Machines.

B.2.1 Euclidean Distance-Based Methods versus Support Vector Machines



(a) (b) (c)

Figure B.8: Image matching results using three different matching methods for the patu artefact with different image transformations: (a) rotation; (b) scale; and (c) tilt.

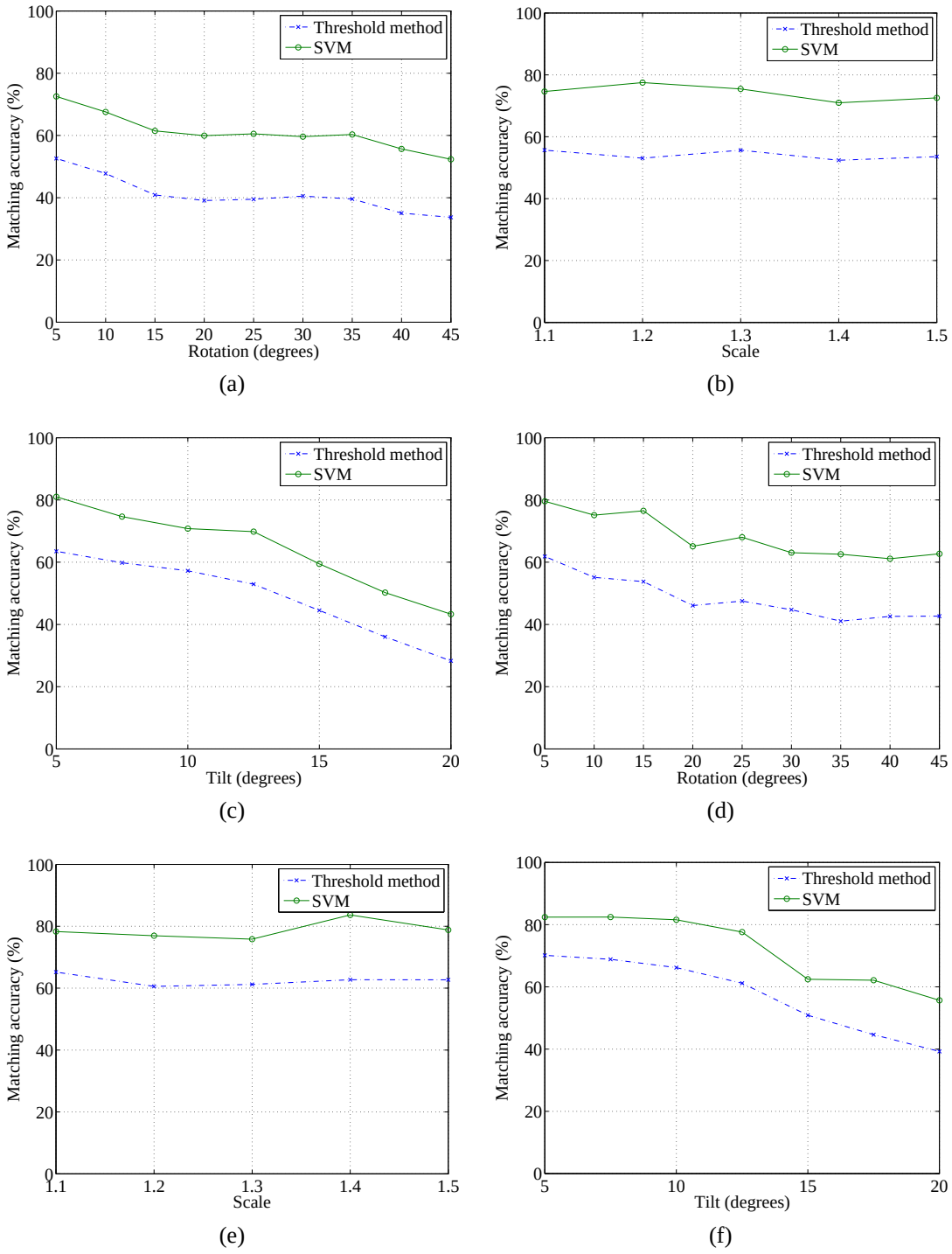
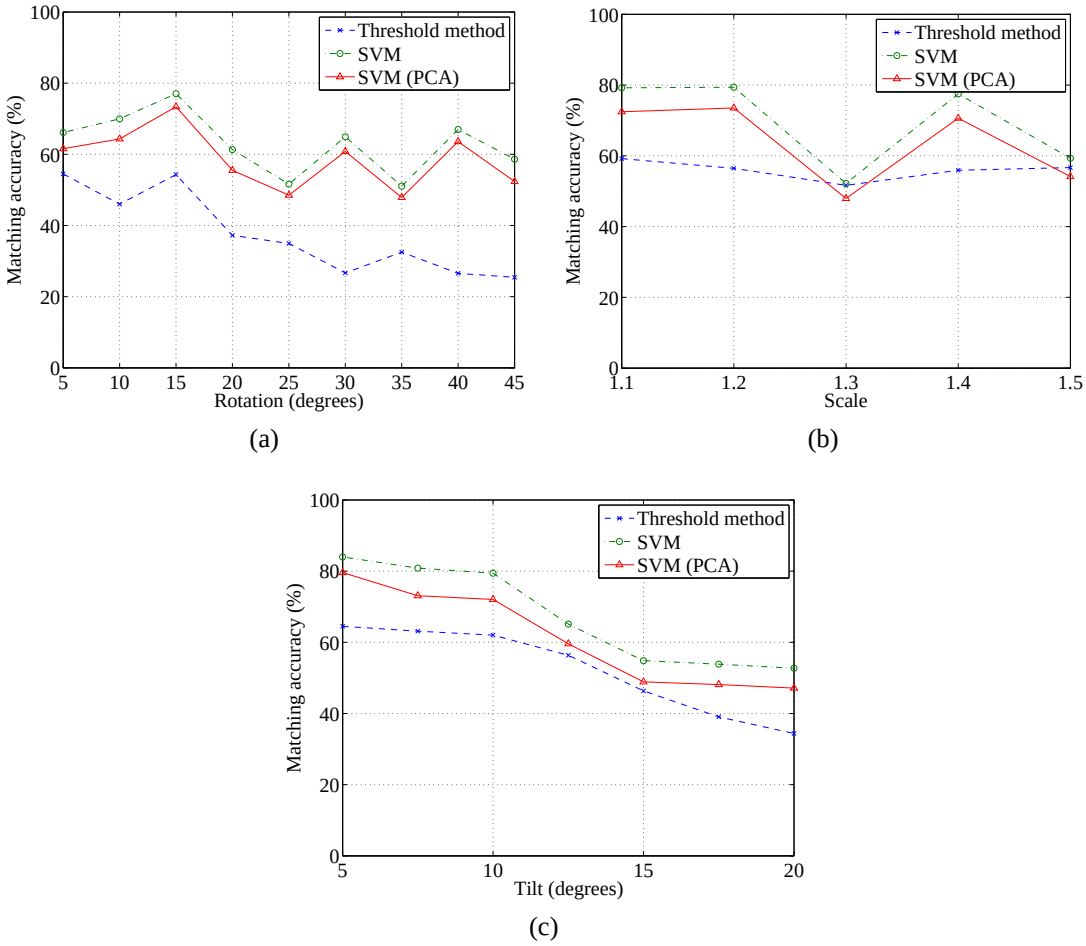


Figure B.9: Image matching results using three different matching methods for the wahaika and tiki artefacts with different image transformations: (a) rotation (wahaika); (b) scale (wahaika); (c) tilt (wahaika); (d) rotation (tiki); (e) scale (tiki); and (f) tilt (tiki).

B.2.2 Feature-Reduced Support Vector Machines



(a) (b) (c)

Figure B.10: Image matching results using three different matching methods for the flute artefact with different image transformations: (a) rotation; (b) scale; and (c) tilt.

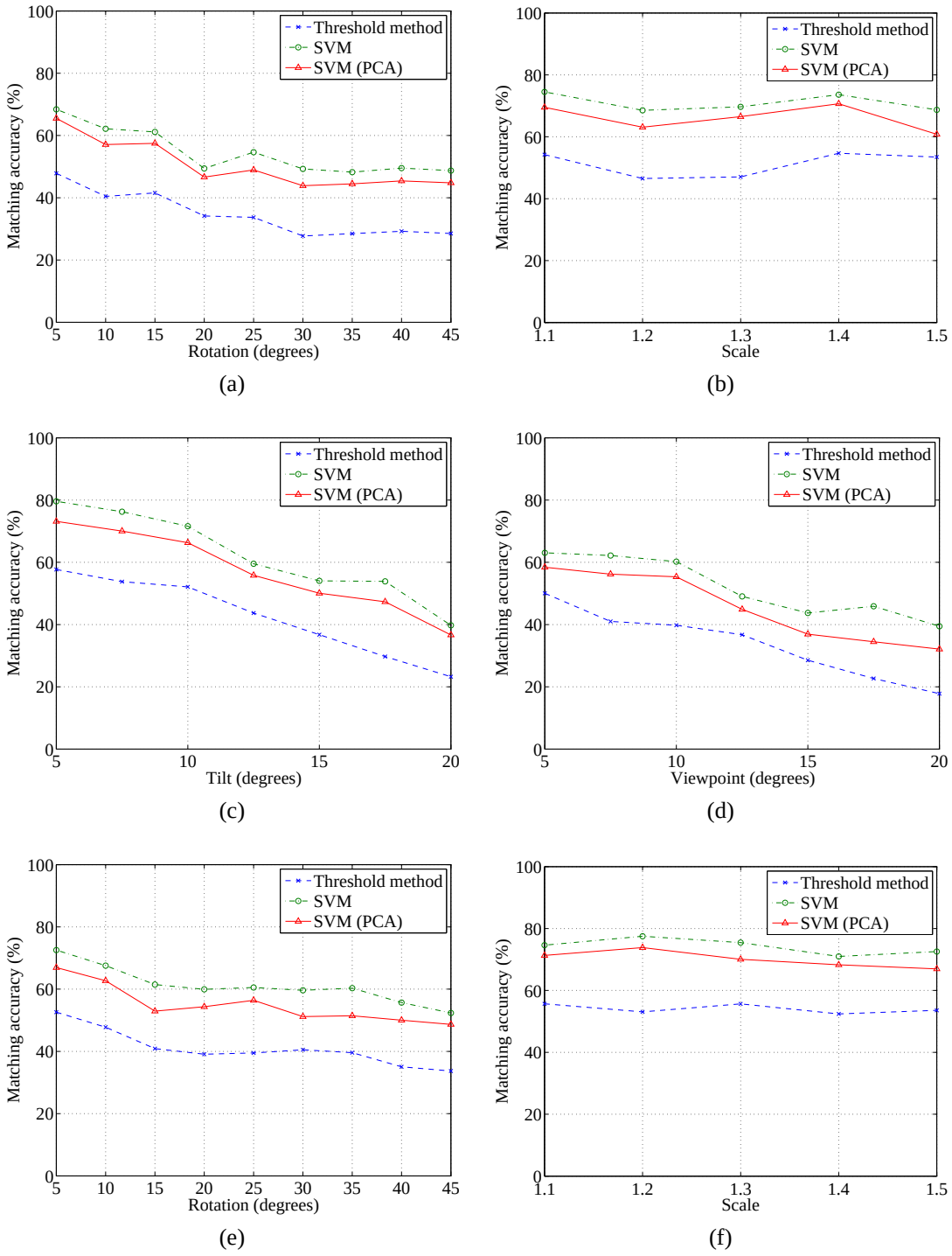


Figure B.11: Image matching results using three different matching methods for the patu and wahaika artefacts with different image transformations: (a) rotation (patu); (b) scale (patu); (c) tilt (patu); (d) viewpoint (patu); (e) rotation (wahaika); and (f) scale (wahaika).

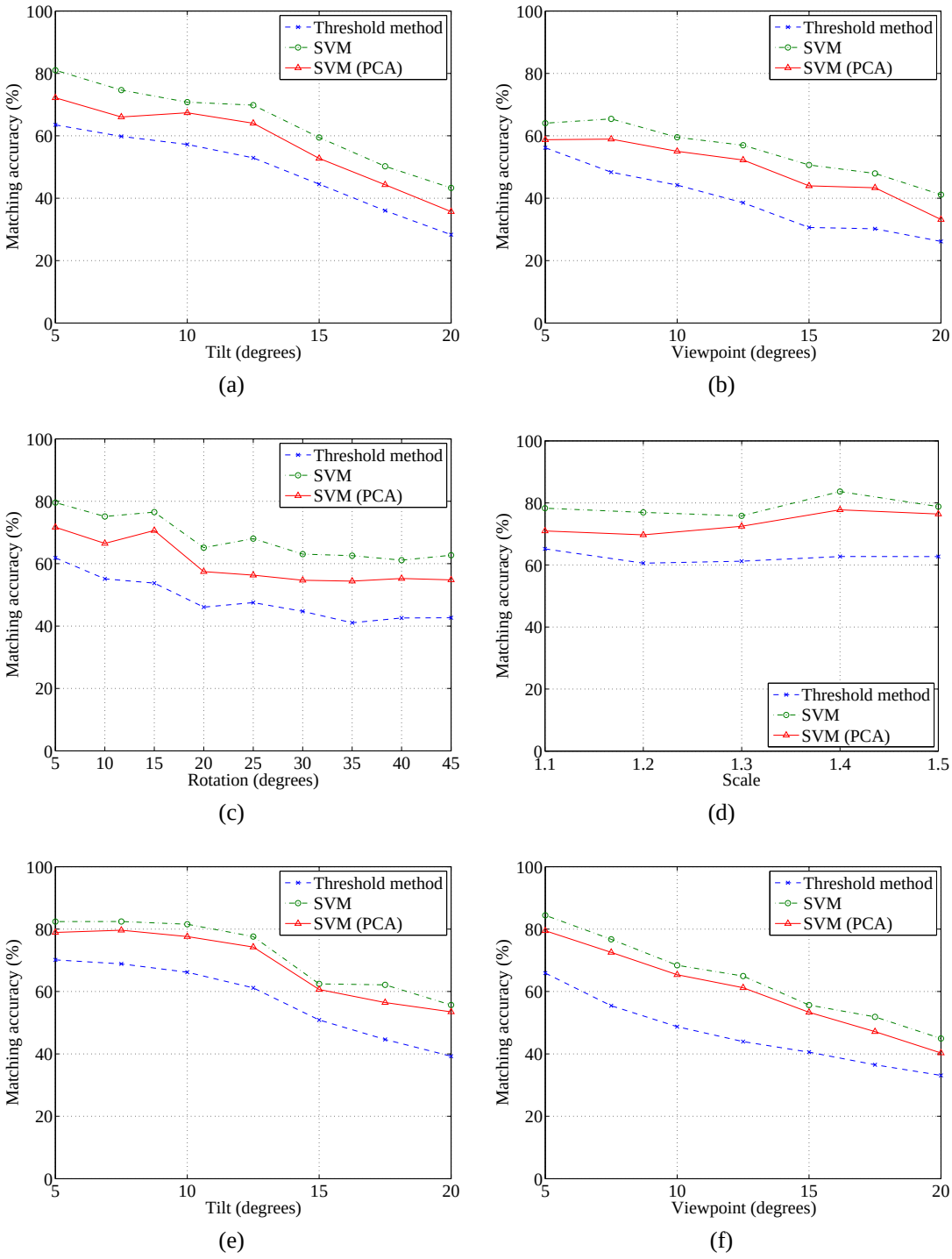


Figure B.12: Image matching results using three different matching methods for the wahaika and tiki artefacts with different image transformations: (a) tilt (wahaika); (b) viewpoint (wahaika); (c) rotation (tiki); (d) scale (tiki); (e) tilt (tiki); and (f) viewpoint (tiki).

B.2.3 Versatility of Machine Learning Algorithms

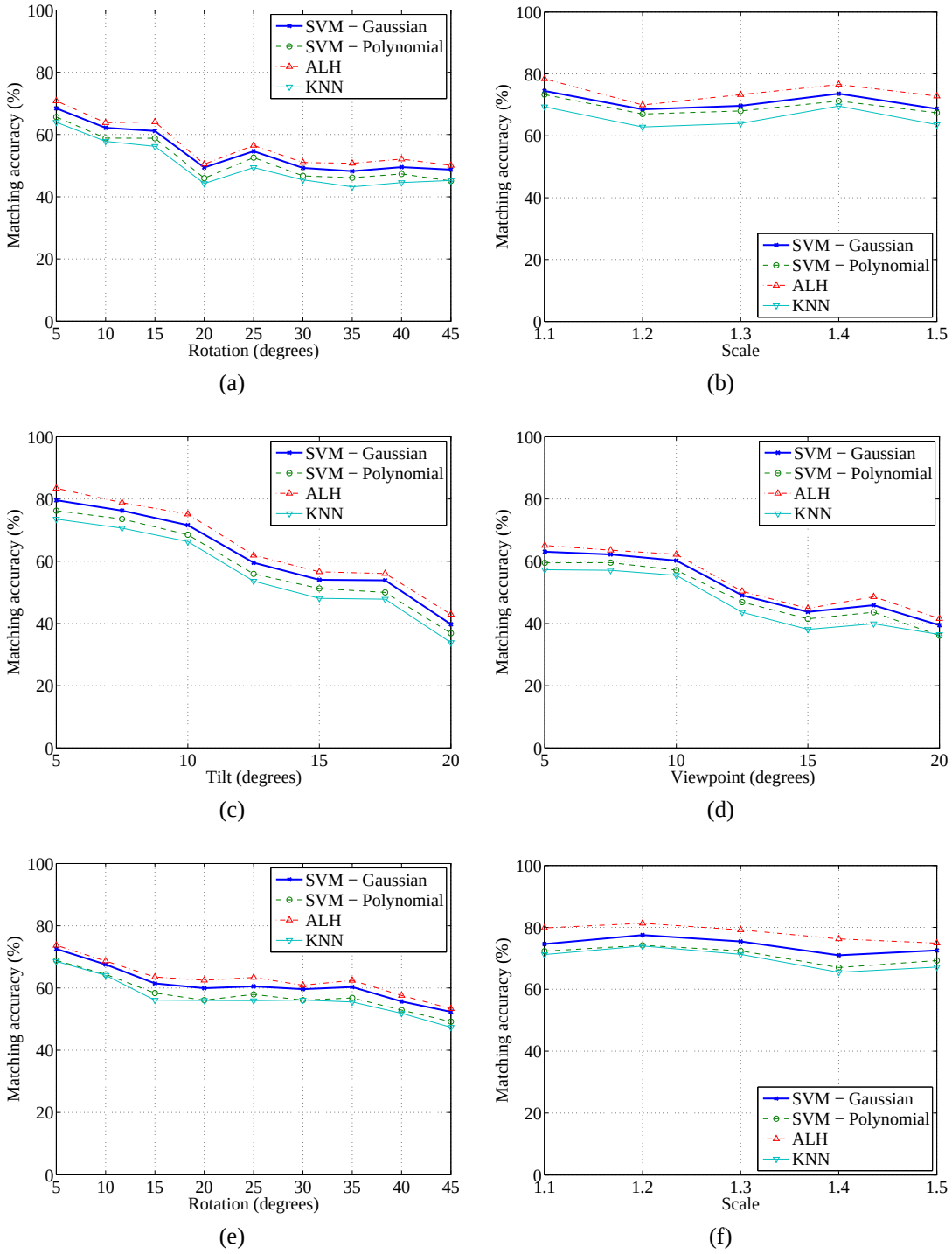


Figure B.13: Image matching results using four different machine learning algorithms for the patu and wahaika artefacts with different image transformations: (a) rotation (patu); (b) scale (patu); (c) tilt (patu); (d) viewpoint (patu); (e) rotation (wahaika); and (f) scale (wahaika).

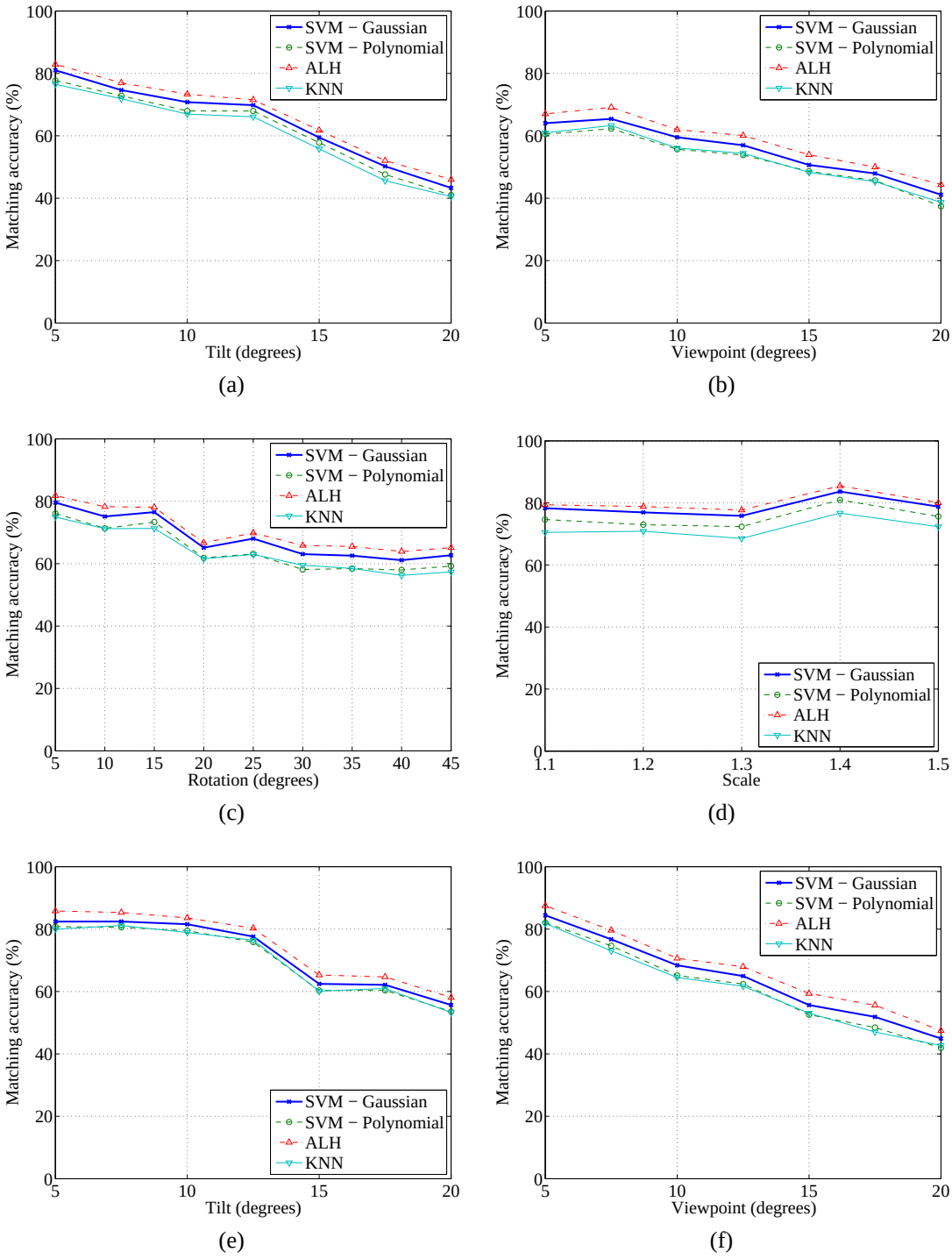


Figure B.14: Image matching results using four different machine learning algorithms for the patu and wahaika artefacts with different image transformations: (a) tilt (wahaika); (b) viewpoint (wahaika); (c) rotation (tiki); (d) scale (tiki); (e) tilt (tiki); and (f) viewpoint (tiki).

B.2.4 SURF versus SIFT

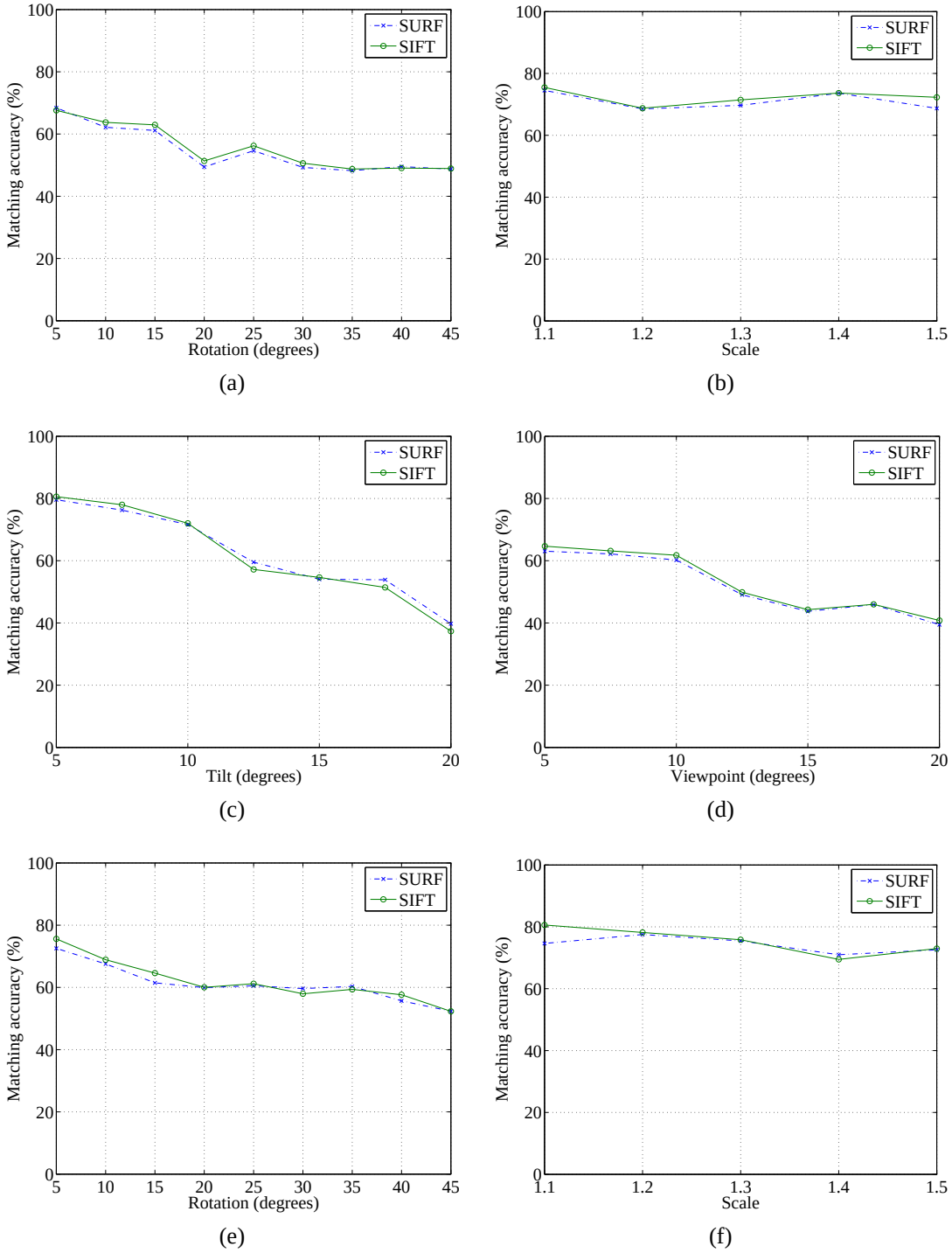


Figure B.15: Image matching results using two different local descriptor methods combined with the SVM matching method for the patu and wahaika artefacts with different image transformations: (a) rotation (patu); (b) scale (patu); (c) tilt (patu); (d) viewpoint (patu); (e) rotation (wahaika); and (f) scale (wahaika).

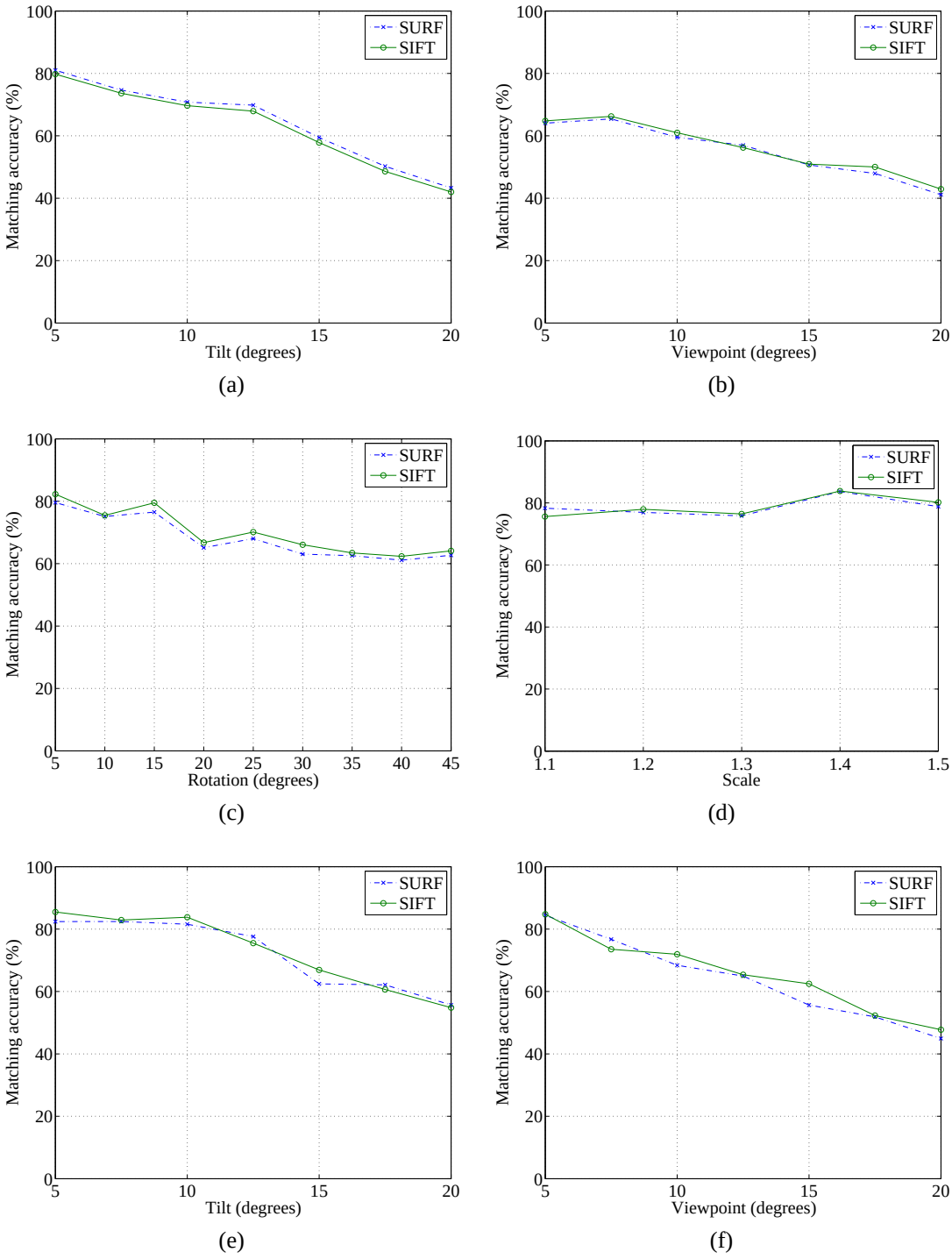


Figure B.16: Image matching results using two different local descriptor methods combined with the SVM matching method for the wahaika and tiki artefacts with different image transformations: (a) tilt (wahaika); (b) viewpoint (wahaika); (c) rotation (tiki); (d) scale (tiki); (e) tilt (tiki); and (f) viewpoint (tiki).

B.3 Integration of Local Descriptor Methods and Assisted Image Registration

Results for Chapter 6: Integration of Local Descriptor Methods and Assisted Image Registration.

B.3.1 Test 1: Local Descriptor Formation Methods Combined with SVM Matching Method

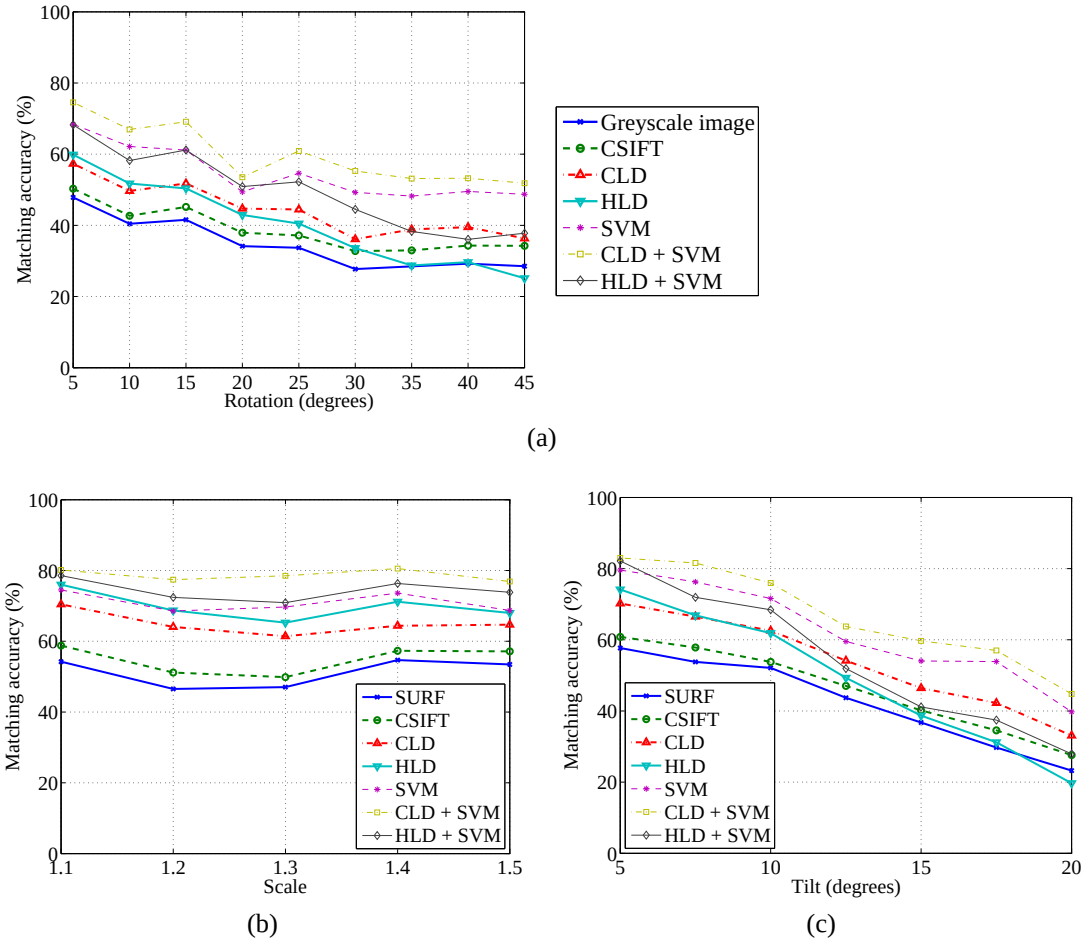
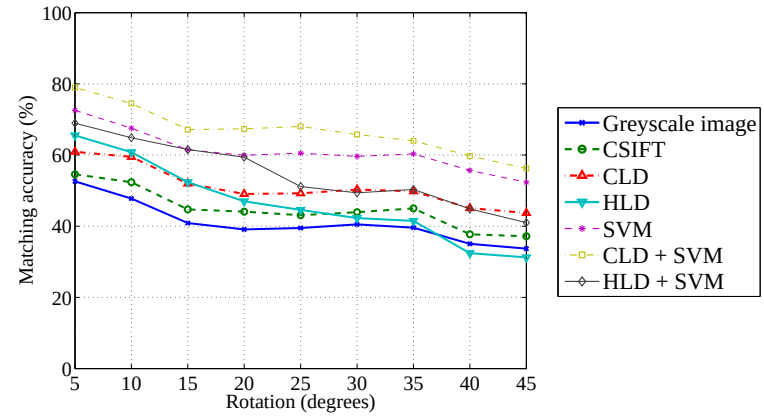
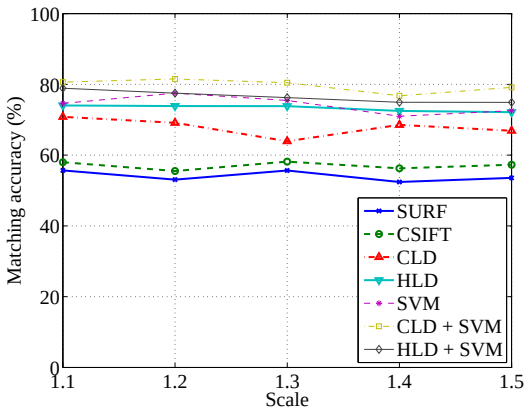


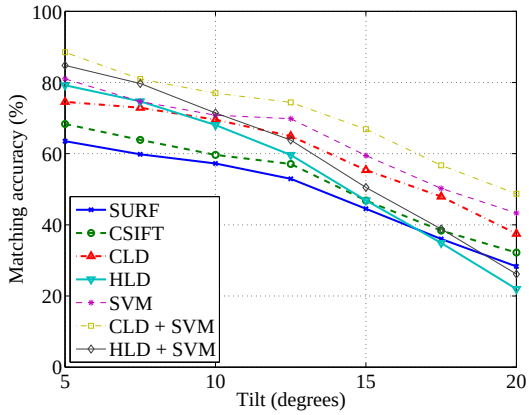
Figure B.17: Image matching results using the colour local descriptor and hybrid local descriptor methods combined with the SVM matching method for the patu artefact with different image transformations: (a) rotation; (b) scale; and (c) tilt.



(a)



(b)



(c)

Figure B.18: Image matching results using the colour local descriptor and hybrid local descriptor methods combined with the SVM matching method for the wahaika artefact with different image transformations: (a) rotation; (b) scale; and (c) tilt.

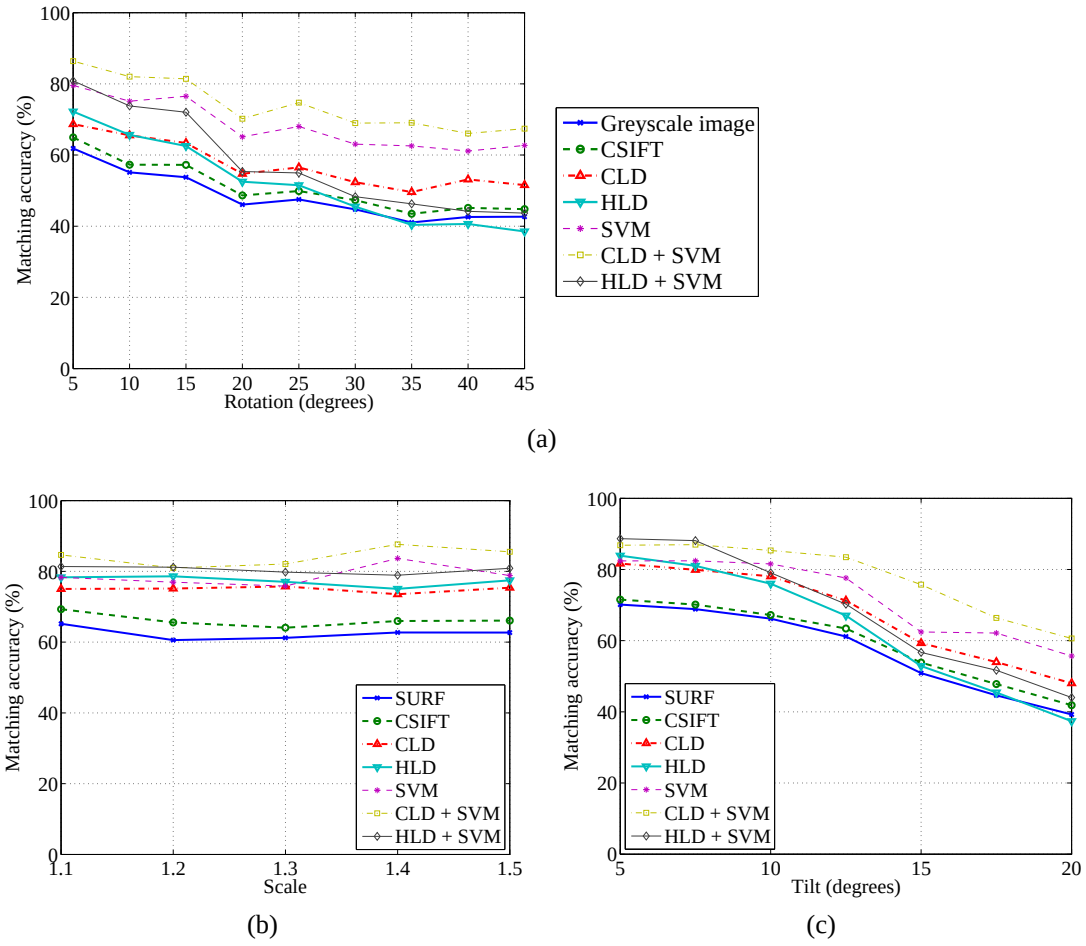


Figure B.19: Image matching results using the colour local descriptor and hybrid local descriptor methods combined with the SVM matching method for the tiki artefact with different image transformations: (a) rotation; (b) scale; and (c) tilt.

B.3.2 Test 2: Local Descriptor Formation Methods Combined With Feature-Reduced SVM Matching Method

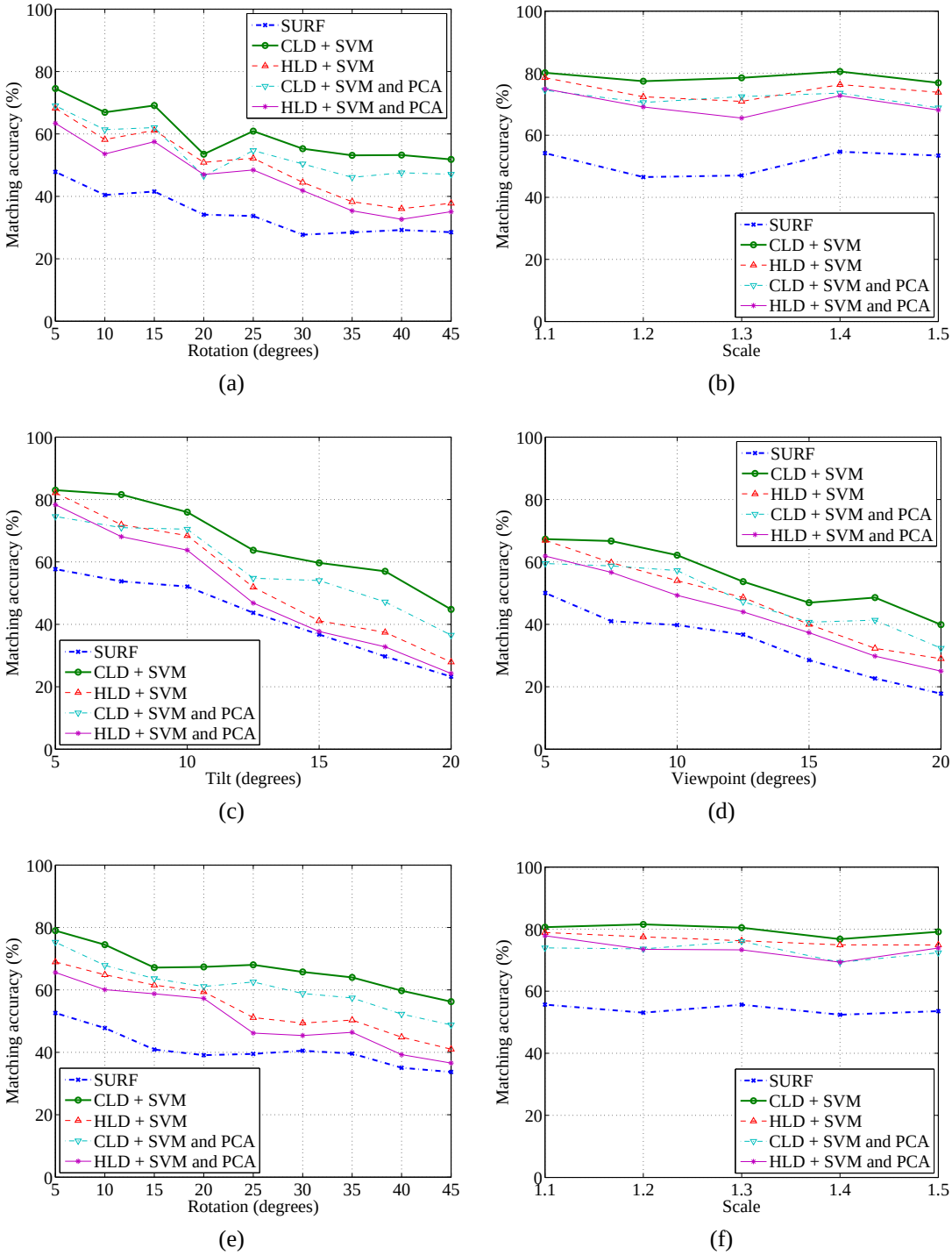


Figure B.20: Image matching results using the colour local descriptor and hybrid local descriptor methods combined with the SVM matching method with the PCA feature-reduction method for the patu and wahaika artefacts with different image transformations: (a) rotation (patu); (b) scale (patu); (c) tilt (patu); (d) viewpoint (patu); (e) rotation (wahaika); and (f) scale (wahaika).

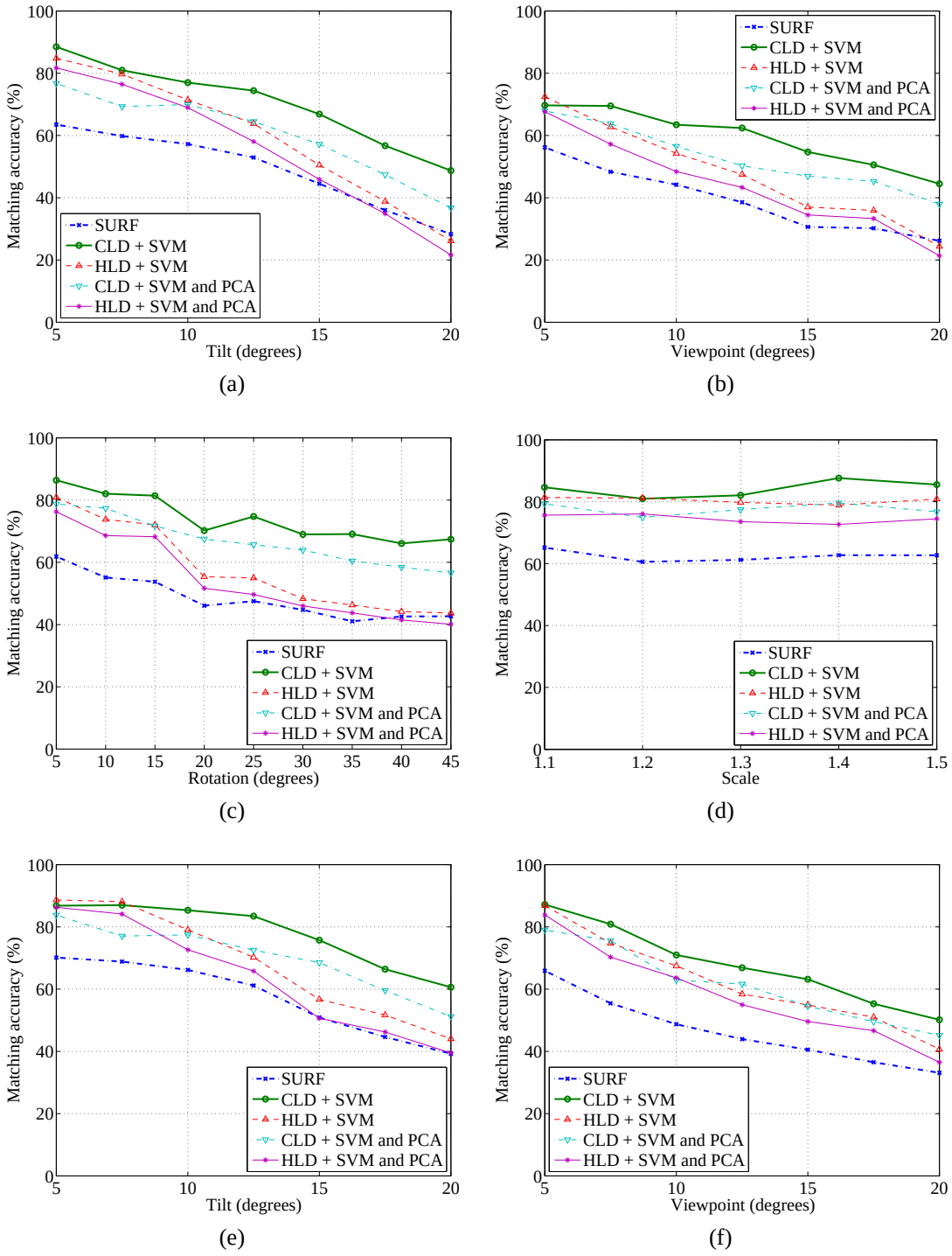


Figure B.21: Image matching results using the colour local descriptor and hybrid local descriptor methods combined with the SVM matching method with the PCA feature-reduction method for the wahaika and tiki artefacts with different image transformations: (a) tilt (wahaika); (b) viewpoint (wahaika); (c) rotation (tiki); (d) scale (tiki); (e) tilt (tiki); and (f) viewpoint (tiki).

B.3.3 Test 3: Feature-Reduced Local Descriptor Formation Methods Combined with SVM Matching Method

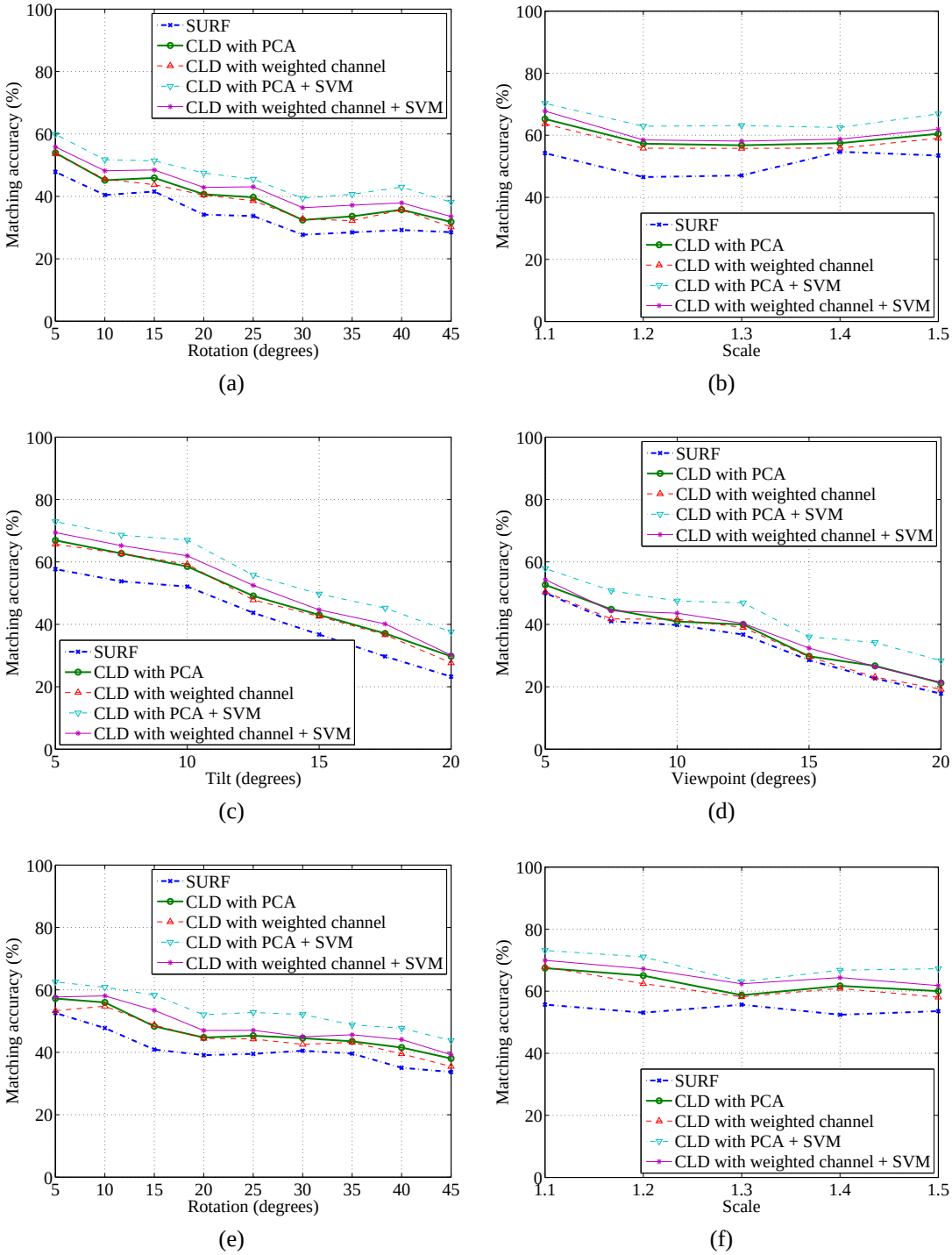


Figure B.22: Image matching results using the colour local descriptor and hybrid local descriptor methods with the two feature-reduction methods combined with the SVM matching method for the patu and wahaika artefacts with different image transformations: (a) rotation (patu); (b) scale (patu); (c) tilt (patu); (d) viewpoint (patu); (e) rotation (wahaika); and (f) scale (wahaika).

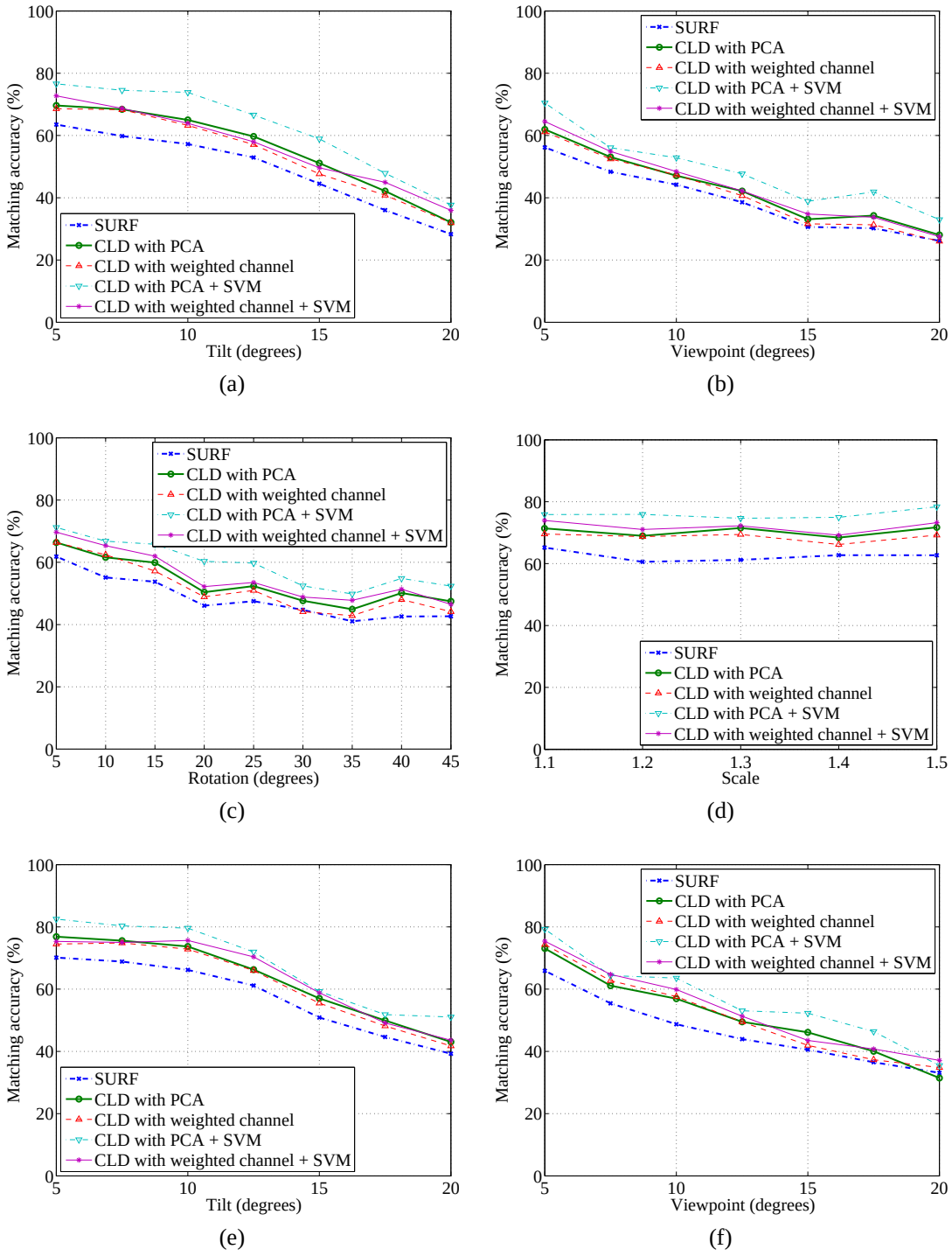


Figure B.23: Image matching results using the colour local descriptor and hybrid local descriptor methods with the two feature-reduction methods combined with the SVM matching method for the wahaika and tiki artefacts with different image transformations: (a) tilt (wahaika); (b) viewpoint (wahaika); (c) rotation (tiki); (d) scale (tiki); (e) tilt (tiki); and (f) viewpoint (tiki).

B.3.4 Test 4: Feature-Reduced Local Descriptor Formation Methods Combined with Feature-Reduced SVM Matching Method

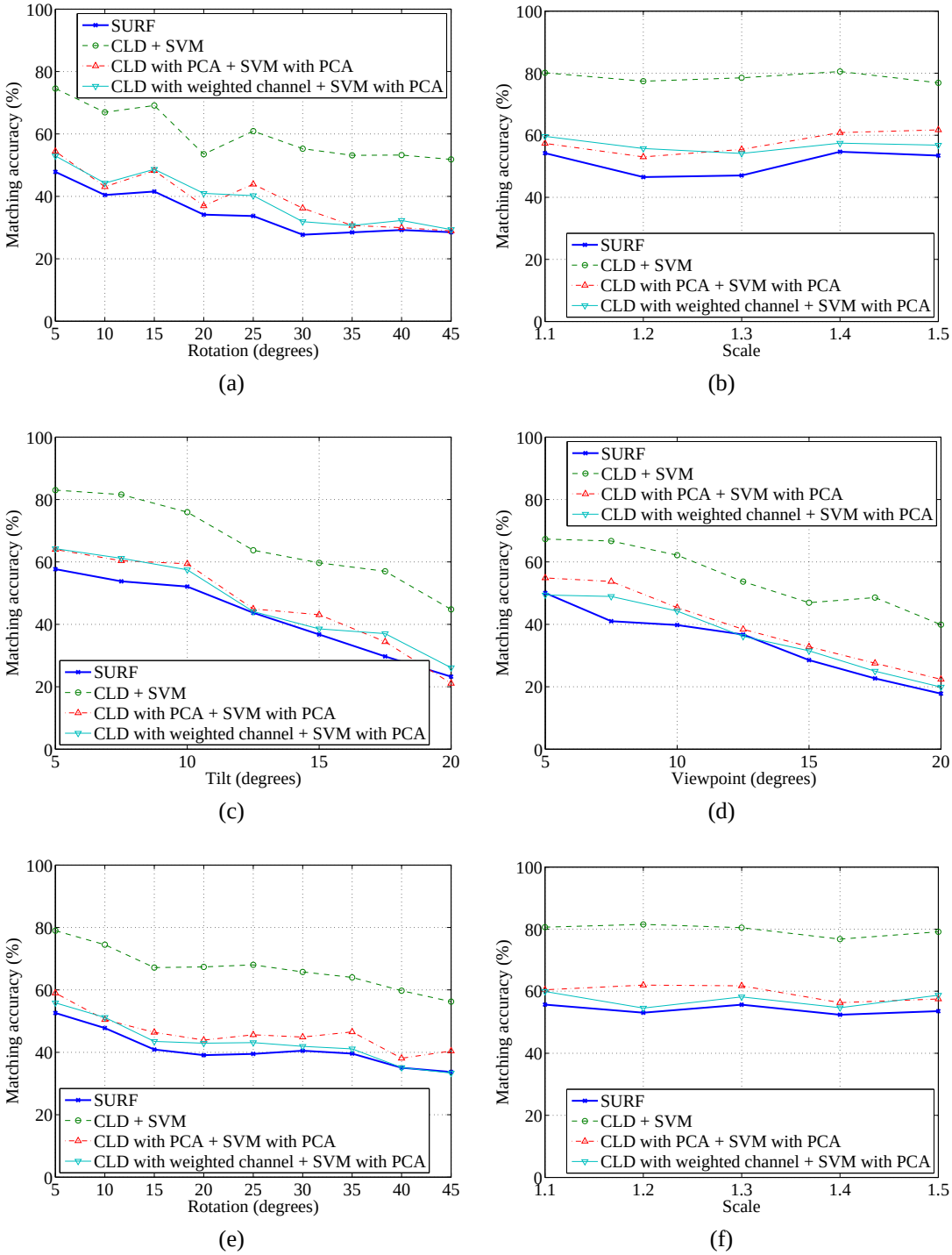


Figure B.24: Image matching results using the colour local descriptor and hybrid local descriptor methods with the two feature-reduction methods combined with the SVM matching method with the PCA feature-reduction method for the patu and wahaika artefacts with different image transformations: (a) rotation (patu); (b) scale (patu); (c) tilt (patu); (d) viewpoint (patu); (e) rotation (wahaika); and (f) scale (wahaika).

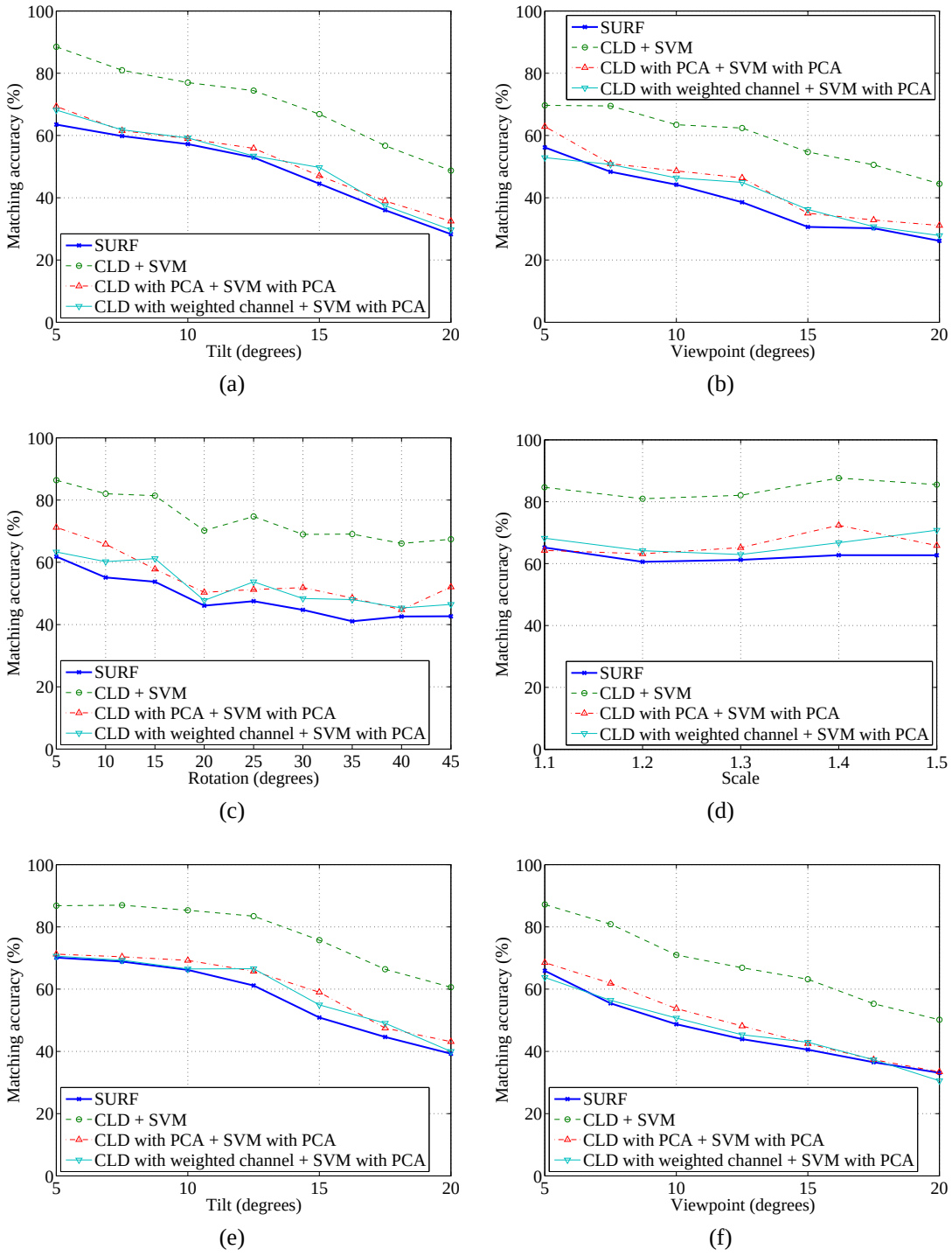


Figure B.25: Image matching results using the colour local descriptor and hybrid local descriptor methods with the two feature-reduction methods combined with the SVM matching method with the PCA feature-reduction method for the wahaika and tiki artefacts with different image transformations: (a) tilt (wahaika); (b) viewpoint (wahaika); (c) rotation (tiki); (d) scale (tiki); (e) tilt (tiki); and (f) viewpoint (tiki).

B.3.5 Test 5: Local Descriptor Methods Combined with SVM Matching Method for Illumination Changes

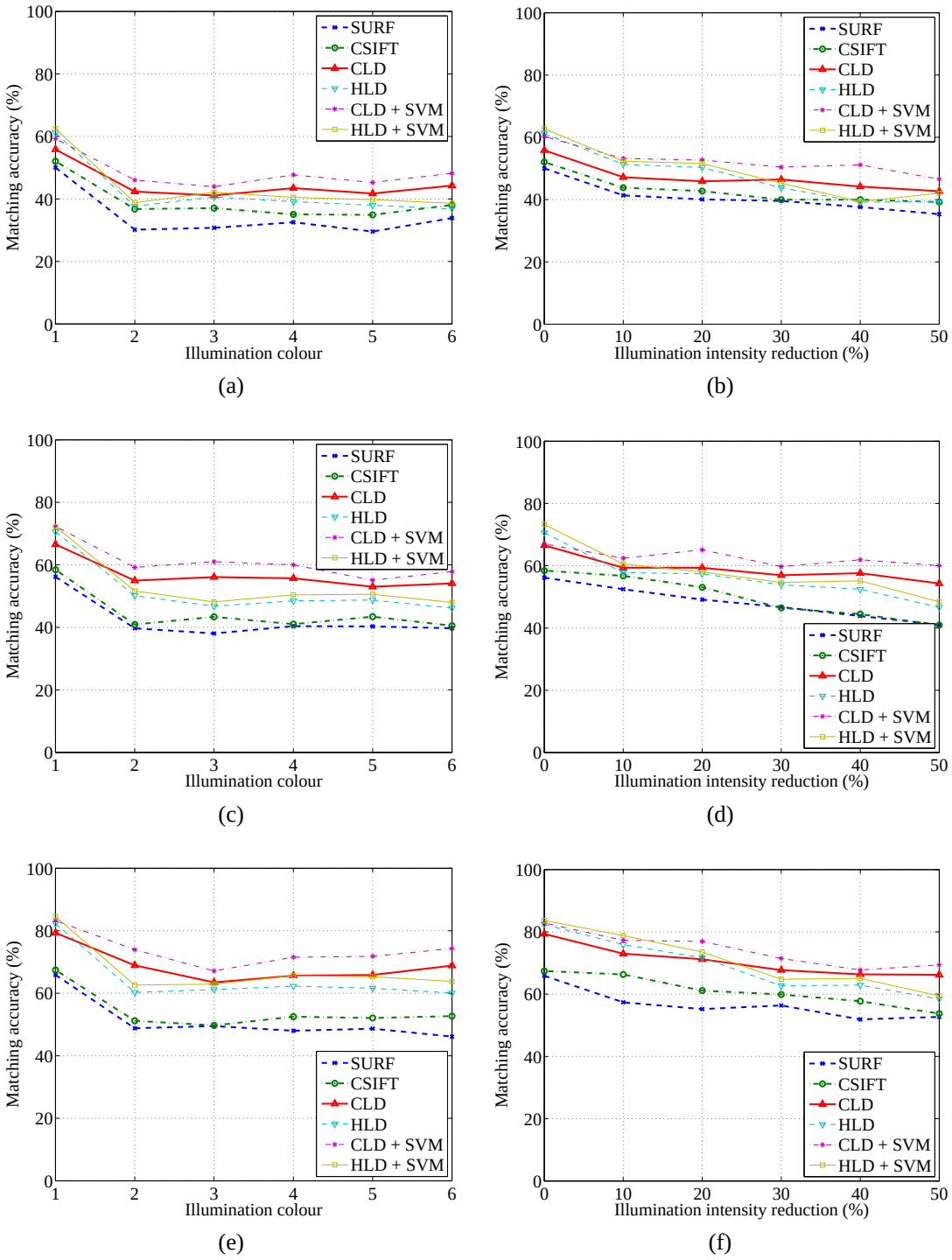


Figure B.26: Image matching results using the colour local descriptor and hybrid local descriptor methods combined with the SVM matching method for the patu, wahaika and tiki artefacts with different illumination changes: (a) colour (patu); (b) intensity (patu); (c) colour (wahaika); (d) intensity (wahaika); (e) colour (tiki); and (f) intensity (tiki).

B.3.6 Assisted Image Registration

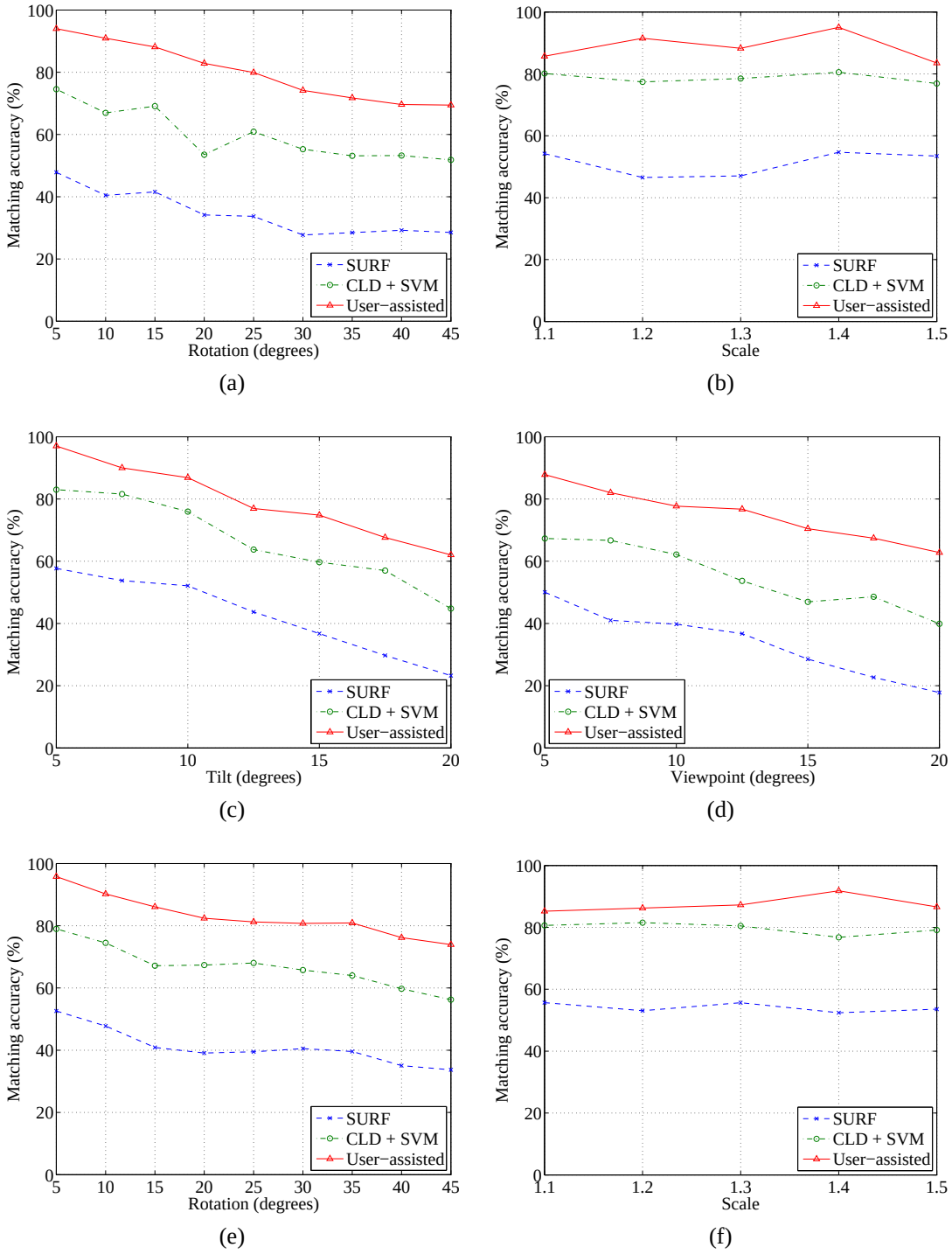


Figure B.27: Image matching results using three different matching methods for the patu and wahaika artefacts with different image transformations: (a) rotation (patu); (b) scale (patu); (c) tilt (patu); (d) viewpoint (patu); (e) rotation (wahaika); and (f) scale (wahaika).

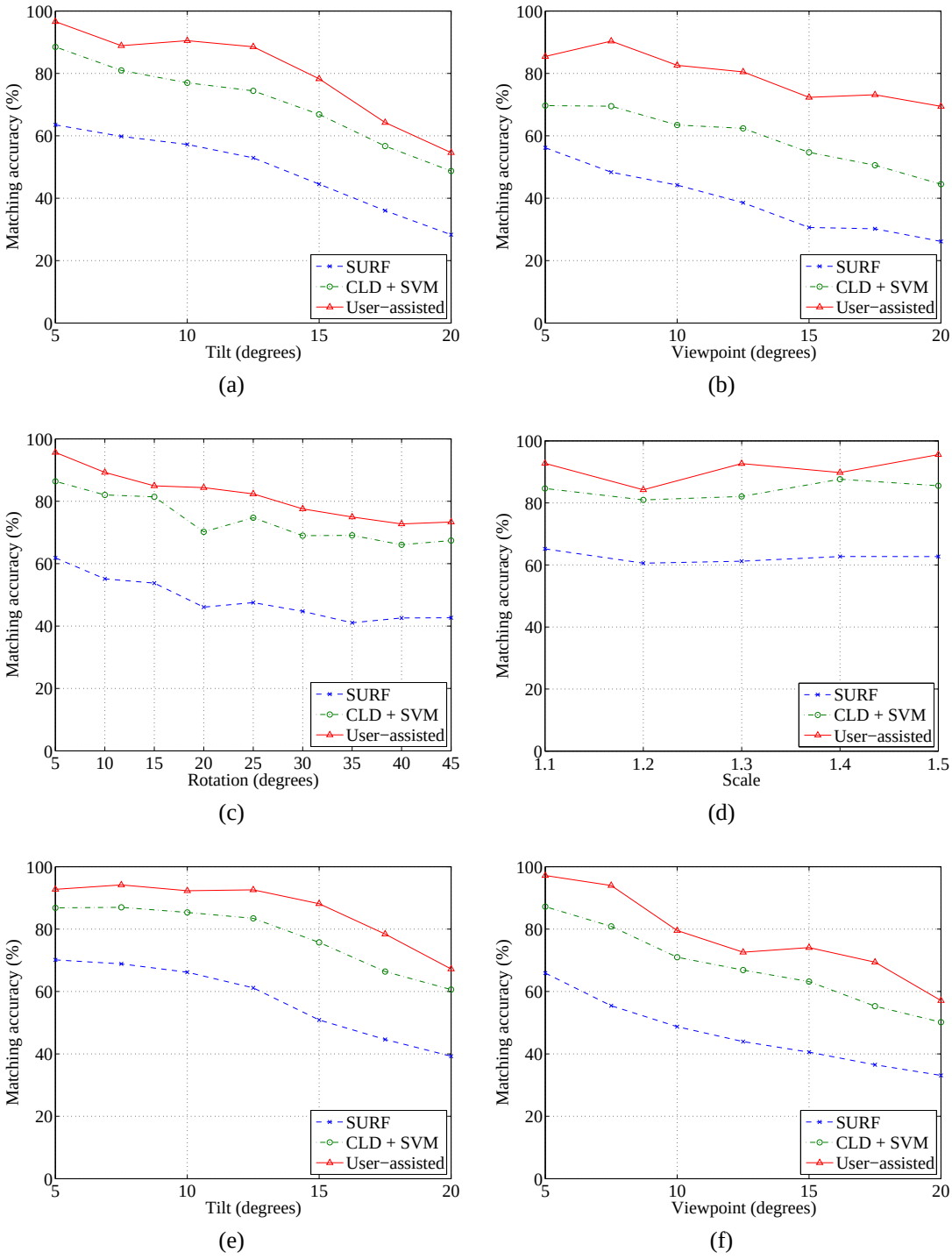


Figure B.28: Image matching results using three different matching methods for the wahaika and tiki artefacts with different image transformations: (a) tilt (wahaika); (b) viewpoint (wahaika); (c) rotation (tiki); (d) scale (tiki); (e) tilt (tiki); and (f) viewpoint (tiki).

References

[1] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 519–526, 2006.

[2] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8, 2007.

[3] D. Scharstein, "vision.middlebury.edu/mview." [Online], August 2008. Available: `http://vision.middlebury.edu/mview/` [Accessed: 7 March, 2009].

[4] University of Virginia, "Scanning monticello - university of virginia." [Online], April 2002. Available: `http://www.cs.virginia.edu/Monticello/` [Accessed: 14 April, 2008].

[5] M. Levoy, "The digital michelangelo project." [Online], June 2006. Available: `http://www-graphics.stanford.edu/projects/mich/` [Accessed: 12 December, 2008].

[6] S. Hammann, "Vihap3d - virtual heritage: High quality 3d acquisition and presentation." [Online], June 2004. Available: `http://www.vihap3d.org/` [Accessed: 2 January, 2008].

[7] X. Xu, "3d digitisation of museum artefacts." [Online], June 2008. Available: `http://www.engineers.auckland.ac.nz/~xxu008/MuseumEPICSProject/index.htm` [Accessed: 8 June 2009].

[8] J. H. Coote, *The Illustrated History of Colour Photography*. Fountain Press, 1993.

[9] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image and Vision Computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[10] Z. Yaniv, L. Joskowicz, A. Simkin, M. Garza-Jinich, and C. Milgrom, "Fluoroscopic image processing for computer-aided orthopaedic surgery," *Lecture Notes in Computer Science*, vol. 1496, pp. 325–334, 1998.

[11] Z. Yaniv and L. Joskowicz, "Long bone panoramas from fluoroscopic x-ray images," *IEEE Transactions on Medical Imaging*, vol. 23, no. 1, pp. 26–35, 2004.

[12] M. Brown, "Autostitch." [Online], June 2009. Available: `http://www.cs.ubc.ca/~mbrown/autostitch/autostitch.html` [Accessed: 10 June, 2009].

[13] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 257–263, 2005.

[14] N. Williams, C. Hantak, K. L. Low, J. Thomas, K. Keller, L. Nyland, D. Luebke, and A. Lastra, "Monticello through the window," *4th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*, 2003.

[15] S. Kumar, D. Snyder, D. Duncan, J. Cohen, and J. Cooper, "Digital preservation of ancient cuneiform tablets using 3d-scanning," *Fourth International Conference on 3-D Digital Imaging and Modeling*, pp. 326–333, 2003.

[16] G. Guidi, L. Micoli, M. Russo, B. Frischer, M. De Simone, A. Spinetti, and L. Carosso, "3d digitization of a large model of imperial rome," *Fifth International Conference on 3-D Digital Imaging and Modeling*, pp. 565–572, 2005.

[17] M. Levoy, J. Ginsberg, J. Shade, D. Fulk, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. Anderson, and J. Davis, "The digital michelangelo project: 3d scanning of large statues," *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 131–144, 2000.

[18] N. Surendran, A. Poolman, R. Raghavan, T. Germain, A. Lau, S. Fairley, N. Siddique, H. Mulchandanni, R. Sharma, G. Bhandari, and C. Cowley, "Engineering projects in community service (epics) - museum project," epics report, The University of Auckland, 13 April 2006.

[19] "Auckland War Memorial Museum tells the story of New Zealand, from our unique flora and fauna and our history at war, to our priceless collection of Maori and Pacific treasures. – Auckland Museum New Zealand." [Online]. Available: `http://www.aucklandmuseum.com/` [Accessed: 21 May, 2007].

[20] M. Kim, K. Lee, C. Lu, S. Wang, D. Wijekoon, Q. Ao, Z. Chen, J. Chu, P. Du, A. Gulati, R. Parameshwar, and J. Zhang, "Engineering projects in community service (epics) – museum project," epics report, The University of Auckland, 2007.

[21] I. Lim, M. Mathur, J. Zhang, Y. S. Tan, K. M. Lin, L. Anwar, M. Bajaj, A. Gulati, K. Lee, and D. Cheng, "Engineering projects in community service (epics) – museum project," epics report, The University of Auckland, 2008.

[22] J. Salvi, C. Matabosch, D. Fofi, and J. Forest, "A review of recent range image registration methods with accuracy evaluation," *Image and Vision Computing*, vol. In press, 2006.

[23] G. Yang, C. V. Steward, M. Sofka, and C. Tsai, "Automatic robust image registration system: Initialization, estimation, and decision," *Fourth IEEE International Conference on Computer Vision Systems (ICVS 2006)*, 2006.

[24] S. Kaneko, Y. Satoh, and S. Igarashi, "Using selective correlation coefficient for robust image registration," *Pattern Recognition*, vol. 36, pp. 1165–1173, 2003.

[25] L. G. Brown, "A survey of image registration techniques," *ACM Computing Surveys*, vol. 24, pp. 325–376, December 1992.

[26] R. Bracewell, "The fourier transform and its applications," *American Association of Physics Teachers*, vol. 34, p. 712, August 1966.

[27] E. De Castro and C. Morandi, "Registration of translated and rotated images using finite fourier transforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 5, pp. 700–703, 1987.

[28] T. M. Lehmann, "A two-stage algorithm for model-based registration of medical images," *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, vol. 1, pp. 344–351, August 1998.

[29] Q. Chen, M. Defrise, and F. Deconinck, "Symmetric phase-only matched filtering of fourier-mellin transform for image registration and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 1156–1168, 1994.

[30] A. Hii, C. E. Hann, J. G. Chase, and E. E. W. Van Houten, "Fast normalized cross correlation for motion tracking using basis functions," *Computer Methods and Programs in Biomedicine*, vol. 82, no. 2, pp. 144–156, 2006.

[31] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, pp. 137–154, 1997.

[32] P. Thévenaz and M. Unser, "An efficient mutual information optimizer for multiresolution image registration," *Proceedings of the IEEE International Conference on Image Processing ICIP'98*, pp. 833–837, 1998.

[33] N. Ritter, R. Owens, J. Cooper, R. H. Eikelboom, and P. P. van Saarloos, "Registration of stereo and temporal images of the retina," *IEEE Transactions of Medical Imaging*, vol. 18, pp. 404–418, 1999.

[34] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3d medical image alignment," *Pattern Recognition*, vol. 32, pp. 71–86, 1999.

[35] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, pp. 187–198, 1997.

[36] M. Roux, "Automatic registration of spot images and digitized maps," *Proceedings of the IEEE International Conference on Image Processing ICIP'96*, pp. 625–628, 1996.

[37] B. Likar and F. Pernus, "Automatic extraction of corresponding points for the registration of medical images," *Medical Physics*, vol. 26, pp. 1678–1686, 1999.

[38] B. Likar and F. Pernus, "A hierarchical approach to elastic registration based on mutual information," *Image and Vision Computing*, vol. 19, pp. 33–44, 2001.

[39] G. Stockman, S. Kopstein, and S. Benett, "Matching images to models for registration and object detection via clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-4, no. 3, pp. 229–241, 1982.

[40] J. Ton and A. K. Jain, "Registering landsat images by point matching," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 27, no. 5, pp. 642–651, 1989.

[41] J. W. Hsieh, H. Y. M. Liao, K. C. Fam, and M. T. Ko, "A fast algorithm for image registration without predetermining correspondences," *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, vol. 1, pp. 765–769, 1996.

[42] L. M. G. Fonseca and M. H. M. Costa, "Automatic registration of satellite images," *Proceedings of the 1997 10th Brazilian Symposium of Computer Graphic and Image Processing*, pp. 219–226, 1997.

[43] H. P. Moravec, *Obstacle avoidance and navigation in the real world by a seeing robot rover*. PhD thesis, Carnegie-Mellon University, 1980.

[44] C. Harris and M. Stephens, "A combined corner and edge detector," *Proceedings of the 4th Alvey Vision Conference*, pp. 147–151, 1988.

[45] C. Tomasi and T. Kanade, “Detection and tracking of point features,” tech. rep., Carnegie-Mellon University, 1991.

[46] S. M. Smith and J. M. Brady, “Susan - a new approach to low level image processing,” *International Journal of Computer Vision*, vol. 23, no. 1, pp. 45–78, 1997.

[47] M. Trajkovic and M. Hedley, “Fast corner detection,” *Image and Vision Computing*, vol. 16, no. 2, pp. 75–87, 1998.

[48] S. Z. Li, J. Kittler, and M. Petrou, “Matching and recognition of road networks from aerial images,” *Lecture Notes in Computer Science*, vol. 588, pp. 857–861, 1992.

[49] H. Maitre and Y. Wu, “Improving dynamic programming to solve image registration,” *Pattern Recognition*, vol. 20, no. 4, pp. 443–462, 1987.

[50] D. Shin, J. K. Pollard, and J. P. Muller, “Accurate geometric correction of atsr images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 4, pp. 997–1006, 1997.

[51] N. Vujovic and D. Brzakovic, “Establishing the correspondence between control points in pairs of mammographic images,” *IEEE Transactions on Image Processing*, vol. 6, no. 10, pp. 1388–1399, 1997.

[52] J. Canny, “A computational approach to edge detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 679–698, November 1986.

[53] D. Ziou and S. Tabbone, “Edge detection techniques - an overview,” *International Journal of Pattern Recognition and Image Analysis*, vol. 8, pp. 537–559, 1998.

[54] M. Basu, “Gaussian-based edge-detection methods – a survey,” *IEEE Transactions on Systems, Manufacturing and Cybernetics Part C: Applications and Reviews*, vol. 32, no. 3, pp. 252–260, 2002.

[55] H. Trichili, M. S. Bouhlel, N. Derbel, and L. Kamoun, “A survey and evaluation of edge detection operators application to medical images,” *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, pp. 706–709, 2002.

[56] M. Holm, “Towards automatic rectification of satellite images using feature based matching,” *Digest – International Geoscience and Remote Sensing Symposium*, vol. 4, pp. 2439–2442, 1991.

[57] A. Goshtasby and G. C. Stockman, “Point pattern matching using convex hull edges,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-15, no. 5, pp. 631–637, 1985.

[58] A. Goshtasby, G. C. Stockman, and C. V. Page, “A region-based approach to digital image registration with subpixel accuracy,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-24, no. 3, pp. 390–399, 1986.

[59] Y. C. Hsieh, D. M. McKeown, and F. P. Perlant, “Performance evaluation of scene registration and stereo matching for cartographic feature extraction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 214–238, 1992.

[60] P. A. Brivio, A. della Ventura, A. Rampini, and R. Schettini, “Automatic selection of control-points from shadow structures,” *International Journal of Remote Sensing*, vol. 13, no. 10, pp. 1853–1860, 1992.

[61] N. R. Pal and S. K. Pal, “A review on image segmentation techniques,” *Pattern Recognition*, vol. 26, no. 9, pp. 1277–1294, 1993.

[62] D. G. Lowe, “Object recognition from local scale-invariant features,” *International Conference on Computer Vision*, pp. 1150–1157, 1999.

[63] A. Goshtasby, “Piecewise linear mapping functions for image registration,” *Pattern Recognition*, vol. 19, no. 6, pp. 459–466, 1986.

[64] A. Collignon, D. Vandermeulen, P. Suetens, and G. Marchal, “3d multi-modality medical image registration using feature space clustering,” *Lecture Notes in Computer Science*, vol. 905, p. 195, 1995.

[65] J. Tsao, “Efficient interpolation for clustering-based multimodality image registration,” *In Proceeding International Society of Magnetic Resonence Medicine*, p. 2195, 1999.

[66] J. Tsao and P. Lauterbur, “Generalized clustering-based image registration for multimodality images,” *Proceedings of the 20th Annual International Conference of the IEEE*, vol. 2, pp. 667–670, 1998.

[67] A. Pitiot, E. Bardinet, P. M. Thompson, and G. Malandain, “Piecewise affine registration of biological images for volume reconstruction,” *Medical Image Analysis*, vol. 10, no. 3, pp. 465–483, 2006.

[68] Y. Rubner, C. Tomasi, and L. J. Guibas, “Earth mover’s distance as a metric for image retrieval,” *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[69] S. H. Chang, F. H. Cheng, W. H. Hsu, and G. Z. Wu, “Fast algorithm for point pattern matching: invariant to translations, rotations and scale changes,” *Pattern Recognition*, vol. 30, no. 2, pp. 311–320, 1997.

[70] X. Yu and H. Sun, "Automatic image registration via clustering and convex hull vertices matching," *Lecture Notes in Artificial Intelligence*, vol. 3584, pp. 439–445, 2005.

[71] H. G. Barrow, *Parametric correspondence and chamfer matching: Two new techniques for image matching*. SRI International, 1977.

[72] A. Goshtasby, *2-D and 3-D image registration*. Wiley, 2005.

[73] G. Borgefors, "Hierarchical chamfer matching: A parametric edge matching algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 6, pp. 849–865, 1988.

[74] A. Thayananthan, B. Stenger, P. Torr, and R. Cipolla, "Shape context and chamfer matching in cluttered scenes," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 127–133, 2003.

[75] M. K. Hu, "Visual pattern recognition by moment invariants," *IRE – Transactions on Information Theory*, vol. IT-8, no. 2, pp. 179–187, 1962.

[76] R. Maghsoodi and B. Rezaie, "Image registration using a fast adaptive algorithm," *Proceedings of SPIE – The International Society for Optical Engineering*, vol. 757, pp. 58–63, 1987.

[77] J. Flusser and T. Suk, "Degraded image analysis: An invariant approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 590–603, 1998.

[78] J. Flusser, J. Boldys, and B. Zitova, "Moment forms invariant to rotation and blur in arbitrary number of dimensions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 234–246, 2003.

[79] S. Ranade and A. Rosenfeld, "Point pattern matching by relaxation," *Pattern Recognition*, vol. 12, no. 4, pp. 269–275, 1980.

[80] D. Cheng, S. Q. Xie, and E. Hämmerle, "Comparison of local descriptors for image registration of geometrically-complex 3d scenes," *14th IEEE International Conference on Mechatronics and Machine Vision in Practice*, pp. 140–145, December 2007.

[81] A. P. Witkin, "Scale-space filtering," *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, vol. 1, pp. 1019–1022, 1983.

[82] T. Lindeberg, "Scale-space theory: A basic tool for analysing structures at different scales," *Journal of Applied Statistics*, vol. 21, no. 2, pp. 224–270, 1994.

[83] T. Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998.

[84] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture Notes in Computer Science*, vol. 3951 NCS, pp. 404–417, 2006.

[85] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," *Lecture Notes in Computer Science*, vol. 2350, pp. 128–142, 2002.

[86] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[87] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10 SPEC ISS, pp. 761–767, 2004.

[88] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector," *Lecture Notes in Computer Science*, vol. 3021, pp. 228–241, 2004.

[89] T. Tuytelaars and L. Van Gool, "Content-based image retrieval based on local affinely invariant regions," *Lecture Notes in Computer Science*, vol. 1614, p. 656, 1999.

[90] T. Tuytelaars and L. Van Gool, "Wide baseline stereo matching based on local, affinely invariant regions," *British Machine Vision Conference*, pp. 415–425, 2000.

[91] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[92] M. Donoser and H. Bischof, "Efficient maximally stable extremal region (mser) tracking," *Proceedings - 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 553–560, 2006.

[93] L. Shao and M. Brady, "Specific object retrieval based on salient regions," *Pattern Recognition*, vol. 39, no. 10, pp. 1932–1948, 2006.

[94] Y. G. Li, Y. F. Ma, and H. J. Zhang, "Salient region detection and tracking in video," *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, vol. 2, pp. 269–272, 2003.

[95] G. Wang, L. Zhang, and C. Wang, “Algorithm for detecting sea-surface targets with sequential salient feature against complicated scene,” *Huazhong Keji Daxue Xuebao (Ziran Kexue Ban)/Journal of Huazhong University of Science and Technology (Natural Science Edition)*, vol. 34, no. 10, pp. 28–30, 2006.

[96] T. Tuytelaars and L. Van Gool, “Matching widely separated views based on affine invariant regions,” *International Journal of Computer Vision*, vol. 59, no. 1, pp. 61–85, 2004.

[97] Y. Ke and R. Sukthankar, “Pca-sift: A more distinctive representation for local image descriptors,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 506–513, 2004.

[98] A. E. Abdel-Hakim and A. A. Farag, “Csift: A sift descriptor with color invariant characteristics,” *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, pp. 1978–1983, 2006.

[99] P. Kubelka, “New contribution to the optics of intensely lightscattering materials, part i,” *Journal of the Optical Society of America*, vol. 38, no. 5, pp. 448–457, 1948.

[100] A. Haar, “Zur theorie der orthogonalen funktionensysteme,” *Mathematische Annalen*, vol. 69, p. 1910, 1910.

[101] M. S. Sarfraz and O. Hellwich, “Head pose estimation in face recognition across pose scenarios,” *International Conference on Computer Vision Theory and Applications*, pp. 235–242, 2008.

[102] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[103] S. Lazebnik, C. Schmid, and J. Ponce, “A sparse texture representation using local affine regions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1265–1278, 2005.

[104] A. E. Johnson and M. Hebert, “Using spin images for efficient object recognition in cluttered 3d scenes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 433–449, May 1999.

[105] F. Schaffalitzky and A. Zisserman, “Multi-view matching for unordered image sets, or “how do i organise my holiday snaps?”,” *Lecture Notes in Computer Science*, vol. 2350, pp. 414–431, 2002.

[106] L. Van Gool, T. Moons, and D. Ungureanu, "Affine/photometric invariants for planar intensity patterns," *Lecture Notes in Computer Science*, vol. 1064, pp. 642–651, 1996.

[107] Y. Li and P. Gu, "Free-form surface inspection techniques state of the art review," *CAD Computer Aided Design*, vol. 36, no. 13, pp. 1395–1417, 2004.

[108] L. Kobbelt and M. Botsch, "A survey of point-based techniques in computer graphics," *Computers and Graphics*, vol. 28, no. 6, pp. 801–814, 2004.

[109] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1067–1073, 1997.

[110] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science*, vol. 194, pp. 283–287, 15 October 1976.

[111] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 150–162, 1994.

[112] K. H. Bae, J. H. Ko, and E. S. Kim, "Regularized stereo matching scheme using adaptive disparity estimation," *Japanese Journal of Applied Physics, Part 1: Regular Papers and Short Notes and Review Papers*, vol. 45, no. 5 A, pp. 4107–4114, 2006.

[113] V. Vaish, M. Levoy, R. Szeliski, C. Zitnick, and S. B. Kang, "Reconstructing occluded surfaces using synthetic apertures: stereo, focus and robust measures," *Proceedings - 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006*, vol. 2, pp. 2331–2338, 2006.

[114] S. Fan and F. P. Ferrie, "Photo hull regularized stereo," *Computer and Robot Vision, 2006. The 3rd Canadian Conference on*, pp. 18–24, 2006.

[115] M. Goesele, B. Curless, and S. M. Seitz, "Multi-view stereo revisited," *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, pp. 2402–2409, 2006.

[116] R. T. Collins, "Space-sweep approach to true multi-image matching," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 358–363, 1996.

[117] G. G. Slabaugh, W. B. Culbertson, T. Malzbender, M. R. Stevens, and R. W. Schafer, "Methods for volumetric reconstruction of visual scenes," *International Journal of Computer Vision*, vol. 57, no. 3, pp. 179–199, 2004.

[118] M. R. Stevens, B. Culbertson, and T. Malzbender, "A histogram-based color consistency test for voxel coloring," *Proceedings – International Conference on Pattern Recognition*, vol. 16, no. 4, pp. 118–121, 2002.

[119] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, second ed., 2003.

[120] K. N. Kutulakos and S. M. Seitz, "What do n photographs tell us about 3d shape?," Technical Report TR680, Computer Science Department, University of Rochester, January 1998.

[121] W. B. Culbertson, T. Malzbender, and G. Slabaugh, "Generalized voxel coloring," *Lecture Notes in Computer Science*, vol. 1883, pp. 100–115, January 2000.

[122] H. Weghorst, G. Hooper, and D. P. Greenberg, "Improving computational methods for ray tracing," *ACM Transactions on Graphics*, vol. 3, no. 1, pp. 52–69, 1984.

[123] J. Shade, S. Gortler, L. He, and R. Szeliski, "Layered depth images," *Proceedings of SIGGRAPH*, pp. 231–242, 1998.

[124] N. Max, "Hierarchical rendering of trees from precomputed multi-layer z-buffers," *Eurographics Rendering Workshop*, pp. 165–174, 1996.

[125] H. Kim and I. S. Kweon, "Appearance-cloning: Photo-consistent scene recovery from multi-view images," *International Journal of Computer Vision*, vol. 66, no. 2, pp. 163–192, 2006.

[126] R. Bhotika, D. J. Fleet, and K. N. Kutulakos, "A probabilistic theory of occupancy and emptiness," *Lecture Notes in Computer Science*, vol. 2352/2002, pp. 112–130, 2002.

[127] R. Bhotika, *Scene-space methods for bayesian inference of 3D shape and motion*. Phd thesis, University of Rochester, 2004.

[128] P. Eisert, E. Steinbach, and B. Girod, "Multi-hypothesis, volumetric reconstruction of 3-d objects from multiple calibrated camera views," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 6, pp. 3509–3515, 1999.

[129] P. Eisert, E. Steinbach, and B. Girod, "Automatic reconstruction of stationary 3-d objects from multiple uncalibrated camera views," *IEEE Transactions on Circuits and Systems For Video Technology*, vol. 10, no. 2, pp. 261–277, 2000.

[130] G. Slabaugh, B. Culbertson, T. Malzbender, and R. Schafer, "A survey of methods for volumetric scene reconstruction from photographs," *Volume Graphics*, pp. 81–100, 2001.

[131] E. Steinbach, B. Girod, P. Eisert, and A. Betz, "3-d reconstruction of real-world objects using extended voxels," *IEEE International Conference on Image Processing*, vol. 1, pp. 569–572, 2000.

[132] A. Broadhurst, *A probabilistic framework for space carving*. PhD thesis, Trinity College, September 2001.

[133] A. Broadhurst and R. Cipolla, "A statistical consistency check for the space carving algorithm," *Proc. ICCV*, vol. 1, pp. 282–291, 2001.

[134] T. Bayes, "An essay towards solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418, 1763.

[135] V. Chhabra, "Reconstructing specular objects with image based rendering using color caching," master thesis, Worcester Polytechnic Institute, 2001.

[136] R. Yang, M. Pollefeys, and G. Welch, "Dealing with textureless regions and specular highlights – a progressive space carving scheme using a novel photo-consistency measure," *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, vol. 1, pp. 576–584, 2003.

[137] T. Bonfort and P. Sturm, "Voxel carving for specular surfaces," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 591–596, 2003.

[138] J. S. De Bonet and P. Viola, "Roxels: Responsibility weighted 3d volume reconstruction," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 418–425, 1999.

[139] J. S. De Bonet and P. Viola, "Poxels: Probabilistic voxelized volume reconstruction," *International Conference on Computer Vision, Proceedings of*, 1999.

[140] A. C. Prock and C. R. Dyer, "Towards real-time voxel coloring," *DARPA Image Understanding Workshop*, pp. 315–312, 1998.

[141] S. Vedula, S. Baker, S. Seitz, and T. Kanade, "Shape and motion carving in 6d," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 592–598, 2000.

[142] B. Goldlücke and M. Magnor, "Space-time isosurface evolution for temporally coherent 3d reconstruction," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 350–355, 2004.

[143] M. Magnor and B. Goldlücke, "Spacetime-coherent geometry reconstruction from multiple video streams," *Proceedings - 2nd International Symposium on 3D Data Processing, Visualization, and Transmission*, pp. 365–372, 2004.

[144] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proceedings of Imaging understanding workshop,*, pp. 121–130, 1981.

[145] C. Leung, B. Appleton, and C. Sun, "Embedded voxel colouring," *Digital Image Computing: Techniques and Applications*, vol. 2, pp. 623–632, 2003.

[146] C. W. Y. Leung, *Efficient Methods for 3D Reconstruction From Multiple Images*. Phd thesis., School of Information Technology and Electrical Engineering, The University of Queensland, February 2006.

[147] O. Batchelor, O. Mukundan, and R. Green, "Incremental voxel colouring by ray traversal," *Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation*, pp. 396–401, 2006.

[148] M. Levoy, "The stanford spherical gantry." [Online], June 2006. Available: `http://graphics.stanford.edu/projects/gantry/` [Accessed: 9 June, 2009].

[149] N. Campbell, G. Vogiatzis, C. Hernandez, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, 2008.

[150] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz, "Multi-view stereo for community photo collections.," *ICCV*, 2007.

[151] M. Habbecke and L. Kobbelt, "A surface-growing approach to multi-view stereo reconstruction," *CVPR*, 2007.

[152] C. Hernandez and F. Schmitt, "Silhouette and stereo fusion for 3d object modeling," *Computer Vision and Image Understanding*, vol. 96, no. 3, pp. 367–392, 2004.

[153] A. Hornung and L. Kobbelt, "Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding," *In IEEE Conference on Computer Vision and Pattern Recognition*, pp. 503–510, 2006.

[154] G. Vogiatzis, C. Hernandez, P. Torr, and R. Cipolla, "Multi-view stereo via volumetric graph-cuts and occlusion robust photo-consistency," *PAMI*, 2007.

[155] C. Zach, "Fast and high quality fusion of depth maps," *3DPVT*, 2008.

[156] K. E. A. Van de Sande, "A practical setup for voxel coloring using off-the-shelf components," bachelors thesis, University of Amsterdam, June 2004.

[157] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[158] "Canadian heritage information network." [Online], November 2008. Available: http://www.chin.gc.ca/English/index.html [Accessed: 2 March, 2008].

[159] "Salzburg research forschungsgesellschaft mbh." [Online]. Available: http://www.salzburgresearch.at/company/organisation_e.php [Accessed: 11 March, 2008].

[160] College of Architecture, Texas Tech University, "Digital statue." [Online]. Available: http://www.arch.ttu.edu/digital_liberty/ [Accessed: 15 March, 2008].

[161] "Museum of the city of New York." [Online]. Available: http://www.mcny.org/ [Accessed: 20 February, 2008].

[162] "Royal Ontario museum." [Online]. Available: http://www.rom.on.ca/ [Accessed: 24 February, 2008].

[163] "Museum of science, Boston." [Online]. Available: http://www.mos.org/ [Accessed: 19 February, 2008].

[164] "American Museum of Natural History." [Online]. Available: http://www.amnh.org/ [Accessed: 26 February, 2008].

[165] "Civilization.ca - cmc home (Canadian museum of civilization)." [Online]. Available: http://www.civilization.ca/cmc/home/cmc-home [Accessed: 27 February, 2008].

[166] B. Museum, "Home page — commissariat 3d reconstruction project." [Online], 2006. Available: http://www.virtualmuseum.ca/Exhibitions/Bytown/index_e.html [Accessed: 27 February, 2008].

[167] "Noma — new orleans museum of art." [Online], 2005. Available: http://www.noma.org/. [Accessed: 14 April, 2008].

[168] "Victoria and Albert museum." [Online]. Available: http://www.vam.ac.uk/ [Accessed: 29 January, 2008].

[169] The University of Auckland, "2009 projects – epics – the university of auckland." [Online], 2009. Available: http://www.epics.auckland.ac.nz/uoa/engineering/undergrad/epics/2009-projects.cfm [Accessed: 8 June, 2009].

[170] Polhemus, "Motion tracking, 3d scanning, and eye tracking solutions from polhemus." [Online]. Available: http://www.polhemus.com/ [Accessed: 30 June, 2009].

[171] Thinglab, "Thinglab – independent uk experts in 3d printing and 3d scanning – full solutions and services in the united kingdom." [Online]. Available: http://www.thinglab.co.uk/ [Accessed: 21 May, 2008].

[172] M. Papakura, *The Old-Time Maori*. Victor Gollancz Ltd, 1938.

[173] K. Mikolajczyk and C. Schmid, "Comparison of affine-invariant local detectors and descriptors," *Proceedings of the 12th European Signal Processing Conference*, pp. 1729–1732, 2004.

[174] J. Serra, *Image Analysis and Mathematical Morphology*. Academic Press, 1984.

[175] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE Transactions on Robotics and Automation*, vol. RA-3, pp. 323–344, August 1987.

[176] Intel, "Open source computer vision library." [Online]. Available: http://www.intel.com/technology/computing/opencv/index.htm [Accessed: 6 June, 2006].

[177] C. H. Chen, L. F. Pau, and P. S. P. Wang, eds., *Handbook of Pattern Recognition and Computer Vision*. World Scientific Publishing Co., Inc., 1993.

[178] G. Medioni and S. B. Kang, *Emerging topics in computer vision*. Prentice Hall PTR, 2005.

[179] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics Magazine*, vol. 38, pp. 114–117, April 1965.

[180] T. Gevers and A. Smeulders, "Color-based object recognition," *Pattern Recognition*, vol. 32, pp. 453–464, 1999.

[181] J. Albers, *Interaction of Color: Revised and Expanded Edition*. Yale University Press, 2006.

[182] J. Van de Weijer and C. Schmid, "Coloring local feature extraction," *Lecture Notes in Computer Science*, vol. 3952/2006, pp. 334–348, 2006.

[183] G. D. Finalyson, B. Schiele, and J. L. Crowley, "Comprehensive color image normalization," *ECCV '98: Proceedings of the 5th European Conference on Computer Vision*, vol. 1, pp. 475–490, 1998.

[184] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 525–531, 2001.

[185] H. F. Kaiser, "The application of electronic computers to factor analysis," *Educational and Psychological Measurement*, vol. 20, pp. 141–151, 1960.

[186] R. B. Cattell, "The scree test for the number of factors," *Multivariate Behavioral Research*, vol. 1, no. 2, pp. 245–276, 1966.

[187] J. P. Lewis, "Fast template matching," *Vision Interface*, pp. 120–123, 1995.

[188] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Prentice Hall, 2nd ed., 2002.

[189] S. J. Sangwine and T. Ell, "Hypercomplex auto- and cross-correlation of color images," *Preceedings of the 1999 IEEE International Conference on Image Processing*, vol. 4, pp. 319–322, 1999.

[190] W. R. Hamilton, *Elements of Quaternions*. Longmans, Green, & co., 2 ed., 1866.

[191] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[192] J. Beis and D. G. Lowe, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pp. 1000–1006, 1997.

[193] J. Friedman, J. Bentley, and R. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM Trans. Math. Software*, vol. 3, pp. 209–226, 1977.

[194] T. Huang, V. Kecman, and I. Kopriva, *Kernel Based Algorithms for Mining Huge Data Sets*. Springer, 2005.

[195] S. Abe, *Support Vector Machines for Pattern Classification*. Springer, 2005.

[196] C. Bahlmann, B. Haasdonk, and H. Burkhardt, "On-line handwriting recognition with support vector machines: A kernel approach," *Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition*, pp. 49–54, 2002.

[197] A. R. Ahmad, M. Khalia, C. Viard-Gaudin, and E. Poisson, "Online handwriting recognition using support vector machine," *TENCON 2004. 2004 IEEE Region 10 Conference*, vol. 1, pp. 311–314, 2004.

[198] L. S. Oliveira and R. Sabourin, "Support vector machines for handwritten numerical string recognition," *Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop on*, pp. 39–44, 2004.

[199] H. Nemmour and Y. Chibani, "New jaccard-distance based support vector machine kernel for handwritten digit recognition," *Information and Communication Technologies: From Theory to Applications*, pp. 1–4, 2008.

[200] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3d and multi-modal 3d + 2d face recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[201] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "An evaluation of multimodal 2d+3d face biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 619–624, April 2005.

[202] S. G. Kong, J. Heo, B. R. Abidi, J. Paik, and M. A. Abidi, "Recent advances in visual and infrared face recognition - a review," *Computer Vision and Image Understanding*, vol. 97, pp. 103–135, January 2005.

[203] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[204] V. Kecman, "Support vector machines basics," tech. rep., The University of Auckland, April 2004.

[205] V. Kecman, *Learning and Soft Computing – Support Vector Machines, Neural Networks, Fuzzy Logic Systems*. The MIT Press, 2001.

[206] M. Vogt and V. Kecman, "Active-set methods for support vector machines," *StuddFuzz*, vol. 177, pp. 133–158, 2005.

[207] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Prentice-Hall, 1982.

[208] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover, 9th ed., 1964.

[209] R. Finkel and J. L. Bentley, "Quad trees: A data structure for retrieval on composite keys," *Acta Informatica*, vol. 1, pp. 1–9, 1974.

[210] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002.

[211] B. V. Dasarathy, ed., *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Ieee Computer Society, 1990.

[212] T. Yang and V. Kecman, "Adaptive local hyperplane classification," *Neurocomputing*, vol. 71, no. 13-15, pp. 3001–3004, 2008.

[213] P. Vincent and Y. Bengio, "*K*-local hyperplane and convex distance nearest neighbor algorithms," *Advances in neural information processing systems*, vol. 14, pp. 985–992, 2002.

[214] L. Nanni, "A novel ensemble of classifiers for protein fold recognition," *Neurocomputing*, vol. 69, pp. 2434–2437, 2006.

[215] L. Nanni and A. Lumini, "An ensemble of *k*-local hyperplanes for predicting protein-protein interactions," *Bioinformatics*, vol. 22, pp. 1207–1210, 2006.

[216] C. Jacobs, *Interactive Panoramas: Techniques for Digital Panoramic Photography*. Springer, 2004.

[217] Adobe Systems Incorporated, "Adobe – compare photoshop cs4 versions." [Online], August 2008. Available: `http://www.adobe.com/products/photoshop/` [Accessed: 10 June, 2009].

[218] New House Internet Services, "Photo stitching software 360 degree panorama image software – ptgui." [Online], 2009. Available: `http://www.ptgui.com/` [Accessed 10 June, 2009].

[219] "hugin – panorama photo stitcher." [Online], 2009. Available: `http://hugin.sourceforge.net/` [Accessed: 10 June, 2009].

[220] H. Dersch, "Panorama tools." [Online], November 2001. Available: `http://www.all-in-one.ee/~dersch/` [Accessed: 10 June, 2009].

[221] P. Andrews, *360 Degree Imaging: The Photographers Panoramic Virtual Reality Manual (Photography on the Web)*. Rotovision, 2003.

[222] J. Gulbins and U. Steinmüller, *Art of RAW Conversion: How to Produce Art-Quality Photos with Adobe Photoshop CS2 and Leading RAW Converters*. No Starch Press, 2006.

[223] B. Postle, "Panorama tools." [Online], July 2007. Available: http://panotools.sourceforge.net/ [Accessed: 10 June, 2009].

[224] M. Brown and D. Lowe, "Automatic panoramic image stitching using invariant features," *International Journal of Computer Vision*, vol. 74, pp. 59–73, August 2007.

[225] L. Joskowicz, "Advances in image-guided targeting for keyhole neurosurgery: a survey paper," *Touch Briefings Reports, Future Directions in Surgery 2006*, vol. 2, 2007.

[226] L. Joskowicz, C. Milgrom, A. Simkin, L. Tockus, and Z. Yaniv, "FRACAS: A system for computer-aided image-guided long bone fracture surgery," *Computer Aided Surgery*, vol. 3, pp. 271–288, 1998.

[227] Z. Yaniv and L. Joskowicz, "Precise robot-assisted guide positioning for distal locking of intramedullary nails," *IEEE Transactions on Medical Imaging*, vol. 24, pp. 624–625, May 2005.

[228] R. J. Althof, M. G. J. Wind, and J. T. Dobbins, "Rapid and automatic image registration algorithm with subpixel accuracy," *IEEE Transactions on Medical Imaging*, vol. 16, no. 3, pp. 308–316, 1997.

[229] M. A. Muquit, T. Shibahar, and T. Aoki, "A high-accuracy passive 3d measurement system using phase-based image matching," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E89-A, no. 3, pp. 686–697, 2006.

[230] S. Choi, T. Kim, and W. Yu, "Performance evaluation of ransac family," *Proceedings – 2009 British Machine Vision Conference*, pp. 1–12, 2009.

[231] S. J. Chapman, *MATLAB Programming for Engineers*. Nelson, Thompson Canada Limited, 2004.