Calibration, Recognition, and Shape from Silhouettes of Stones

Keith Forbes

BSc(Eng), University of Cape Town, 1997
MSc(Eng), University of Cape Town, 2000

Thesis presented for the Degree of
Doctor of Philosophy,
Department of Electrical Engineering,
University of Cape Town

June, 2007

Calibration, Recognition, and Shape from Silhouettes of Stones

Keith Forbes

June, 2007

Abstract

Multi-view shape-from-silhouette systems are increasingly used for analysing stones. This thesis presents methods to estimate stone shape and to recognise individual stones from their silhouettes.

Calibration of two image capture setups is investigated. First, a setup consisting of two mirrors and a camera is introduced. Pose and camera internal parameters are inferred from silhouettes alone. Second, the configuration and calibration of a high throughput multi-camera setup is covered.

Multiple silhouette sets of a stone are merged into a single set by inferring relative poses between sets. This is achieved by adjusting pose parameters to maximise geometrical consistency specified by the epipolar tangency constraint. Shape properties (such as volume, flatness, and elongation) are inferred more accurately from the merged silhouette sets than from the original silhouette sets.

Merging is used to recognise individual stones from pairs of silhouette sets captured on different occasions. Merged sets with sufficient geometrical consistency are classified as matches (produced by the same stone), whereas inconsistent sets are classified as mismatches.

Batch matching is determining the one-to-one correspondence between two unordered batches of silhouette sets of the same batch of stones. A probabilistic framework is used to combine recognition by merging (which is slow, but accurate) with the efficiency of computing shape distribution-based dissimilarity values. Two unordered batches of 1200 six-view silhouette sets of uncut gemstones are correctly matched in approximately 68 seconds (using a 3.2 GHz Pentium 4 machine).

An experiment that compares silhouette-based shape estimates with mechanical sieving demonstrates an application using the developed methods. A batch of 494 garnets is sieved 15 times. After each sieving, silhouette sets are captured for sub-batches in each bin. Batch matching is used to determine the 15 sieve bins per stone. Better estimates of repeatability, and better understanding of the variability of the sieving process is obtained than if only histograms (the natural output of sieving) were considered. Silhouette-based sieve emulation is found to be more repeatable than mechanical sieving.

Declaration

This thesis is being presented for the degree of Doctor of Philosophy in the Department of Electrical Engineering at the University of Cape Town. It has not been submitted before for any degree or examination at any university. I confirm that it is my original work. Portions of the work have been published in condensed form in several conference papers [44–47]. I confirm that I am the primary researcher in all instances where work described in this thesis was published under joint authorship.

Keith Forbes
June, 2007

Acknowledgements

I am grateful to those who have played a role in this thesis project. The advice of my supervisor, Fred Nicolls, has been invaluable. My co-supervisor, Gerhard de Jager, has organised interesting work for me, and has given me the opportunity to travel to the four corners of the world. I have benefited enormously by working with my collaborators from industry: Colin Andrew, Ndimi Bodika, Thomas Landgrebe, Garry Morrison, and Anthon Voigt. I have also enjoyed numerous discussions with my officemate, Mathew Price, and other members of UCT's Digital Image Processing Group. The helpful observations and enthusiasm of Pavel Paclík, Michael Taylor, and Jaco Vermaak are also much appreciated. I thank Colin Andrew, George Daniel, Lee-Ann Foley, Gordon Forbes, Marianne Forbes, Henry Foulds, Stephen Haddad, Phillip Milne, John Morkel, Fred Nicolls, and Mathew Price for their help with proofreading and data capture.

Contents

Glossary

AUC	area under the ROC curve
CDF	cumulative density function
CDRH	constant depth rim hull
CIP	cone intersection projection
EMD	earth mover's distance
ET	epipolar tangency
PDF	probability density function
rim	the locus of points on an object's surface that project onto a silhouette outline (also known as a contour generator)
RMS	root mean square
ROC	receiver operating characteristic
run	a single instance of data capture associated with a stone, e.g., five runs of silhouette sets for a batch of stones implies that a silhouette set was captured for each stone in the batch on five different occasions
silhouette set	a set of silhouettes and associated camera internal parameters whose poses are specified in a common reference frame
STD	standard deviation
UIAIA	University of Illinois Aggregate Image Analyser
VAR	variance
VEMH	viewing edge midpoint hull
VH	visual hull

Chapter 1

Introduction

1.1 Overview and Motivation

Silhouette images of a stone provide cues for (1) inferring properties of the imaging system, (2) inferring properties of the 3D shape of the stone, and (3) recognising the stone from previously stored silhouettes. This thesis addresses these inference and recognition problems.

Silhouette images are frequently used in computer vision applications as a simple and robust means for inferring the shape properties of 3D objects. For instance, the *visual hull* is the largest object consistent with a set of silhouettes captured from known viewpoints. Shape-from-silhouette often involves using the visual hull to approximate the 3D shape of the object that produced the silhouettes.

Since, under controlled conditions, foreground and background regions in an image can be distinguished using simple and reliable methods, shape-from-silhouette approaches have become popular in the geosciences for measuring 3D size and shape properties (such as volume, elongation, and flatness) of individual stones or other rigid particles*. Such information is useful for many purposes ranging from value estimation of gemstones to predicting the strength of concrete.

This thesis aims to extend the functionality of silhouette-based particle analysis by developing and analysing new algorithms that are based on recently-developed ideas in the field of computer vision. The application of silhouette-based techniques to stones rather than general objects provides the useful constraint of rigidity: the 3D shape of the imaged object is assumed not to vary over time.

Multiple silhouette views of individual particles provide information that will be used for different purposes:

1. inferring characteristics of the imaging system (camera calibration),

*The term *particle* is commonly used in the geosciences literature to refer to stones, rock fragments, coarse aggregate, mineral grains, pebbles, and so on.

2. inferring particle size and shape properties, and

3. recognising individual particles from their silhouettes.

The following sections briefly overview these three topics.

1.1.1 Camera Calibration

When the imaging characteristics (such as the camera's focal length and principal point) and pose (position and orientation) associated with silhouettes are known, then the silhouette set is *calibrated*. Once the values of calibration parameters are known, it is possible to determine the 3D rays corresponding to 2D points on the silhouette images in a common reference frame. Camera calibration[†] is an important first step for both the recognition and 3D shape analysis algorithms developed in this thesis.

Traditionally, camera calibration has been achieved by observing the image locations of points with known 3D coordinates. Camera parameters are estimated by minimising the difference between observed image points and those predicted by the parameterised camera model.

More recently, there has been interest in self-calibration [40, 58]. Self-calibration solves the calibration problem without using images of marker patterns whose 3D coordinates are known in advance; instead, the images themselves are used (e.g., images of stones in the context of this thesis). Corresponding scene points whose 3D coordinates are initially unknown are used to simultaneously compute both the 3D coordinates and the camera parameters in a process known as bundle adjustment. There has also been activity in self-calibration using silhouettes instead of point correspondences. To render the problem tractable, some form of additional information is incorporated, such as knowledge that the silhouette set is a circular motion sequence. In many approaches to calibration, an initial non-optimal solution is computed using a closed-form solution. The solution is then refined using iterative optimisation. This is the approach taken for calibrating the setups that are used in this work.

This thesis investigates the possibility of self-calibrating camera setups for capturing multiple silhouette views of stones. Two types of setups are used for capturing silhouette sets of particles: a setup consisting of two mirrors and a single camera, and a setup consisting of multiple simultaneously triggered cameras (see Figure 1.1).

The mirror setup provides a simple means for capturing silhouette images of stones using only readily available equipment. Two mirrors are used to create a scene containing five views of an object. The five views are captured in a single image. It will be shown that the silhouettes impose geometrical constraints that can be used to calibrate each silhouette view.

[†]In certain contexts, camera calibration may refer to *radiometric* camera calibration. In this thesis, camera calibration is limited to *geometric* camera calibration: inferring camera poses and internal parameters.

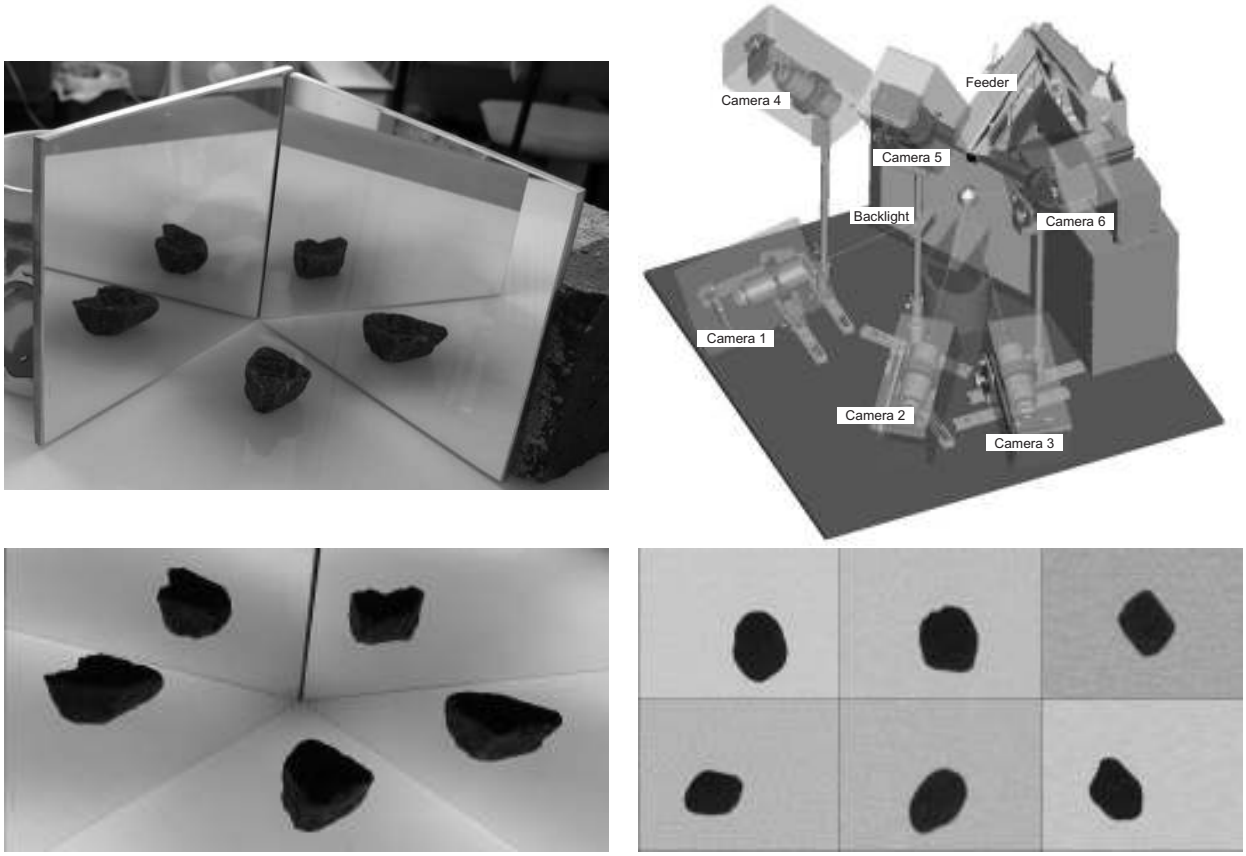


Figure 1.1: The two image capture setups considered in this thesis *(top)*, and examples of corresponding captured images *(bottom)*. The two-mirror setup *(left)* provides a simple low cost means of capturing silhouette sets of stones, whereas the six-camera setup *(right)* enables high throughput imaging.

The multi-camera setup is a high throughput alternative to the mirror setup. It was constructed by a team of engineers from the company that commissioned part of the work described in this thesis. The multi-camera setup is calibrated using images of balls (spheres). The use of ball images aids two aspects of the calibration procedure: (1) forming an initial parameter estimate, and (2) enforcing absolute scale. Since the distance from the cameras to the ball is large with respect to the ball size, the Tomasi-Kanade [129] factorisation method can be used to give a good initial estimate to the calibration parameters. Silhouette centres are used as approximate point correspondences across multiple views. The calibration parameters are then iteratively refined using geometrical constraints imposed by the silhouette boundaries.

1.1.2 Size and Shape Properties

Information about particle size and shape is used in the gem industries, mining, and the geological sciences. The longest, intermediate, and shortest diameter of individual particles are typically recorded, and properties such as flatness, elongation, sphericity, or compactness are derived from the three diameter values. Manually measuring the three diameter values is tedious, time-consuming, and error prone. Machine vision systems

that estimate shape properties from multiple silhouette views therefore provide the potential for saving time and removing the element of human error.

Particle size is also often one of the most important properties of interest. Volume is usually the most desirable measure of size [133], yet sizing of particles has been historically carried out using sieves. Sieves provide only a distribution of sizes of a batch of particles (a histogram), rather than individual per-particle measurements. Machine vision systems can estimate particle volume as well as emulate sieving. Since machine vision systems can consider one stone at a time, different shape properties can be measured for each stone, allowing multi-dimensional distributions to be derived for a batch of stones.

It is *not* the goal of this thesis to analyse the shape of particles with respect to any industrial or environmental process, but rather to investigate algorithms and methods that will provide this means (and other related tools) to particle shape analysts. These include geologists, civil engineers, as well as technicians and researchers from the gem industries, mining, and the geological sciences.

Shape measurements such as particle volume, elongation, or flatness are not the ultimate output of the silhouette-based methods described in this thesis. These are a set of measurements that are often useful to particle shape analysts. Since these shape measurements are commonly-used they are selected as one of the means of quantifying the performance of the silhouette-based methods. For instance, the performance of the new self-calibration methods is quantified in the terms of the accuracy with which these shape properties can be estimated.

It is worth noting that in recent years, particle shape analysts increasingly require 3D shape models of particles (typically triangular mesh models) rather than values of shape properties (e.g., volume, elongation, flatness) that summarise particle shape. The 3D shape models may be used as input to simulations carried out using a finite element analysis software package, for instance. Using 3D mesh models of particles rather than (say) ellipsoids with the same moments up to order two, provides the potential for more accurate simulations.

1.1.3 Recognition

Although the computer vision literature contains an abundance of articles on image-based biometrics applications, such as recognising people from their faces or fingerprints, individual particle recognition does not appear to have received attention in academic literature.

This thesis introduces silhouette-based recognition of individual particles as a research and processing tool for particle analysis. Recognition (or matching) systems are commonly used for *verification* or *identification*.

Identification and verification of stones from silhouette sets is potentially useful for (1) verifying gemstone origin, and (2) tag stone identification:

1. **Gemstone origin verification.** Verification is a potentially useful tool for high value particles such as uncut gemstones. A silhouette set of a stone can be compared with a silhouette set on record to confirm that the two silhouette sets correspond to the same stone.

2. **Tag stone identification.** Gemstone miners often 'spike' mines with gemstones of known mass (so-called *tag stones*). The tag stones are retrieved after processing to audit the performance of the recovery process. Currently, tag stones are recovered by humans who identify them by their mass and by manually comparing them with previously captured photographs. This is a time-consuming procedure.

The methods developed in this thesis are applicable to the problems of gemstone origin verification and tag stone identification. However, the main recognition task addressed by this thesis is the one-to-one matching of an unordered batch of stones captured on two separate occasions: a *square assignment* problem. The problem is potentially more difficult than identification or verification, since each of n silhouette sets in the first run[‡] must be matched to one of the n silhouette sets in the second run. The matching can be specified by an $n \times n$ permutation matrix in which each element is either one or zero (indicating match or mismatch), and each row and each column sums to one. The nature of the batch matching problem is illustrated in Figure 1.2.

The ability to match up silhouette sets of an unordered batch of stones across two runs (batch matching) is potentially useful for several applications:

1. Batch matching can be used to measure the repeatability or accuracy of a stone classifier. The classifier could be, for example, a mechanical classifier such as a sieve which classifies stones into different sieve bins according to size, or a human classifier, such as a person who sorts gemstones into different piles according to colour. (Piling the stones enables efficient sorting, since there are far fewer classes than stones.) Stones are passed through the multi-camera setup after class labels have been assigned. (To keep a record of class labels for each silhouette set, it is easiest to pass the stones through the camera setup in sub-batches of the same class label.) Batch matching will determine the different class labels that each stone has received after being classified on multiple occasions.

2. Batches of stones are used by various laboratories for research purposes. The stones are often stored in trays with one stone per compartment so that each stone can be uniquely identified. This means of storage can become impractical for large batches of stones (of more than about 100 stones). Properties of the individual stones (such as volume, density, or hardness) may be measured and recorded for the individual stones at different times. With batch matching technology, the stones need not be separately stored as the matching process can be used to reconcile the information.

3. This thesis will demonstrate how to merge several silhouette sets of the same particle into a single large silhouette set in which all silhouettes are specified in a common reference frame. More accurate estimates of the 3D particle shape can be made from the merged set than from any of the individual

[‡]In this thesis, the term *run* is used to refer to a batch of silhouette sets in which one silhouette set is captured for each stone.

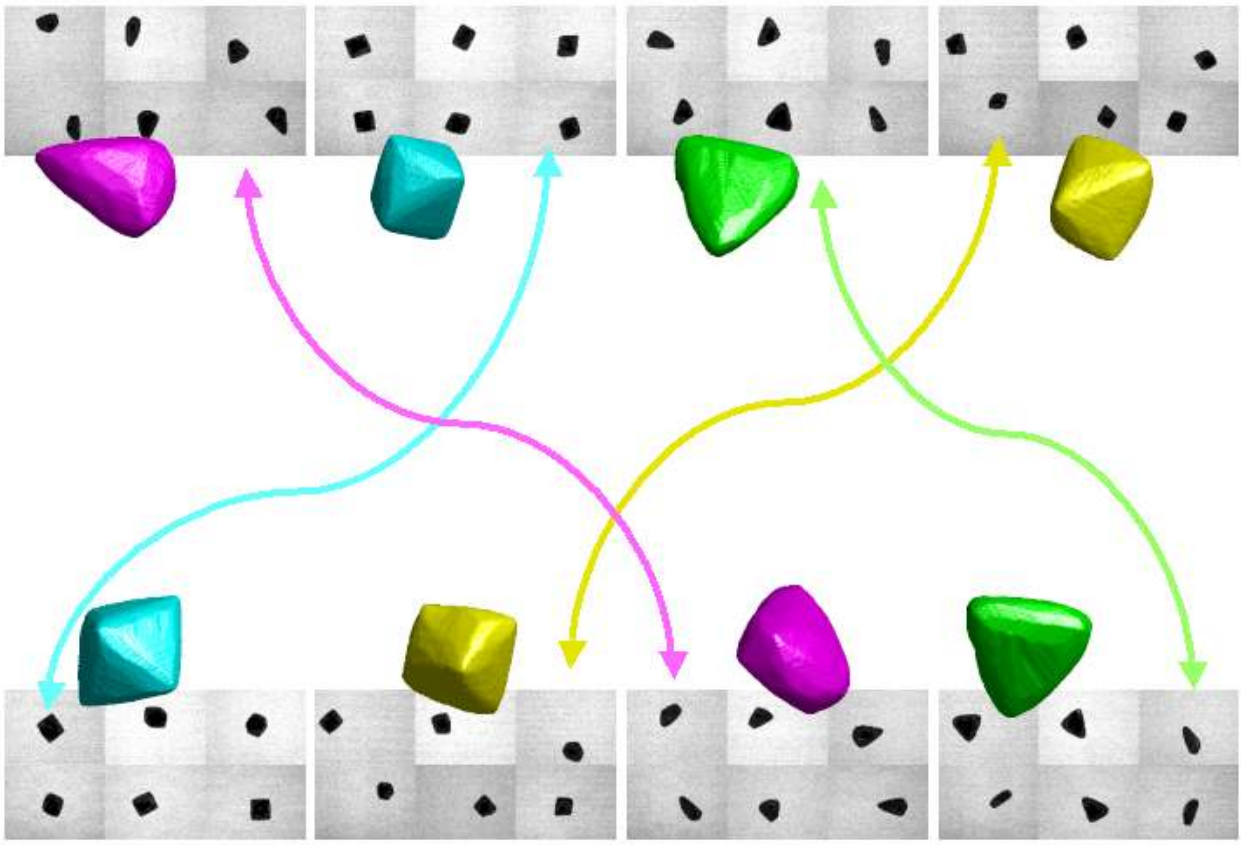


Figure 1.2: The batch matching problem: each six-view silhouette set in the first run *(top row)* must be matched to the corresponding silhouette set in the second run *(bottom row)* using only the silhouette images and corresponding camera calibration information. Coloured arrows show the desired unknown correspondences: pairs of silhouette sets generated by the same stone. The problem is difficult because (1) the stones are ordered arbitrarily, and (2) the stones are oriented arbitrarily. The efficient batch matching algorithm developed in this thesis rapidly estimates the 3D shape of a stone from its silhouettes to identify likely matches. Pairs of silhouette sets that are geometrically consistent with being produced by the same stone are then sought. This illustration shows a small data set of $n = 4$ stones; in practice, data sets will contain hundreds or possibly thousands of stones.

sets. Batch matching allows an unordered batch of stones to be passed through the multi-camera setup several times so that merged silhouette sets can be formed for each stone. Passing unordered batches of stones through the multi-camera setup is quicker than passing individual stones through one at a time.

1.2 Research Objectives

The principal objective of this thesis is to develop new algorithms to solve the problems of self-calibration, recognition, and particle shape analysis using multi-view silhouette sets of particles.

A portion of the work presented in this thesis was carried out as part of a project commissioned by a company that wishes to remain anonymous. The nature of this company's specific uses for the developed methods lie outside the scope of this thesis. However, the methods are by no means applicable only to gemstones. Data sets of uncut gemstones (in addition to garnets and gravel) were used as test sets in this work as these were made available by the commissioning company. Indeed, many of the methods developed here have broader application scope than particle analysis, and can be applied to other objects. Three-dimensional shape reconstruction for multimedia content creation is an example of an application that will benefit from some of the methods developed in this thesis. For cases in which the methods are applicable to general objects, experiments and examples will therefore be given for objects other than stones. Particle analysis, however, is the unifying theme for the topics covered.

Within the topic of *shape and size*, the aim is to develop algorithms for estimating properties that are commonly used by particle shape analysts. These methods are to then to be used in conjunction with *calibration* and *recognition* methods to quantify the accuracy and repeatability of such systems.

Systems that compute the 3D shape of particles must trade off the desirable characteristics of accuracy, throughput and affordability (in terms of monetary cost). This thesis investigates two multi-camera setups: (1) a highly affordable setup that uses two mirrors to generate multiple views, and (2) a high-throughput system that uses six simultaneously triggered cameras. The goal is to develop separate *self-calibration* algorithms for the two setups.

A further goal within the topic of *calibration* is to demonstrate that multiple silhouette sets of a particle can be merged into a single large silhouette set in which all silhouettes are specified in a common reference frame.

The major objective within the *recognition* component of this thesis is to develop an *efficient* method for solving the batch matching problem (as illustrated in Figure 1.2). (In this thesis efficiency will always refer to the speed of execution, as opposed to, for example, memory efficiency.) To achieve this objective, it is useful to break it down into several components:

1. The aim of the first component is to determine, as efficiently as possible and preferably without error, whether a pair of silhouette sets corresponds to the same stone (a match) or not (a mismatch).

2. The aim of the second component is to develop a rapid means of identifying candidate matches by assigning a dissimilarity score to silhouette set pairs.

3. The aim of the third component is to combine the first two components to create an algorithm that makes use of the accuracy of the first component and the efficiency of the second component to solve the batch matching problem.

The final objective of this thesis is to demonstrate the use of the *calibration*, *recognition*, and *shape analysis* tools by showing how they can be used together to solve a practical problem: estimating the repeatability of mechanical sieves and comparing the repeatability with a machine vision emulation of sieving.

1.3 Contributions

The most important novel components of this thesis are the following:

1. The analysis of viewing edges is introduced as an alternative to the visual hull for efficiently estimating 3D shape properties of stones. Viewing edge midpoints are demonstrated to provide more accurate estimates of 3D properties such as caliper diameter measurements (longest, shortest and intermediate diameters). The viewing edges are demonstrated to impose geometrical constraints from which the upper and lower bounds of a stone's longest and shortest diameters can be computed from its silhouette set.

2. A novel, low cost mirror-based setup for capturing multiple silhouette views is described, and algorithms for self-calibration are developed. The method provides an accessible and affordable method for 3D shape reconstruction of stones. The method is not limited to 3D reconstruction of stones and has been applied to objects other than stones (e.g., toy animals). It can be used as a simple method for creating 3D multimedia content for people who do not have access to expensive equipment.

3. Calibration of a simultaneously-triggered six-camera setup is achieved by combining two existing approaches to calibration. Initial parameter estimates are determined using approximate point correspondences and the Tomasi-Kanade method [129]. The initial parameter estimates are then refined by minimising a cost function based on the outer epipolar tangents [138].

4. A new pose optimisation method for merging several silhouette sets of the same object into a large silhouette set is developed. The method allows one to generate an arbitrarily large number of silhouettes of an object in a common reference frame using an image capture setup that generates a small number of views. A merged silhouette set provides a more accurate 3D reconstruction and tighter constraints on 3D shape than any of the original silhouettes sets from which it was formed.

5. The use of the residual error associated with a merged pair of silhouette sets is demonstrated to be an effective indicator of whether the pair corresponds to two silhouette sets of the same stone (a match) or to two silhouette sets of two different stones (a mismatch).

6. An existing shape matching method [103] based on shape distributions is adapted to create a rapidly computable method for assigning a measure of dissimilarity between two silhouette sets. A batch matching algorithm is then developed to use both the rapidly computable dissimilarity measures and the pose optimisation method, to match two runs of silhouette sets of an unordered batch of stones. The important feature of this batch matching algorithm is its efficiency: a test set of 1200 stones is correctly matched across two runs in approximately 68 seconds on a 3.2 GHz Pentium 4 machine.

1.4 Thesis Organisation

The remainder of this thesis is organised as follows.

Chapter 2 provides a short historical overview of particle shape analysis in the geosciences. This is provided because (1) computer vision researchers are unlikely to be familiar with the geosciences literature on this topic, and (2) this presents a historical background of the work that this thesis extends by developing new algorithms and methods. First, definitions of shape properties that are of interest are covered, and some examples of their uses are given to demonstrate that particle shape analysis is a broad field with diverse goals. Next, silhouette-based machine vision systems that have been designed to measure particle shape properties are covered.

Chapter 3 introduces background theory on the geometry of silhouette sets that will be used to develop the methods described in later chapters. The concept of *silhouette consistency* is introduced and two methods that will be used throughout the thesis are described: (1) a new silhouette-consistent estimate of 3D shape, the viewing edge midpoint hull (VEMH), which will be used for estimating 3D shape properties and as a component of the matching process, and (2) an existing measure of silhouette consistency based on the epipolar tangency constraint that will be used for calibration and for matching.

Chapter 4 describes a novel low cost image capture setup based on two plane mirrors. The chapter describes how the camera parameters associated with silhouette views of an object can be computed from the silhouettes alone: there is no need for calibration markers. Since the method can be used to reconstruct the 3D shape of a broader class of objects than stones, results are demonstrated using both stones and other objects.

Chapter 5 covers the geometric configuration and calibration of a high-throughput alternative to the image capture setup described in Chapter 4. Heuristics are introduced that are used to determine the positioning of the cameras. A calibration routine based on silhouette images of a ball is described. Balls of known dimension allow scale to be enforced, and ball images allow an approximation based on Tomasi-Kanade factorisation method to be used for forming initial parameter estimates.

Chapter 6 shows how silhouette sets of the same particle captured in different poses can be merged into a single large silhouette set by minimising the degree of geometrical inconsistency across the silhouettes. Results computed using both stones and other objects are presented. The method is applied to objects captured with the mirror-based setup described in Chapter 4 as well as the six-camera setup described in Chapter 5.

Chapter 7 shows how the pose optimisation and associated error described in the previous chapter is used for matching or *recognising* particles from their silhouettes. Other measures of silhouette consistency are covered. The methods are applied to data sets of stones captured using the mirror setup and the six-camera setup.

Chapter 8 develops a method of rapidly computing a measure of dissimilarity between two silhouette sets. The method is based on the shape distributions of Osada et al. [103], but is modified to improve efficiency in the context of silhouette sets of stones. This includes using the VEMH introduced in Chapter 3 as an estimate of the 3D convex hull of the stone.

Chapter 9 describes a method for efficiently finding the one-to-one correspondences between silhouette sets from two runs of the same batch of stones. The method makes use of the efficiency of the matching approach described in Chapter 8 together with the accuracy of the slower method described in Chapter 7. A probabilistic framework is used to achieve efficiency: a likelihood ratio (indicating the likelihood of being a match) is associated with each silhouette set pairing across the two runs. Likelihood ratios are updated using Bayes's rule as new information is added from the results of pose optimisations. A greedy algorithm is shown to provide a tractable solution that produces excellent results in terms of running time and accuracy.

Chapter 10 describes an experiment that makes use of the main ideas developed in this thesis: batch matching, estimating shape properties, and calibration. The experiment estimates the repeatability of mechanical sieving by determining which stones fall into which bins over multiple runs of sieving. Knowing the sieve bins associated with each particle allows repeatability to be estimated more accurately than if only the bin counts were known for each run. The repeatability of the mechanical sieving process is compared with a machine vision emulation in which sieve bin classification is computed using silhouette sets.

Chapter 11 concludes the thesis by reviewing the main contributions and summarising the work. Ideas for future work are identified.

Chapter 2

An Overview of Particle Shape Analysis

2.1 Introduction

This chapter provides a historical overview of particle shape and size analysis that is drawn mainly from the geosciences literature. It is shown how interest has grown in using silhouette-based machine vision methods to quantify particle shape properties. Initially, single-view systems were used, and more recently there has been interest in multi-view systems.

The content of this chapter is not required for understanding the methods developed in this thesis. Readers who are not interested in a historical overview may wish to skip this chapter, and continue reading Chapter 3 on page 25.

2.2 Quantifying Particle Shape

Particle shape analysts in a range of different fields (for example, geomorphologists, civil engineers, process engineers, hydrologists) are interested in summarising the size and shape (sometimes termed ‘form’) of particles using a small number of features.

Volume is usually the preferred measure of size [133]. The volume of a particle can be used to estimate weight (if density is known), or to estimate density (if weight is known). Size distributions play an important role in determining particle packing and porosity characteristics in asphalt mixes [109]. In the gem industry, individual particle volume is closely (and nonlinearly) related to the monetary value of each gemstone. Historically, sieving has been used to characterise the size distributions of large batches of particles, because of the high throughput that can be achieved.

Although there are some differences in their precise definitions, the *long*, *intermediate*, and *short diameters* of a particle are frequently used to summarise its shape. These three diameters are sometimes referred to as the *a*, *b*, and *c* diameters, respectively [70]. Often *a* is defined as the longest diameter (*a* is termed simply the *diameter* by computational geometers), *c* is the shortest diameter (termed the *width* by computational geometers), and *b* is the diameter measured in the direction that is perpendicular to the directions corresponding to *a* and *c* [131]. Note that these diameters are *caliper diameters*; in other words, they represent distances between parallel plane pairs that are tangential to the particle. Different variations on the definitions include measuring the caliper diameters along the principal directions (as determined by the inertia tensor) [4, 127], and requiring the long diameter to be measured perpendicular to the shortest diameter [71], or requiring the short diameter to be measured perpendicular to the longest diameter [41].

The *a*, *b*, and *c* diameters are measured in various ways. Manual measurements include the use of a sliding rod caliper [70], Vernier calipers [59], and a ruler [83]. Automated methods include the use of 3D laser scanning [71], X-ray tomography [82], and silhouette-based machine vision [87].

The three diameter values are frequently used to provide dimensionless quantifications of particle *elongation* and *flatness*. Two common formulations specify elongation as the ratio a/b, and flatness as the ratio b/c [6].

A measure of sphericity (the degree of compactness) is also often derived from the three parameters. Krumbein's commonly used definition [70] is

$$\text{sphericity} = \sqrt[3]{\frac{bc}{a^2}}. \tag{2.1}$$

Zingg's diagram [144] is a popular means for classifying particles into one of four shape categories and for visualising the distribution of shape for a batch of particles (see Figure 2.1). Zingg classified particles into the classes oblate (disk-shaped), spherical (compact), bladed (triaxial) and prolate (rod-shaped), based on the ratios b/a and c/b. For each particle, the *y*-coordinate of its data point is b/a, and the *x*-coordinate is c/b. Using a threshold of $2/3$, the data points that lie in the top left quadrant are oblate; the lower left are bladed; the upper right are spherical; and the lower right are prolate. Hyperbolic contour lines can be plotted on the standard chart so that sphericity values can be read off.

It is interesting to note Joshi and Bajcsy's discussion (within the field of linguistics) on the ways in which humans interpret shape [66]. The terms 'flat', 'elongated' and 'round' are listed as some of the few non-template-based terms that humans tend to use to describe 3D shapes. Joshi and Bajcsy's 'roundness' refers to what is termed 'sphericity' in the geosciences literature. People prefer template-based descriptions such as 'star-like'.

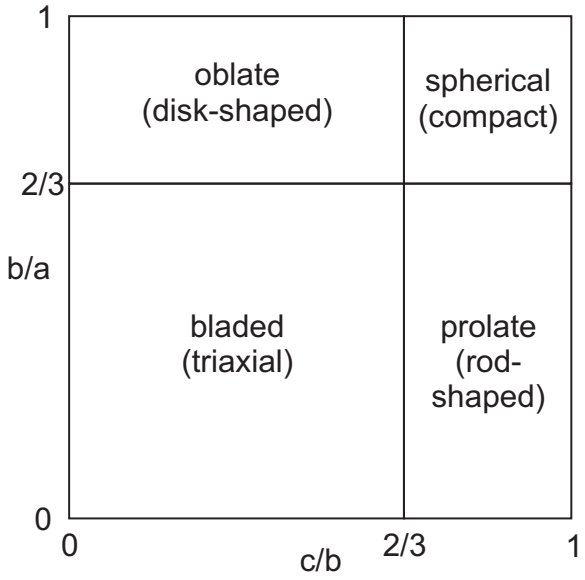


Figure 2.1: Zingg's diagram [144] for classifying particle shapes from the *a*, *b*, and *c* diameters.

2.3 A Range of Analyses of Particle Shape

To provide an indication of the wide range of subjects that make use of measurements of individual particle shape, this section provides a brief description of a few of the studies described in the literature. In many cases, it appears that these types of studies will benefit from modern silhouette-based machine vision methods for quantifying particle shape.

2.3.1 Ice-Rafted Pebbles

Hassler and Cowan [59] collected 331 pebbles from drill sites on the Antarctic Peninsula. The long, intermediate, and short axes were manually measured using Vernier calipers. Together with other evidence, the shape measurements were used to support the hypothesis that the pebbles had been transported as supraglacial debris.

2.3.2 Alluvial Gravel

Lindsey and Shary [83] assessed alluvial gravel deposits by measuring the long, intermediate and short diameters of 150 pebbles from three locations along the South Platte River in Colorado. The measurements were performed manually using a ruler. They show that the proportion of equidimensional particles increases downstream. The study aims to predict the downstream limit of gravel production (mining) and of post-mining land uses.

2.3.3 Gold Grains

Wierchowiec [136] uses Zingg diagrams to visualise the shapes of gold grains from different sources. Gold grains from preglacial and alluvial deposits are observed to be bladed, whereas those from piedmont fan sediments tend to be oblate. Factors such as hammering and folding during transport, and reflattening after folding account for the variations in shape.

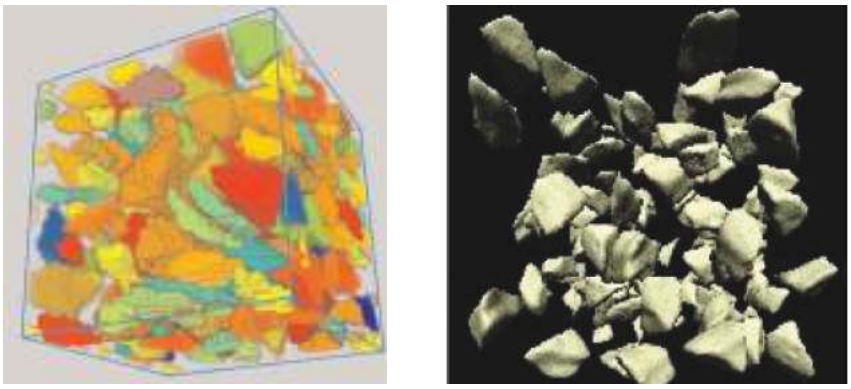
2.3.4 Anthropogenic Fragment Redistribution

Nyssen et al. [100] monitored rock fragment transportation in the stepped mountains of Ethiopia over a four year period. Limestone rocks were used as tracers since the existing rocks were basalt and sandstone (painted rocks were not used since they may have been picked up by shepherds). The long, intermediate, and short diameters were used to replace basalt and sandstone rocks with limestone tracers of approximately the same shape. The authors show that the degree of tracer transportation over the years is related to the degree of over-grazing by livestock and conclude that livestock trampling appears to be an important geomorphic process.

2.3.5 Estimating Particle Properties with Computer Simulations

Computer simulations of a large number of particles often make use of simplified models of particle shape. For instance, a sample of particles may be modelled using ellipsoids with the same volume, flatness, and elongation. Sims et al. [119] use ellipsoidal models of the aggregate particles in concrete to investigate strain rate. They demonstrate that particle flatness and elongation play an important role in determining concrete viscosity.

Rather than using simple ellipsoidal models, Bullard, Garboczi, and coworkers [20, 52] take advantage of the power of modern desktop computers to model concrete using 3D shapes based on real aggregate particles (see Figure 2.2). Using particle shape based on real particles rather than simpler ellipsoidal models has the potential to provide more accurate simulations. The 3D shape of real aggregate can be determined using X-ray computed tomography, and then be included in their computer simulations. The authors aim to use computer modelling to replace empirical testing for predicting concrete properties such as the degree of hydration, pore percolation, diffusivity, and yield stress viscosity. Simulation predictions of certain properties such as elastic moduli have been shown to agree closely with values obtained in real experiments.



(a) (b)

Figure 2.2: Computer simulations of concrete using 3D shapes based on real particles: (a) modelling concrete flow (picture from Bullard et al. [20]), (b) simulation of coarse aggregate in a mortar matrix flowing under mixing forces (picture from Garboczi et al. [52]).

2.4 Single View Silhouette-Based Particle Analysis

Other than shape, particle shape analysts are also interested in *angularity* and *roughness*. These properties are not addressed by this thesis, but are mentioned here since the first attempts at image-based particle analysis were attempts to measure these properties. Before image-based methods were used, angularity was determined by comparing particles with Krumbein's standard chart [70].

Schwarcz and Shane [117] use Fourier coefficients of the boundary of a particle projection (Fourier descriptors) to derive several procedures for quantifying angularity. First, they describe how measurements of sphericity and angularity might be derived from a 3D model of the stone. They point out that 3D models are rarely available and proceed to present their measurements that are based on computing the Fourier descriptors of a 2D projection of the stone. A measurement for sphericity is given as the mean squared deviation between the silhouette boundary and the circular boundary defined by the first Fourier descriptor. The authors investigate several methods for measuring angularity based on Fourier descriptors. One such method involves determining the number of Fourier coefficients that are required to reconstruct the boundary so that it fits the original boundary to within a specified tolerance. This type of measurement varies according to surface *roughness* as opposed to angularity.

Ehrlich and Weinberg [37] show how Fourier descriptors can be used to discriminate grain differences arising from geographic, stratigraphic, and process factors. Plots of the average values of the first ten harmonics are used to discriminate between grains from three different geographical regions with a high success rate. The same method is used to show how grain shape varies according to its position in the soil profile. Various means for defining roughness coefficients based on summing a range of Fourier coefficients are also suggested.

Diepenbroek et al. [31] give yet another definition of roundness using Fourier descriptors. They discard the first two Fourier coefficients, which describe an ellipse, and form a weighted sum of the remaining coefficients, with the higher order coefficients receiving greater weight. The method was used to detect changes in roundness of gravel clasts being transported down mountain rivers. Changes over distances as small as 7 km could be detected. Drevin et al. [32, 33] investigate means other than Fourier descriptors for determining particle roundness. Wavelet and granulometric methods are considered. They show that their methods and the method of Diepenbroek et al. both produce results that correlate well with the values indicated by Krumbein's chart.

2.5 Multi-View Silhouette-Based Particle Analysis

Since using single silhouette views results in the loss of information about the third dimension, there has recently been research directed towards *multi-view* silhouette-based particle analysis. The goal of these methods is to extract information about the three-dimensional shape of individual particles from multiple silhouette views.

It is the objective of this thesis to extend this line of research by designing algorithms that are based on the shape-from-silhouette ideas that have been developed in the field of computer vision.

2.5.1 Multiple Views from a Single Camera

Several groups of researchers have considered means for obtaining multiple silhouette views of a particle using a single camera. Typically this involves moving the particle and capturing images at different instants in time (although Chapter 4 of this thesis introduces a method in which different silhouette views are captured simultaneously using mirrors). Using a single camera and moving the particle has the advantage of lower monetary expense than a multiple camera setup, but this comes at the cost of requiring more time to capture the images.

Motivated by the high monetary expense of laser scanning and tomographic methods, Taylor [126] and Lau [72] investigate the use of silhouettes as a cheaper alternative to quantify particle shape. A setup consisting of a turntable with two orthogonal axes of rotation (see Figure 2.3) is used to view a rock from any direction. Individual rocks are glued to a rod, and images are captured from well-distributed viewpoints. A ball of known diameter is used to calibrate the setup. The calibration simply provides a conversion from pixel units to millimetres (and therefore implicitly assumes that depth variation is sufficiently small to have negligible effect on scale). Taylor and Lau are aware of the visual hull concept, as it is noted that silhouettes place a restriction on the volume of space that contains the object, and a computer vision paper of Laurentini [74] is cited. However, they decide to limit their initial investigations to estimating volume using silhouette area. Silhouette area averaged over 13 views is computed for 126 rocks (crushed granite and

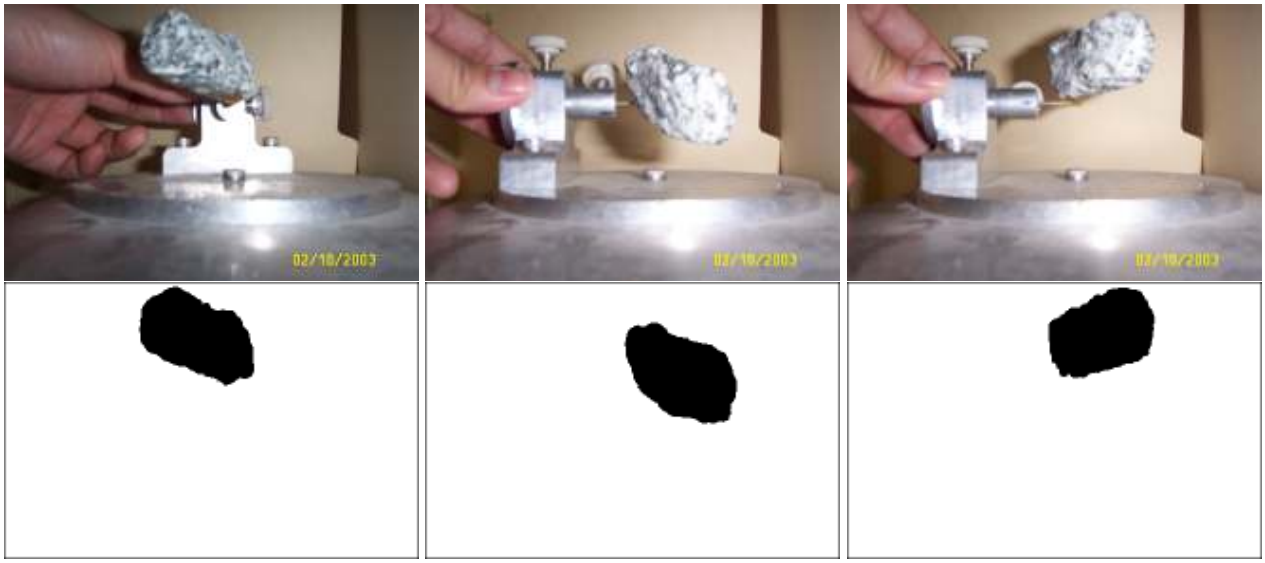


Figure 2.3: Three images from a sequence captured using a turntable device for rotating individual rocks about two different axes *(top row)*, and the corresponding manually segmented silhouettes *(bottom row)* (pictures from Lau [72]). The images were used to investigate volume prediction from multiple silhouette views.

rounded conglomerate rocks from a river bed). Plots of average silhouette area versus weight show a high degree of correlation.

Chen et al. [25] measure the short, intermediate, and long diameters of a sample of aggregate particles by attaching the particles to a clear plastic tray with two perpendicular faces. The particles are imaged from two perpendicular directions by rolling the tray onto each of the two faces. Diameter values are measured from the silhouette images. Elongated and flat particles are demonstrated to produce hot-mix asphalt with lower compactibility and higher breakage than compact particles. The use of a tray with the perpendicular faces for imaging stones from perpendicular directions is also described by Frost and Lai [50].

Fernlund [41] describes a method for capturing multiple views in which particles are moved by hand. Two views are captured for the particles: a side-on view and a top view. To capture the side-on view, the particles are manually positioned on a luminous background in a stable position so that their maximum projected area is observed by an overhead camera. To capture the top view, the particles are manually positioned in an upright position in a bed of luminous beads and sand. The bed allows the particles to be placed in a stable position with their longest axes parallel to the viewing direction. The principal benefit of the method is its low cost. Multiple particle silhouettes are captured in each image. Longest and intermediate diameters are measured from the side-on image, and shortest and intermediate diameters are measured from the top-view image. To find the corresponding silhouette pairs for each particle, the silhouettes are sorted by intermediate diameter value, which is assumed to be the same across the two images. Although it is acknowledged that this assumption may not hold in all cases, the method is reported to provide results that correlate well with manual measurements.

Commercial shape-from-silhouette systems for characterising gemstone shape are produced by Sarin, an Israeli company, and by Octonus, a Russian company [78]. These systems build 3D visual hull models of individual rough gemstones to aid gemstone cutters. Multiple silhouette images of the rough gemstone are captured by a single camera as the stone is rotated on a turntable. The rotation of a stone takes approximately 25 seconds. An optional laser range-finder can also be used to build 3D models of rough gemstones with concavities.

2.5.2 Multiple Views from Multiple Cameras

Multiple simultaneously triggered cameras provide the potential of greater throughput than multi-view single camera setups. For this reason, various multi-camera setups have been designed over the last decade. The two most prominent multi-camera silhouette-based particle analysis systems described in the academic literature are the WipFrag system and the University of Illinois Aggregate Image Analyser.

WipFrag

The WipFrag system was developed at the University of Missouri-Rolla by Maerz et al. [86, 87]. It consists of two orthogonally mounted cameras that simultaneously image individual particles (see Figure 2.4).

Figure 2.4: The WipFrag system. (Picture from Al Rousan [111])

The WipFrag system is used to estimate the aspect ratios and volumes of individual particles from silhouette images. Elongation, flatness, and volumes are derived from measurements of length, width and height. Length and width are measured from the top-view image and height is measured from the other. Length is the longest caliper diameter of the silhouette, and width is the caliper diameter measured perpendicular to the length. Height is measured from the side-view image. It is the caliper diameter measured in the direction of the top-view camera. The measured lengths are reordered if necessary so that length is longer than width, which is in turn longer than height.

Aspect ratio is the ratio of length to height. Volume is estimated with the following experimentally determined equation:

$$\text{volume} = 0.8 \times \text{length} \times \text{width} \times \text{height}. \tag{2.2}$$

The vision-based methods were compared with manual caliper measurements of aspect ratio [86], whereas results for volume estimation are not shown. The vision-based methods are found to be close to the manual measurements in most cases; closeness is not, however, quantified.

The University of Illinois Aggregate Image Analyser

The University of Illinois Aggregate Image Analyser (UIAIA) is the most sophisticated system for estimating stone shape from silhouettes that is described in the academic literature. It is the only setup that creates a 3D model of each stone from multiple silhouettes. The 3D models are used for volume estimation.

The UIAIA setup consists of three orthogonally mounted cameras. A conveyor system presents the stones to the cameras (see Figure 2.5). Images are captured as each particle triggers a motion sensor. Explicit

Figure 2.5: The UIAIA: a three-camera setup at the University of Illinois (picture from Rao [109]).

calibration of camera poses is not carried out. Rather, the cameras are orthogonally positioned, and images of spheres are used to ensure that the effective scale factors are the same across the three views. This approach implicitly assumes that the depth variation of each stone is sufficiently small with respect to the distances to the cameras that perspective distortion can be ignored. The only explicit calibration that is carried out is to use images of a sphere to determine the scale factor (that is, a mapping of pixels to millimetres).

Volume estimates are made by computing a three-view visual hull of the stone. However, the term visual hull is not used, and the method seems to have been developed independently from (and without reference to) shape-from-silhouette approaches described in the computer vision literature. To compute the visual hull,

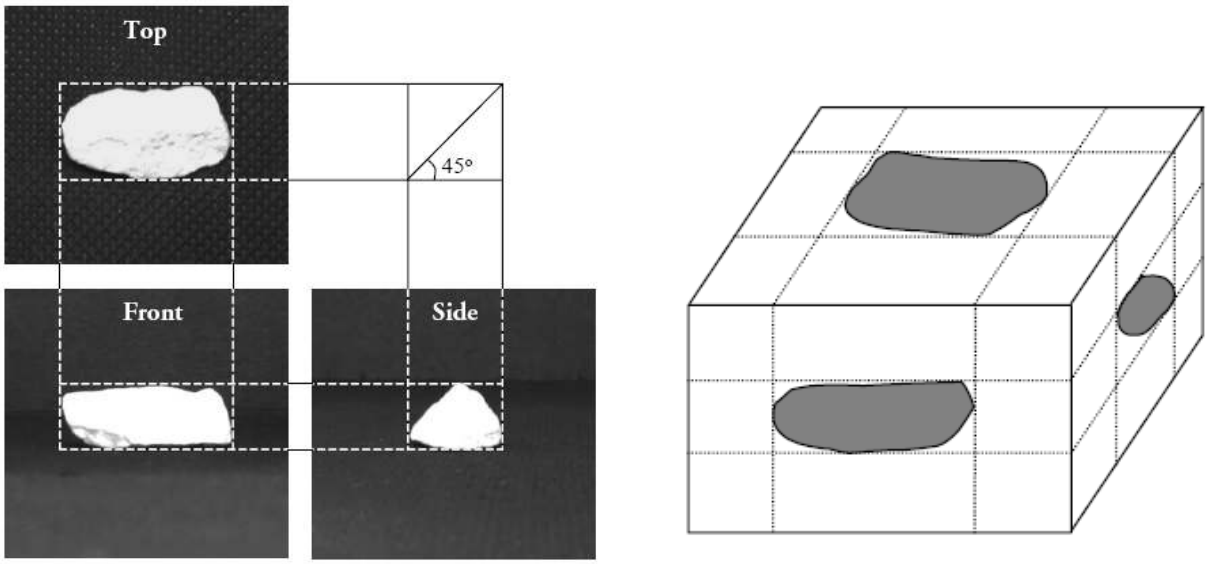


Figure 2.6: Computing the visual hull from three orthogonal views (picture from Rao [109]).

voxels that do not project onto the silhouette foreground in all three images are removed, leaving an estimate of the 3D shape of the stone (see Figure 2.6).

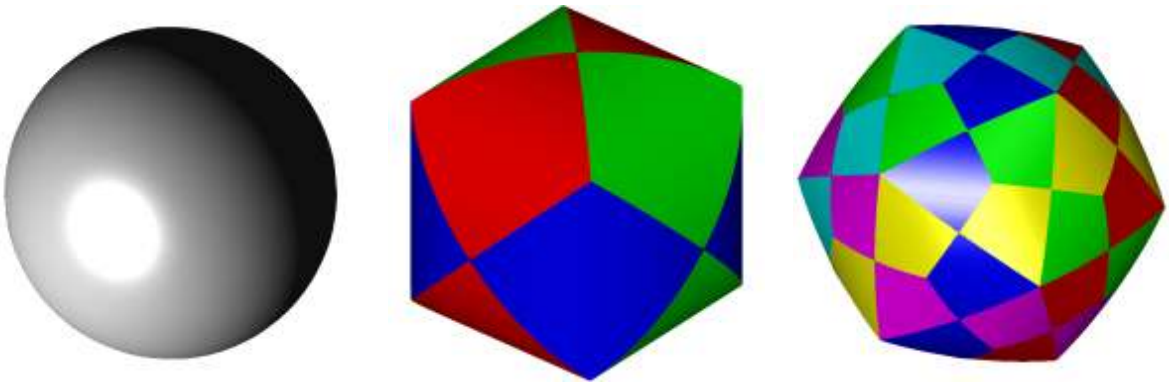
Volume computation is applied to four spheres of known volume to test the accuracy of the voxel-based visual hull volume estimates [109]. The largest sphere's computed visual hull volume ranges from 101.38% to 102.84% of the sphere volume (5 trials), whereas the smallest sphere's computed visual hull volume ranges from 105.77% to 107.60% of the sphere volume (5 trials).

Since the exact 3-view visual hull of a sphere (from three orthogonal orthographic views) is 11.9% larger than the sphere (see Figure 2.7), it is unsurprising that the visual hull volume produces an overestimate when used as an approximation of the volume of the imaged object. Inaccuracies in the assumed orientation of the cameras and image noise tend to result in a computed visual hull that is smaller than the exact visual hull, since visual hull voxels are required to project to foreground region in *all* views. Because of this, a real setup can be expected to produce values lower than 11.9%.

The description of the UIAIA experiments makes no mention that the 3-view visual hull volume is expected to be larger than the sphere. Spatial quantisation error is given as the reason that the smallest sphere's results (which are closest to the noise-free ideal of 11.9%) correspond to the greatest error when using visual hull volume as an estimate of the volume of the imaged object.

In a further experiment, visual hull volumes are used as an estimate of stone volume for 50 pieces of aggregate. Ground truth values are obtained by weighing the stones and using the known density values to compute volume. A mean absolute percentage error of 8.74% is reported.

The authors cite the inability of silhouettes to capture information about concavities in the stone as the



(a) (b) (c)

Figure 2.7: (a) A sphere, (b) a 3-view visual hull of the sphere, and (c) a 6-view-visual hull of the sphere. The 3-view visual hull is computed from three orthogonal views (the UIAIA camera configuration), and its volume is 11.9% larger than the sphere. The 6-view visual hull is computed using the camera configuration of the high throughput, six-camera setup considered in this thesis (each camera looks onto one of six parallel face pairs of a regular dodecahedron; details are given in Chapter 5), and its volume is 4.5% larger than the sphere. (Visual hull surface regions are coloured to correspond to the camera view for which the surface region projects to the silhouette outline.)

reason for consistent overestimation of volume by the visual hull. Curiously, the tendency for a visual hull computed from a finite number of views (three views in the case of the UIAIA) to be larger than the imaged object (whether convex or nonconvex) is not mentioned as a possible cause for the consistent overestimates observed in both the experiments with stones and with spheres.

The UIAIA is also used to estimate the ratio a/c (termed the flat and elongated ratio), where a is the longest diameter of the stone and c is the shortest diameter that is perpendicular to the longest diameter. The longest diameter, and the diameter perpendicular to the longest diameter is computed for all three views. The largest of these six diameter values is used to estimate a, and the smallest is used to estimate c. Approximately one thousand aggregate particles were classified into three classes of a/c: smaller than 3:1, 3:1–5:1, and greater than 5:1. This was done both manually with a caliper device, and using the UIAIA. The UIAIA is found to produce more repeatable results than the manual measurements in terms of the proportion of particles in each class by weight. The class proportions obtained by the UIAIA are found to be in good agreement with the manually determined classes, but are not quantified.

The UIAIA is also used to emulate sieving. The smallest of the longest diameters from each 3-view image set is used to predict the sieve class for each particle. The aggregate particles are sieved into five sieve classes using square-aperture sieves. Plots of histograms from UIAIA sieve emulations are compared with those obtained from manual sieving and are found to match closely.

The UIAIA has also been used to approximate local shape properties such as the angularity and texture of stones.

2.6 Recognising Individual Particles

The academic literature makes few references to the problem of recognising individual particles. The existing references either are speculative and do not provide quantitative evaluations of proposed methods, or simply describe the need for particle recognition rather than proposing solutions to the problem. Note, however, that object recognition is one of the main goals of computer vision, and a wealth of literature exists on the subject.

In a theoretical paper [127], Taylor proposes to describe the 3D shape of particles by their principal moments. He states that it is extremely unlikely for two particles to be congruent, and proposes that the principal moments can be used to uniquely identify individual particles. Several shapes are demonstrated to have the same sieve size, yet the shapes are uniquely identifiable by their moments. The author aims to test his proposed formulation using real particles and tomographic shape reconstruction in the future.

In a later paper [126], Taylor points out that it is not easy to "confirm that one has selected a given particle from a group" and proposes that moments of two voxel representations are used to determine whether or not the representations correspond to the same particle. For irregular particles, each voxel representation will have a unique shape if sufficient voxels are used. Taylor and his coworkers currently identify individual rocks by imprinting a number on each rock. Note that this approach is impractical for smaller stones (such as the garnets and gemstones used in this thesis), and requires manual identification, whereas this thesis provides methods to enable automated identification.

Fernlund [41] mentions identifying particles from silhouettes using the intermediate diameter (as described in Section 2.5.1). This is done to reconcile pairs of silhouettes of the same particle captured from approximately orthogonal viewpoints. However, no quantitative assessment of the accuracy of this approach is given.

2.7 Reconstruction Techniques Not Based on Silhouettes

Shape-from-silhouette methods are by no means the only methods that have been considered for determining the 3D shape characteristics of particles. Shape-from-silhouette methods are typically favoured over competing methods because of their low monetary cost, simplicity, and robustness. Other methods may achieve greater accuracy; for instance, they may be able to model surface concavities whose shape cannot be captured by silhouettes from any viewpoint.

A few examples of particle shape reconstruction techniques that are not based on silhouettes follow.

Bouguet [13] demonstrates the use of point-based *stereo reconstruction* of a rock from a turntable sequence (see Figure 2.8). Using the texture of the rock's surface, points are tracked across multiple frames and 3D coordinates are inferred from the points' 2D image locations. The method is not amenable to high throughput

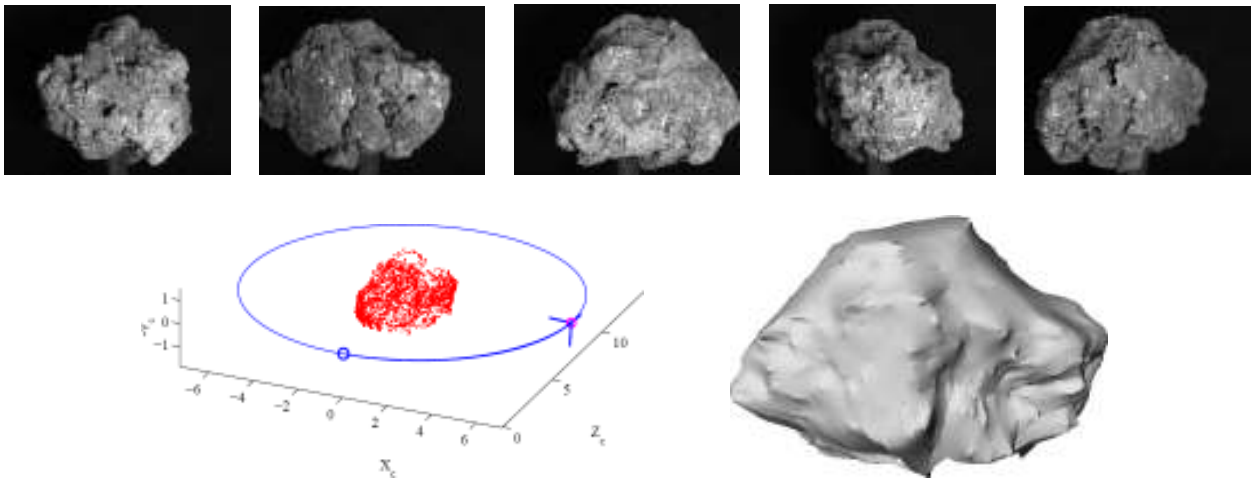


Figure 2.8: Rock shape reconstruction. Images are captured as a rock is rotated on a turntable. The top row of images shows 5 of the 226 images used. The camera trajectory and reconstructed 3D points *(bottom, left)* and triangular mesh model formed with Delaunay triangulation *(bottom, right)* are also shown (pictures from Bouguet [13]).

modelling since the rock must be rotated on a turntable, but the 3D models are potentially highly accurate. The method relies on the rock's texture for point tracking and is therefore unsuitable for textureless particles.

Erdogan et al. [38] describe the use of X-ray tomography for acquiring 3D particle shape. The particles must be embedded in a cement-like matrix, and are rotated in front of the X-ray scanner for several hours, so that multiple slices can be collated to form a 3D shape model. Multiple particles are imaged simultaneously, and individual particles are segmented from the 3D image. It is important that the matrix have significantly different X-ray absorption properties from the particle. The authors manually measured long, intermediate and short diameters for three rocks using digital calipers. The maximum discrepancy between the X-ray models and the manual measurements was 2.6 mm for a longest diameter of 74.5 mm (i.e., an error of approximately 3.5%).

Lanaro and Tolppanen describe an alternative X-ray imaging setup. A cone beam is used (as opposed to collating slices.) The authors cite greater accuracy and the ability to model the interior of opaque solids as the reasons for preferring the cone beam approach to slices. Samples of rock particles, glass beads and quartz sand are demonstrated to have different shape properties in terms of elongation and flatness measured from 3D reconstructions. It is suggested that their method can be used for creating approximate 3D models for detailed numerical modelling of particulate processes. Since realistic particle simulations typically require a large number of particles to be considered, it is suggested that simple ellipsoidal models that share volume, elongation, and flatness properties with the 3D reconstructions be used. This reduces the computational load.

Unlike X-ray tomography methods, laser scanning acquires surface data points one at a time. Lanaro and Tolppanen [71] describe a laser scanning setup in which the surface of individual stones is scanned with a laser and viewed by two cameras. Triangulation of the projection of the laser line yields the corresponding 3D surface coordinates. Since a scan only captures one side of each particle, it must be turned over and rescanned

to capture the hidden side. To register the two surfaces in a common reference frame, at least three point correspondences are required. These are determined by gluing three ball bearings to each particle. The centre points of the balls are identified and used as reference points. Seven railroad ballast particles (32–64 mm) were reconstructed. The computed volumes differed from the ground truth values (measured manually) with a greatest underestimate of 5.3% and a greatest overestimate of 3.2%.

2.8 Summary

This chapter has illustrated the broad range of interest in particle shape analysis from many different fields, and the range of solutions that have been devised to estimate particle shape. This provides the historical background to the work that is presented in this thesis.

The principal shape features of interest are particle volume, and the long, intermediate, and short diameters. Many approaches, both silhouette-based and others, have been carried out to estimate these shape features. Researchers have tried various different approaches to capturing silhouettes from multiple viewpoints (multi-camera setups, turntables, manual repositioning of particles, perpendicular faced trays).

All the multi-view silhouette-based setups (with the possible exception of the commercial turntable systems for which explanations of methodology are not available) rely on accurate positioning of the apparatus, rather than calibration. Calibration is limited to estimating scale so that pixel coordinates may be converted to world coordinates such as millimetres. This makes the implicit assumption that a weak perspective approximation is appropriate. This thesis proposes new methods to calibrate multi-view setups so that principled estimates of particle shape can be made using geometric reasoning, and so that individual particles can be efficiently and effectively recognised from their silhouettes.

Although silhouette-based particle analysis makes use of concepts that are covered in the computer vision literature (such as the visual hull), there appears to be little awareness amongst particle analysis researchers of the shape-from-silhouette research from the field of computer vision.

The problem of individual particle recognition has been mentioned a few times in the particle analysis literature, but no quantitative studies appear to have been carried out.

Particle shape reconstruction has been attempted with non-silhouette-based approaches. Examples include X-ray tomography, laser scanning, and stereo reconstruction. Although these systems have the potential to reconstruct shape more accurately than silhouette-based methods, they tend to be both expensive and slow. Attention has been paid to computational efficiency for the algorithms presented in this thesis. In practice, feeding a batch of particles through the six-camera setup (at a rate of ten particles per second) takes more time than computing the silhouette-based estimates of shape properties, or matching the silhouette sets across two runs.

Chapter 3

The Geometry of Silhouette Sets

3.1 Introduction

This chapter overviews the geometry of silhouette sets and introduces important concepts that will be used in this thesis. First the *visual hull*, the simple and widely-used method for approximating 3D shape from silhouettes, is covered. The explanation of the visual hull allows the related concepts of visual cones, cone strips, frontier points and viewing edges to be introduced.

Next, two concepts that are central to the methods developed in this thesis are covered: (1) the viewing edge midpoint hull (VEMH) as a means for efficiently estimating the 3D shape of the convex hull of a stone from its silhouettes, and (2) outer epipolar tangency error (ET error) as a measure of silhouette inconsistency.

The VEMH plays a central role in this thesis. However it is less important than ET error since an obvious alternative for efficiently approximating 3D shape from silhouettes exists: the visual hull. Efficiently approximating 3D shape from silhouettes will play a role in the following chapters:

1. In Chapter 6, moments computed from approximate 3D shapes will be used to form an initial pose estimate between silhouette set pairs of the same stone. The pose estimate is subsequently refined using ET error. The *a*, *b*, and *c* diameters which are widely used by particle shape analysts (as discussed in Chapter 2) will then be measured from approximated 3D shapes.

2. Chapter 8 describes a computationally efficient method for computing approximate dissimilarity between silhouette sets. The method is based on an estimate of the 3D shape of a stone computed from its silhouettes.

3. Chapter 10, the VEMH will be used to emulate sieve sizing.

In this chapter, the VEMH is introduced by first demonstrating how viewing edges impose bounds on the caliper diameter of the corresponding stone. This allows an upper and lower bound to be computed for the longest and shortest diameter (given a noise-free silhouette set). Next, the VEMH is presented as an alternative to the visual hull for estimating 3D stone shape from silhouettes.

In later chapters, the ET error will form the basis for developing the following methods:

1. In Chapter 4, the ET error is used to calibrate a mirror-based setup from silhouettes.
2. In Chapter 5, the ET error is used to calibrate a high throughput six-camera setup from silhouettes.
3. Chapter 6 demonstrates how the ET error can be used to infer the relative pose between two silhouette sets of the same object.
4. Chapter 7 shows how the pose estimation method of Chapter 6 can be used to distinguish two silhouette sets of the same object (a match) from two silhouette sets of two different objects (a mismatch).

The ET error is introduced by first briefly covering silhouette consistency in general. The ET error, which is based on the epipolar tangency constraint (a necessary, but insufficient condition for consistency) is then described.

The methods described in this chapter have been chosen for their simplicity, which leads to efficient computation. Efficient computation is important for online computations of the high-throughput six-camera setup that captures image sets at a rate of ten stones per second. Efficiency is also crucial for solving the batch matching problem for realistic sized stone batches (hundreds to thousands of stones per batch) without making use of unreasonably long running times (computing the matching should not take longer than it takes to feed the stones through the camera setup).

3.2 Visual Hulls

3.2.1 The Visual Hull Concept

The term *visual hull* was coined by Laurentini [73] in the 1990s, but the use of the largest silhouette-consistent object as a means for 3D modelling dates back to the work of Baumgart in the 1970s [7]. Laurentini's initial use of the term visual hull described the largest object consistent with all possible silhouettes, but the term is now usually used to refer to the largest object that is consistent with a finite set of available silhouettes [76, 91]. In this thesis, the visual hull is the largest object that is consistent with a given set of silhouette views. The *line hull* is the complement of space covered by all lines that do not pass through the object. Line hull is a term from the field of computational geometry that is equivalent to Laurentini's visual hull computed from all viewpoints outside an object's convex hull [90].

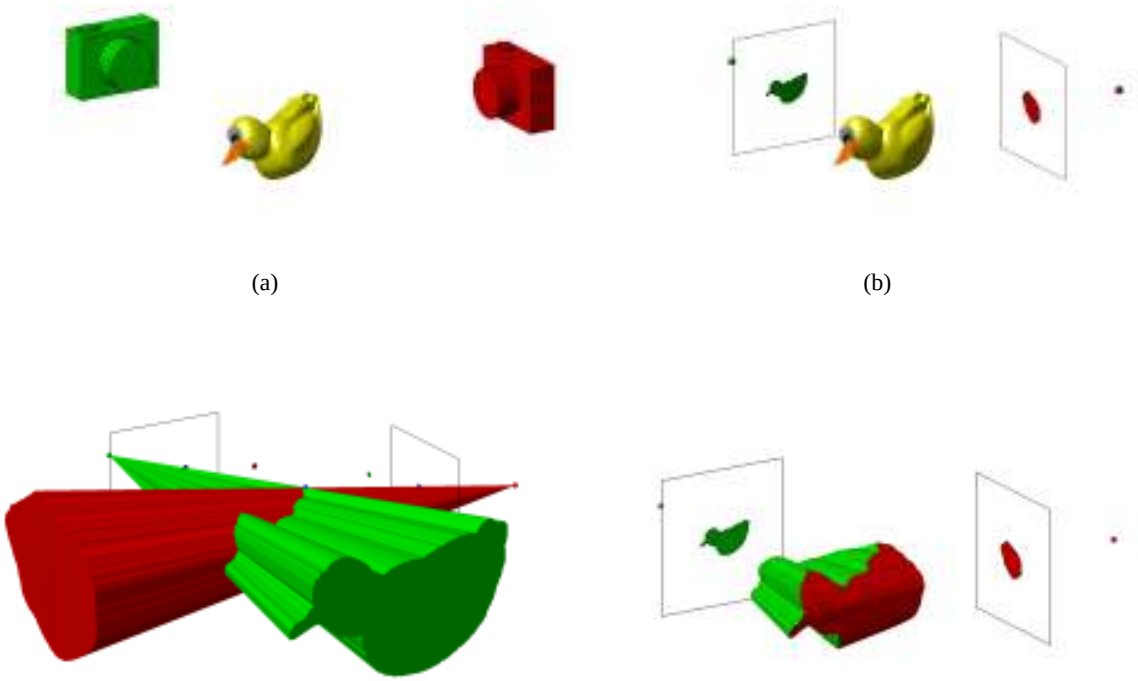
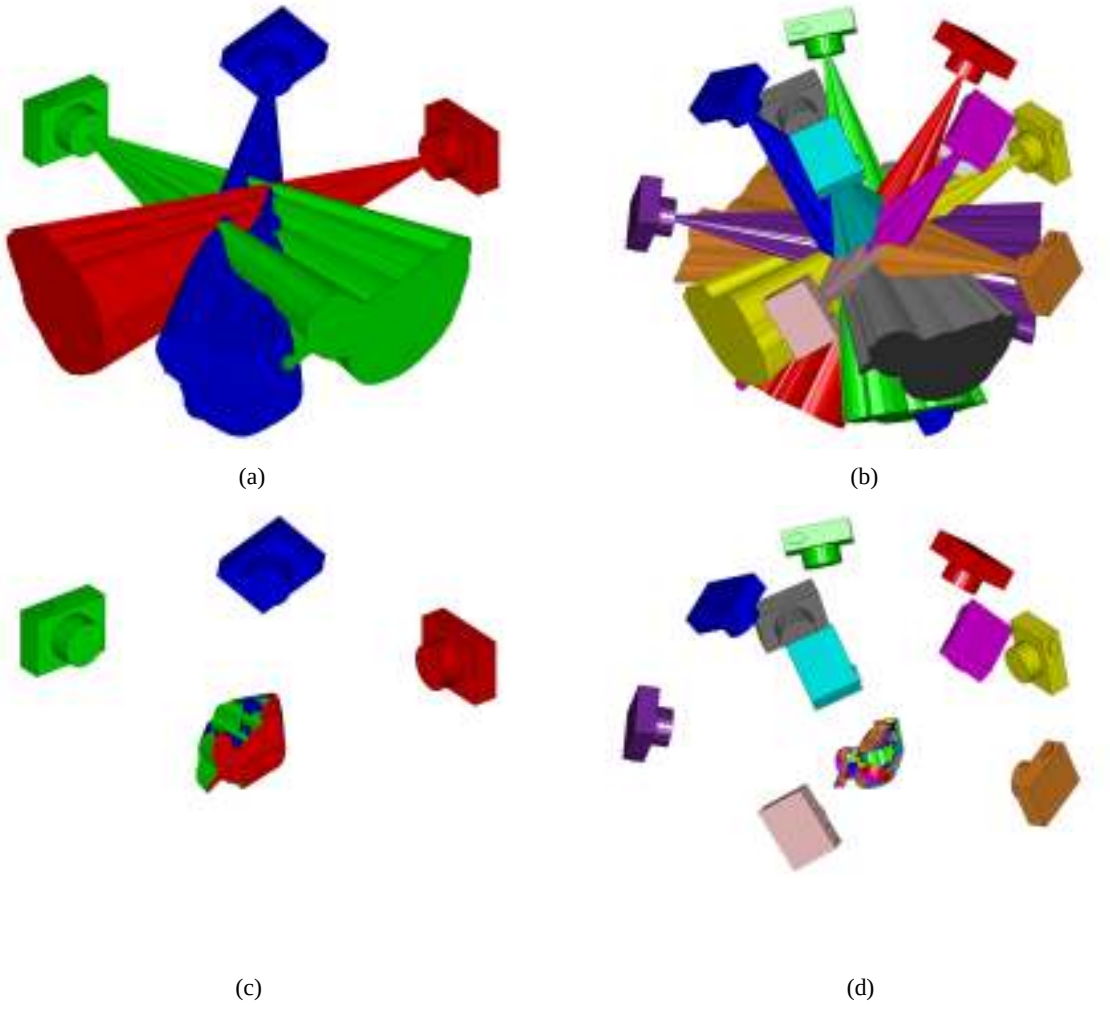


Figure 3.1: The visual hull concept: (a) a duck viewed by two cameras, (b) two silhouette views of the duck, (c) the two visual cones associated with the two silhouette views, (d) the visual hull formed from the two silhouette views.

The visual hull concept is illustrated in Figure 3.1. Figure 3.1b shows two silhouette views of a duck (the object being imaged). Camera centres are represented by small spheres. For convenience, the image planes are placed *in front* of the camera centres, and the projected silhouette views are shown non-inverted; for the purposes of this thesis, such a setup is geometrically equivalent to placing the image planes behind the camera centres. *Visual cones* corresponding to each silhouette are shown in Figure 3.1c. A visual cone is the volume of space that the object cannot lie outside of, given the observed silhouette. The intersection of the visual cones is the visual hull (shown in Figure 3.1d). The visual hull cannot be smaller than the object. With two silhouettes, the visual hull is often a poor approximation to the object. However, if further silhouette views are added, more information about which volumes of space are empty is added, and the visual hull becomes a better approximation to the object. Figure 3.2 shows visual hulls of the duck formed from three and from ten cameras. With the additional camera views, more visual cones carve away 3D regions that do not form part of the object, leaving a closer approximation to the object. Concave surface regions, however, cannot be reconstructed by the visual hull, since such regions (the interior of a coffee mug, for instance) do not affect silhouette shape. In a sense, it is the line hull of the object that is approximated by the visual hull. Fortunately, most particles are well-approximated by their line hulls. In addition, many properties of interest (such as caliper diameters) have the same value when measured from the object, its line hull, or its convex hull.



(a) (b)

(c) (d)

Figure 3.2: More accurate shape from additional views: (a) visual cones from three silhouette views, (b) visual cones from ten silhouette views, (c) the visual hull from the three silhouette views, (d) the visual hull from the ten silhouette views.

The surface of the visual hull is made up of surface regions from the visual cones. The part of the visual hull surface associated with each visual cone is a *cone strip*. Since this thesis only considers single objects of genus zero (i.e., objects without holes), each cone strip forms a single ring around the visual hull. At certain points, the rings are of zero width. These points are called *frontier points* and are important for the methods developed in this work. Frontier points are discussed in more detail in Section 3.5. Note that in practice, camera parameters and silhouettes will not be known exactly (i.e., there will be some degree of noise). This means that some cone strips will be discontinuous if computed directly from visual cone intersections.

Visual hull approximations have been popular as a relatively simple and robust technique for 3D modelling, since silhouettes can be easily extracted under controlled lighting conditions. For instance, if diffuse back-lights are used so that the background appears much lighter than the foreground, then the silhouette can be extracted using a simple threshold on the pixel intensity values.

3.2.2 Computing the Visual Hull

In order to determine the visual hull corresponding to a set of silhouettes, the cameras that produced the images must be calibrated. This means that the internal camera parameters (such as focal length, principal point) and the pose must be (at least approximately) known. In this thesis, the term *silhouette set* is used to refer to a calibrated set of silhouettes (i.e., the view poses are known in a common reference frame, and the camera internals are known). Furthermore, the silhouettes will be approximated by *polygons*. As pointed out by Lazebnik [75], the use of polygons rather than higher order spline curves allows simpler and more efficient methods to be developed.

Voxel-Based Approaches

A simple means of approximating the visual hull from a silhouette set is to consider the voxels that tessellate the common field of view. The size of the voxels will determine the resolution of the computed visual hull. Only voxels that project into the silhouette foreground in a *all* views are classified as part of the visual hull. Other voxels are classified as empty.

The efficiency of the voxel-based method can be improved by using an octree decomposition as described by Szeliski [123]. Initially, a coarse voxel grid is considered. Any voxel that projects entirely into the background in *any* view is classified as empty. Any voxel that projects entirely into the foreground in *all* views is classified as visual hull. The remaining voxels are each subdivided into eight smaller voxels that are then classified as empty, visual hull, or subdivide. Subdivision ceases once a sufficiently high resolution has been achieved.

Once a voxel representation has been computed, a technique such as marching cubes [84] can be used to create a polygonal surface. This approach considers all voxels through which the surface passes. (If octree subdivision is used, these are the smallest voxels.) A surface patch is created for each of these voxels. The shape of the patch is determined by which of the voxel vertices lie inside the visual hull.

Surface-Based Approaches

A second group of approaches proceed by considering the surface rather than the volume of the visual hull. This includes some of the original approaches [7] in which constructive solid geometric techniques were used to directly intersect the visual cones.

Since intersecting general polyhedra is slow, methods have been developed to take into account the specific geometry of the visual cones, so that they may be efficiently intersected. Matusik et al. [91] make use of an edge-bin data structure to store edges associated with angular ranges of lines through the epipole. The method achieves efficiency by computing intersections in 2D: polygonal intersections are first formed in the

image plane, and then these intersections are intersected with one another on planes defined by facets of the viewing cones.

Franco and Boyer [48] describe another efficient method for computing polygonal surface models of silhouettes. The first step is to compute the *viewing edges* from a silhouette set. A *viewing line* is a line that passes through a silhouette vertex and its camera centre. Intersecting a viewing line with the visual cones from all other cameras leaves a viewing edge. The vertices of the viewing edge endpoints are vertices of the visual hull polyhedron. Franco and Boyer show how the connectivity of these vertices and the remaining surface points can be determined by using local orientation and connectivity rules to walk along the viewing cone intersection boundaries.

Several alternative approaches for computing the visual hull are described in the computer vision literature [17, 76, 85, 125].

It is interesting to note that the convex hull of the visual hull can be computed relatively simply by forming an intersection of all halfspaces defined by the edges of the silhouette polygons. The plane specifying a halfspace is formed by the edge and its camera centre. Many efficient halfspace intersection algorithms exist. For instance, the Quickhull algorithm [5] is of $O(n \log n)$ time complexity for n halfspaces. If the planes are treated as points in dual space, then the duals of the facets of their 3D convex hull specify the visual hull vertices in primal space.

3.3 Constraints Imposed by Viewing Edges

This section demonstrates that silhouette sets impose both an upper and lower bound on the caliper diameter in a given direction. These bounds are derived by considering viewing edges.

By considering the upper and lower bounds over all directions, it is possible to compute upper and lower bounds for the longest and shortest diameter of a stone from its silhouette set. Although estimating the longest and shortest diameter of a particle from its silhouette set is of interest to particle shape analysts (as discussed in Chapter 2), it does not appear to have been pointed out that a silhouette set imposes bounds on these properties.

3.3.1 Bounds on Caliper Diameters in a Given Direction

Since the caliper diameter of a stone in a given direction is the same as that of its convex hull, for simplicity the convex hull of the stone will be considered. The convex hulls of the observed silhouettes are the projections of the convex hull of the stone.

Let the caliper diameter of a stone in direction $\mathbf{r}$ be $d_{\mathbf{r}}$. The value $d_{\mathbf{r}}$ is the distance between two parallel support planes that are tangent to and enclose the object (Figure 3.3a). The tangent plane normals are parallel to $\mathbf{r}$.

The 3D shape of the object is unknown; all that is available is a silhouette set. The upper bound $d_{U\mathbf{r}}$ for the caliper diameter is the largest $d_{\mathbf{r}}$ value that can be computed from an object that could have produced the observed silhouettes. The visual hull provides the upper bound for $d_{\mathbf{r}}$. No greater value is possible, since if either support plane were moved away from the object, no object would be able to be both tangent to the support planes and able to produce the observed silhouettes.

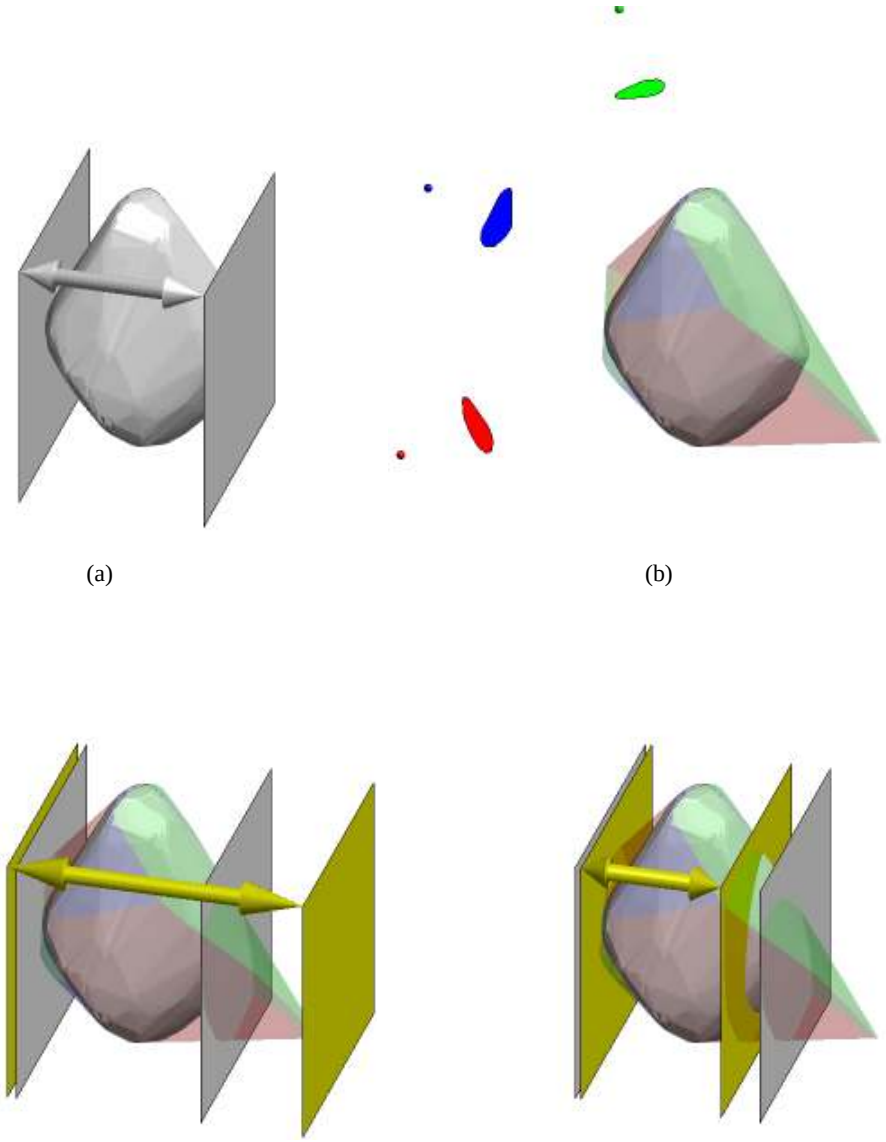
The method for computing a caliper interval for a given direction is illustrated in Figure 3.3. The figure shows the support planes of the actual caliper diameter for a given direction (Figure 3.3a) and three silhouette views that are used to construct a visual hull (Figure 3.3b). The visual hull is the largest object that is consistent with the silhouettes. It can be used to compute the largest caliper diameter along the given direction that is consistent with the silhouettes (Figure 3.3c).

Identifying the lower bound $d_{L\mathbf{r}}$ (Figure 3.3d) on $d_{\mathbf{r}}$ from the silhouette set is less obvious. The support planes of $d_{L\mathbf{r}}$ must be as close as possible without destroying any cone strips that generate the observed silhouettes.

Finding the smallest consistent caliper diameter along a given direction is illustrated in more detail in Figure 3.4. Figure 3.4a shows the visual hull model that is made up of cone strips corresponding to the three silhouettes. In the noise-free case, the cone strips project exactly onto the corresponding silhouette outlines. Each cone strip represents the only regions in 3D space that may generate the corresponding silhouette outline and remain consistent with all silhouettes. A *viewing edge* is the portion of a ray through the silhouette outline that coincides with the corresponding cone strip. Some part of each viewing edge must be tangent to the object, so that the point on the silhouette outline is generated. No viewing edge can therefore lie outside the support planes that contain the object. This provides a means for calculating the smallest consistent caliper diameter: the support planes must be as close together as possible, without any viewing edge lying entirely outside the region between the support planes. In Figure 3.4, the upper support plane is limited by the viewing edges that form the green cone strip (Figure 3.4c): if the support plane were moved any closer, viewing edges from the green cone strip would lie entirely outside the region between the support planes. Note that the portion of the visual hull that lies between the support planes generates the observed silhouettes, and is therefore an example of a silhouette-consistent object with a diameter $d_{L\mathbf{r}}$ in direction $\mathbf{r}$.

3.3.2 Bounds on the Longest and Shortest Diameters

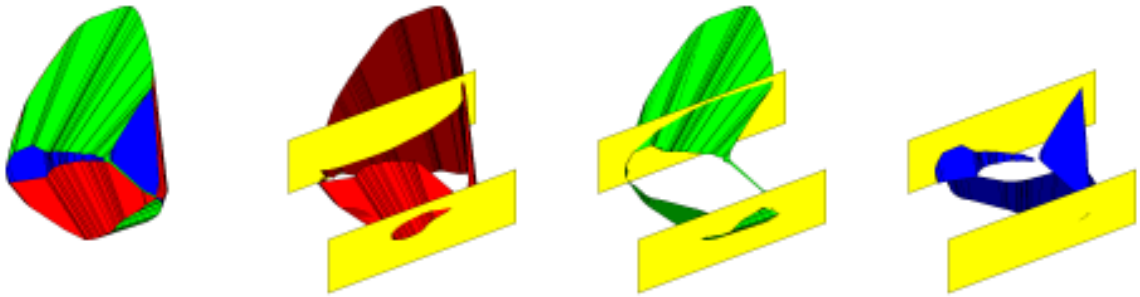
Since silhouette sets impose bounds on the diameter in a given direction, it is interesting to note that a silhouette set imposes bounds on the longest and shortest diameters (quantities of interest to particle shape



(a) (b)

(c) (d)

Figure 3.3: Caliper intervals: (a) the caliper diameter of a stone for a given direction, (b) the available information: three silhouettes from which a visual hull consisting of cone strips from each silhouette can be constructed, (c) the maximum caliper diameter along the given direction that is consistent with the silhouette set, (d) the minimum caliper diameter along the given direction that is consistent with the silhouette set.



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

Figure 3.4: Diagram showing (a) the visual hull, and (b–d) the three constituent cone strip components along with the support planes for the minimum consistent caliper diameter. The example uses the same stone, silhouettes, and caliper direction as Figure 3.3.

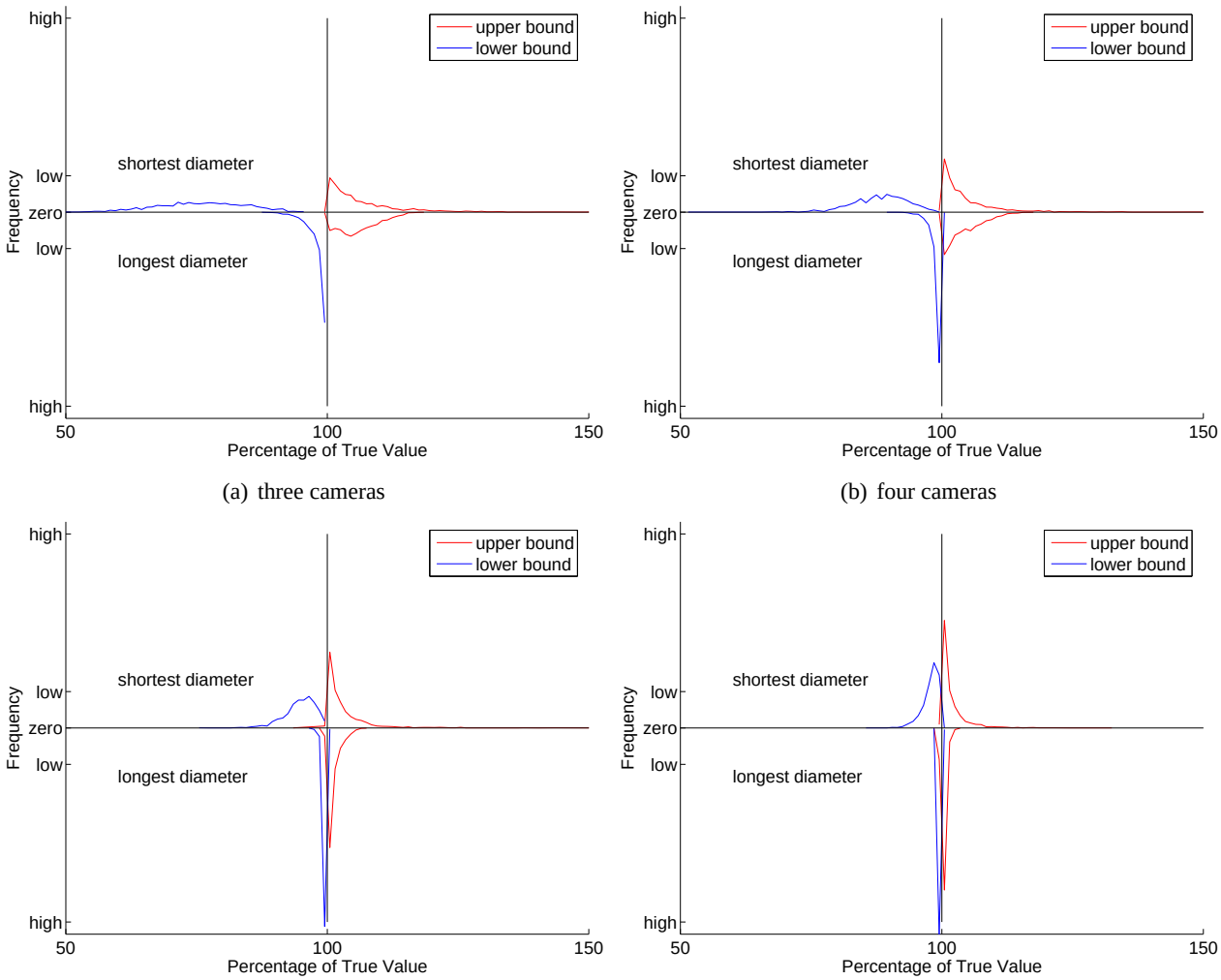
analysts). The bounds are, however, geometrical bounds and are only valid for noise-free silhouettes. Nevertheless, the bounds provide insight into the inherent (i.e., geometrical) limits of the extent to which longest and shortest diameters can be computed from silhouette sets.

The upper bound for both the longest diameter and the shortest diameter are simply computed by finding the longest diameter and the shortest diameter of the visual hull, because no larger shape is consistent with the silhouettes. Computational geometers have discovered exact solutions for determining the longest diameter (termed simply the *diameter*) and the shortest diameter (the *width*) of arbitrary polyhedra [24, 57]. These methods can be applied directly to a polyhedral representation of the visual hull to obtain upper bound values.

Lower bounds for the longest and shortest diameter are approximated by considering an approximately uniform dense sampling of directions obtained using subdivisions of an icosahedron [61]. The best solution from dense sampling is then refined using a conjugate gradient optimiser which makes use of an analytical expression for the partial derivatives of the lower bound diameters with respect to an azimuth-elevation angle parameterisation of direction.

An experiment was carried out using synthetic data in which the longest and shortest diameter of a polyhedral stone model is compared with the bounds computed from its silhouette set. Synthetic silhouettes were generated using 3-, 4-, 6- and 10-camera setups. To provide viewpoints that are well distributed about the viewing hemisphere, setups with n cameras are positioned to look onto the parallel face pairs of a $2n$-faced Platonic solid (such setups are described in more detail in Chapter 5). The refined visual hull models of a data set of uncut gemstones, illustrated in Appendix C (pages 222–224), were used as polyhedral stone models. The stones were randomly oriented. For each polyhedral stone model, the longest and shortest diameter was computed. The upper and lower bounds were then computed from silhouette sets of the stone. These bounds are expressed as a percentage of the actual value. Ideally, lower bounds should be less than 100% of the true value, and upper bounds should be greater than 100% of the true value, but since the bounds are only approximated, there are a small number of cases in which this is not true.

Figure 3.5 shows plots of the distributions of bounds for the four different camera setups considered. To aid comparison, the upper half of each sub-plot shows distributions for the bounds on the smallest diameter, and the lower half of each sub-plot shows distributions for the bounds on the longest diameter. As the number of cameras is increased, the bounds move closer to 100%. This is because the additional views place tighter constraints on the range of possible values. Notice that the bounds on the longest diameter are closer to 100% than those on the shortest diameter, indicating that there is less uncertainty on its value. Interestingly, the plots indicate that the longest diameter is better approximated by its lower bound, whereas the shortest diameter is better approximated by its upper bound.



(a) three cameras

(b) four cameras

(c) six cameras

(d) ten cameras

Figure 3.5: Distributions of bounding values computed from silhouette sets as specified as a percentage of the actual values. Silhouette sets were computed from the 1423 polyhedral stone models illustrated on pages 222–224.

3.4 Viewing Edge Midpoint Hulls for Approximating Shape

The viewing edge midpoint hull (VEMH) is proposed as an alternative to the visual hull for approximating the 3D convex hull of a stone from the 2D convex hulls of its silhouettes. The VEMH is the convex hull of the midpoints of all viewing edges. The silhouette projections of the VEMH are the same as the convex hulls of the observed silhouettes in the noise-free case, so the VEMH is a silhouette-consistent object.

3.4.1 Advantages of the VEMH

Use of the Convex Hull

The approach taken in this thesis is to attempt to reconstruct the 3D shape of the convex hull of a stone from its silhouettes rather than the possibly nonconvex shape of the stone. Using convex hulls simplifies computations and allows for 3D shape to be approximated more efficiently than if nonconvex shapes are considered. This approach is useful in two contexts:

1. Since the caliper diameter of a stone in a given direction is the same as that of its convex hull, the VEMH can be used to estimate caliper diameters. This will be done for both estimating the short, intermediate, and long diameters of a stone and for estimating a caliper diameter distribution to aid recognition.
2. Since the principal axes of the convex hull of a stone can be used to specify its pose with respect to some reference frame, the VEMH is used to approximate the pose of a stone from its silhouette set. This provides an initial pose estimate that will be used to align silhouette sets of the same stone in a common reference frame.

Comparison with the Visual Hull

The aim of the VEMH is to provide a more accurate estimate of the 3D shape of stones from silhouettes than the visual hull.

Visual hulls often have sharp edges where cone strips meet. Although geometrically the visual hull could be the object that generated the silhouettes, more often the sharp edges are artefacts that do not exist on the actual object. Unless by chance a stone's surface is tangent to the cone strip near the regions where cone strips meet, the volume of the visual hull near the cone strip intersections and far from the frontier points is not shared by the stone.

In general, an object will be tangent to the viewing edge at one point along the viewing edge. Using the visual hull to approximate stone shape considers the stone to be tangent to the entire viewing edge (this is an extremely unlikely coincidental alignment of the stone). Since stones are not in general smooth, no use of the silhouette curvature is used for interpolation, and the midpoint of the viewing edge is simply used as the point of tangency for the shape approximation. The convex hull of the midpoints is used as the shape approximation. Although additional volume could be incorporated into the shape approximation, this is not done for two reasons:

1. Since the VEMH is silhouette-consistent, the silhouettes do not provide any evidence of the presence or absence of additional volume. One would have to make use of *a priori* knowledge of shape. Since stones are irregular in shape there is no obvious *a priori* knowledge to incorporate.

2. To a certain extent, many stone surfaces consist of low-curvature regions (flattish faces) that are joined by high curvature regions (edges). Since the stones are arbitrarily oriented with respect to the cameras, high curvature regions are most likely to form contour generators, with the flatter regions in between. This parallels with the VEMH in which rims are joined by flat faces, and is unlike the visual hull in which the volume extends to the limit of silhouette consistency.

Figure 3.6 illustrates the differences between visual hulls and VEMHs. Note that the much of the visual hull volume in the regions where cone strips meet, and which is absent in the VEMHs, is also absent in the original stones.

3.4.2 Alternative 3D Shape Estimates from Silhouette Sets

Several other approaches are described in the literature for estimating the 3D shape of an object from its silhouettes. The main advantage of the VEMH over these methods is it computational efficiency (how this is achieved is described in Section 3.4.3) and its simplicity.

Visual Shapes

Franco et al. [49] introduce a family of silhouette-consistent 3D objects that they term *visual shapes*. Their approach is similar to the VEMH in that a portion of the viewing edges is included to ensure silhouette consistency. Three approaches for selecting portions are suggested: (1) thinning the viewing edges, (2) selecting a single random contact point, and (3) choosing the contact point corresponding to a local order 2 surface. Of the three approaches, the VEMH is most similar to the second. However, the VEMH approach makes use of the midpoint instead of a random contact point. Compared with the random approach, the midpoint approach reduces by a factor of two the maximum possible distance between the actual contact point and the assumed contact point. (Despite the similarity between the VEMH and visual shapes, the VEMH was developed independently and prior to the publication of the visual shapes.) To determine a polyhedron from the visual shape points, Franco et al. compute the Delaunay tetrahedrisation of the points, and then carve tetrahedra that project outside any silhouette.

Dual-Space Approaches

Another approach to approximating 3D shape from silhouettes is to represent tangent planes to the object (that are defined by the silhouette outlines) as points in dual space, and then to estimate the dual surface of

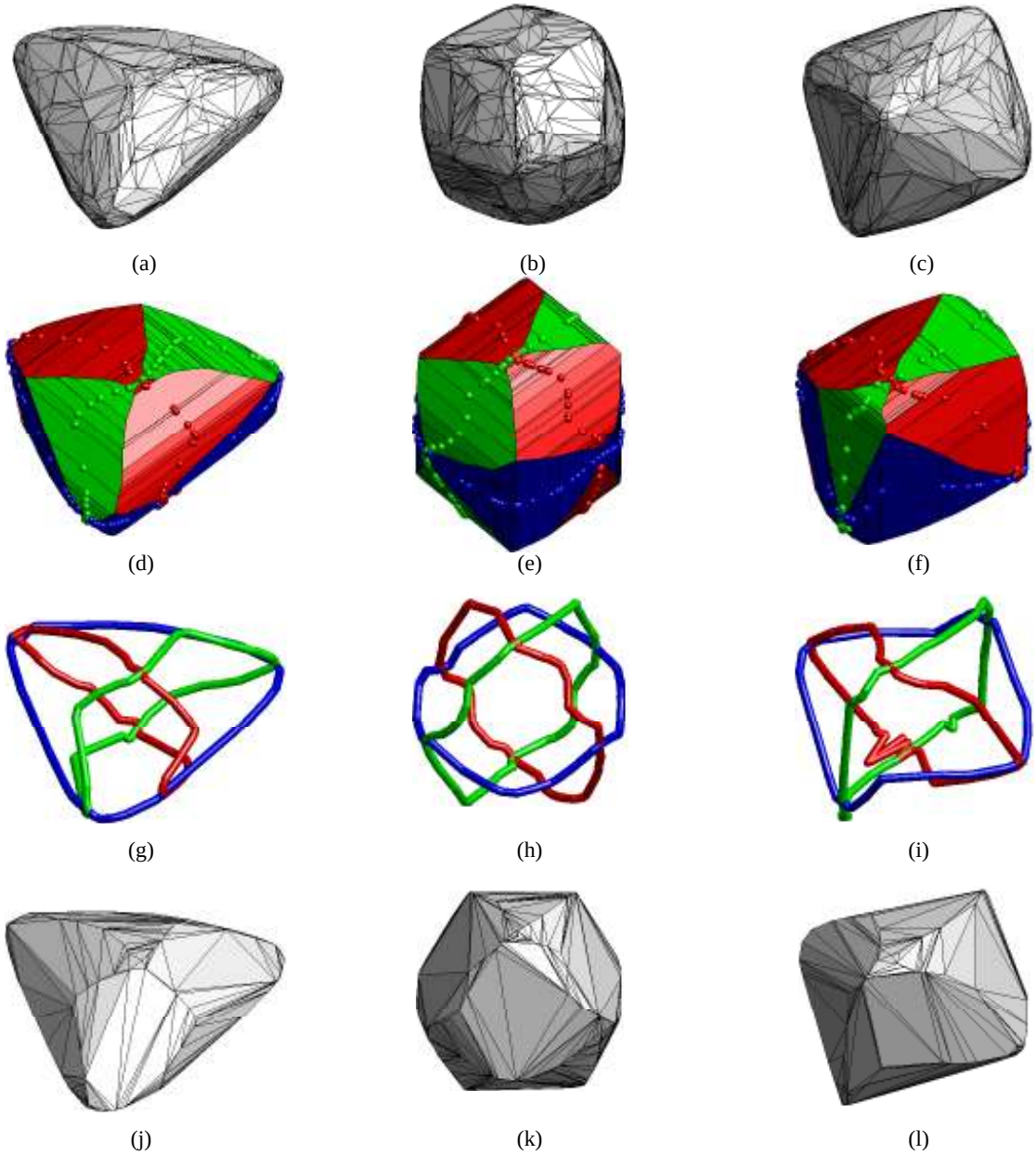


Figure 3.6: Visual hulls and VEMHs generated from three orthogonal silhouette views of stones. The first row (a–c) shows three stones. The second row (d–f) shows the 3-view visual hulls computed from 3-view silhouette sets of the above stones. The visual hull surfaces are coloured according to the cone strips that they are made up from. Viewing edge midpoints are shown as small spheres. The third row (g–i) shows the rims of the VEMHs. These are the loci of the viewing edge midpoints; it is the rims that generate the silhouette outlines (i.e., the silhouette outlines are projections of the rims). The fourth row (j–l) shows the VEMHs: the convex hulls of the viewing edge midpoints.

the object [17, 68, 80]. However, as pointed out by Franco et al. [49], these approaches do not enforce the constraint that other silhouettes limit the position of tangency on the viewing line (i.e., the tangency must occur on a viewing edge, rather than anywhere along a viewing line); these approaches are therefore unsuitable for sparse silhouette sets in which the viewpoints are well-distributed (as is the case for the silhouette sets considered in this thesis.) In addition, the dual-space approaches assume that surfaces can be locally modelled with a quadric; this approach is unlikely to work well with stones, since they are not in general smooth.

Nevertheless, it is noted that a dual-space approach may yield a good solution to the problem of estimating the *convex hull* of a stone from its silhouette set. The tangent envelope corresponding to the convex polygonal representation of each silhouette boundary is a planar convex polygon in dual space [110]. (The tangent planes at the crossing points of these planar convex polygons correspond to frontier points in primal space.) The convex hull of these polygons corresponds to the visual hull in primal space. (This arises from the duality between halfspace intersections in primal space and convex hull in dual space.) This approach provides two useful properties:

1. Points may be added in dual space (to the original points that are the vertices of the planar convex polygons). The convex hull of all points corresponds to a polyhedron in primal space that is a carved version of the original visual hull. Ensuring that all points lie on the surface of the convex hull in dual space ensures that the corresponding primal space polyhedron is silhouette-consistent (i.e., it generates the observed silhouettes). Convexity preserving interpolation of the planar convex polygon vertices may therefore provide a smooth silhouette-consistent shape.

2. Since the convex polygons corresponding to each silhouette are planar, methods for interpolating cross sections [9] may provide a means for computing a smooth silhouette-consistent shape.

Radial Basis Functions

As with the VEMH, Collings et al. [29, 30] impose the restriction of approximating the 3D shape of convex objects from convex silhouettes. They approximate a convex solid from its silhouettes by fitting implicit radial basis functions. This is achieved by computing the positions of frontier points, which are assumed to lie on the surface, and by incorporating local curvature at frontier points. The method relies on the solid being sufficiently smooth that local curvature can be used to interpolate the surface regions between frontier points, and is therefore not applicable to stones, for which this assumption is not generally valid. Unlike the VEMH approximation, the method does not enforce the constraint that the object is tangent to the viewing edge interval on each viewing line. The reconstructed shape is therefore not constrained to be silhouette-consistent as it is not constrained to lie within the visual hull.

Triangular Spline Models

Sullivan and Ponce [121] describe a method in which triangular spline models are used to approximate the 3D shape of an object from its silhouettes. The spline model is deformed using an iterative minimisation of the average distance between the surface and viewing lines defined by the observed silhouette set.

The Constant Depth Rim Hull

Possibly the simplest and most efficient estimate of 3D shape from multiple silhouettes is the constant depth rim hull (CDRH). Marr [88] speculates that the human visual system may infer 3D shape from silhouettes by assuming that the rim (contour generator) is planar (i.e., constant depth). (This is however disputed in a later article by Koenderink [69]). Regardless of whether or not the human visual system may infer shape by assuming planar rims, the assumption of planar rims provides a simple and computationally efficient means for approximating the 3D shape of stones from multiple silhouette views. First the object depth is approximated by triangulating the centres of each silhouette to provide an approximate centre point. The polygonal silhouette boundaries are then backprojected to the depth of the centre point to form planar rims at that depth. The convex hull of the planar rims is the CDRH. Note that although the CDRH is used to approximate 3D shape, it is not necessarily silhouette-consistent: although the planar rims ensure that the CDRH projections are sufficiently large to cover the silhouettes, the projections may be larger than the silhouettes.

Assuming constant depth rims for approximating 3D shape does not appear to have been used for stones, but has been used for other objects such as fruit [67].

The CDRH is introduced in this thesis mainly to justify the additional complexity used in computing the VEMH. The CDRH and the VEMH are similar in that both compute a rim for each silhouette in the silhouette set. (The rim projection is the corresponding silhouette outline.) The CDRH and the VEMH differ in that CDRH rims are of constant depth, whereas VEMH rims may vary in depth. To justify the additional complexity of the VEMH it will be demonstrated that it provides more accurate estimates of 3D shape (for the tasks relevant to this thesis) than the CDRH.

Figure 3.7 shows an example that shows the VEMH and CDRH computed from three orthogonal silhouette sets of convex stones.

3.4.3 Computing the VEMH

The VEMH is computed by considering, in turn, the viewing line passing through each vertex of the polygons representing each of the silhouettes in the set. Each remaining silhouette in the set (i.e., the silhouettes other than that of the viewing line under consideration) is used to identify segments of the line that the object may

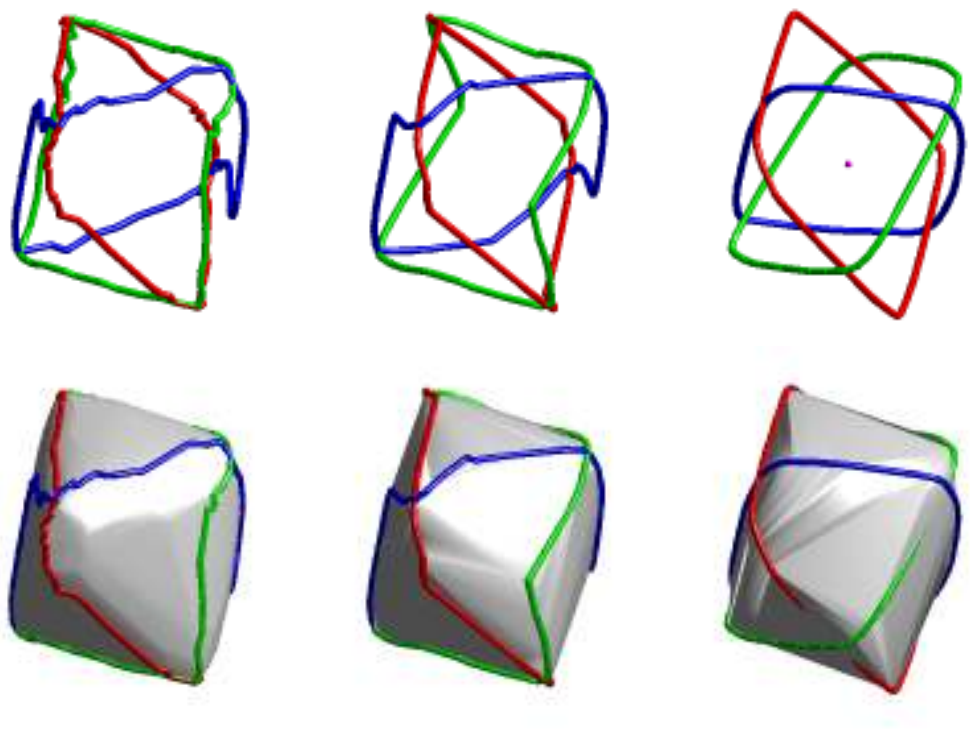


Figure 3.7: Estimating convex shape with rims using the VEMH and the CDRH. The first column shows the rims generated by a 3-view silhouette formed from three orthogonal cameras. Rims are shown with *(above)* and without *(below)* the imaged stone. The second column shows rims calculated using the viewing edge midpoints that are computed using the 3-view silhouette set. The convex hull of the midpoints (the VEMH) is also shown. The third column shows rims that are calculated by backprojecting the silhouette boundaries to a constant depth that is determined by triangulating the three silhouette centroids. The triangulated point that defines the constant depth is shown in purple. The convex hull of the constant depth rims (the CDRH) is also shown. Note that some portions of the constant depth rims lie within the CDRH indicating that it is not silhouette-consistent.

lie within. The intersection of all of the segments is the viewing edge. The convex hull of all viewing edge midpoints is the VEMH. For computing caliper diameters of the VEMH, it is not necessary to explicitly compute the convex hull of the midpoints, as the caliper diameter of the 3D point set (consisting of all viewing edge midpoints) can be used instead of a polyhedral representation of the VEMH.

Viewing Line Projections

To identify the segment of the viewing line that a silhouette does not imply as being empty, the viewing line is projected onto the silhouette. This is illustrated in Figure 3.8 in which the viewing line corresponding to vertex m is under consideration.

The viewing line passes through the points C_1 (the camera centre of Camera 1), P_1, M, and P_2 in the figure. Its projection is easily computed by projecting C_1 and m on to Camera 2's image plane. The viewing line projection is illustrated by the line passing through e_{21} (an epipole: Camera 1's projection onto Camera 2's

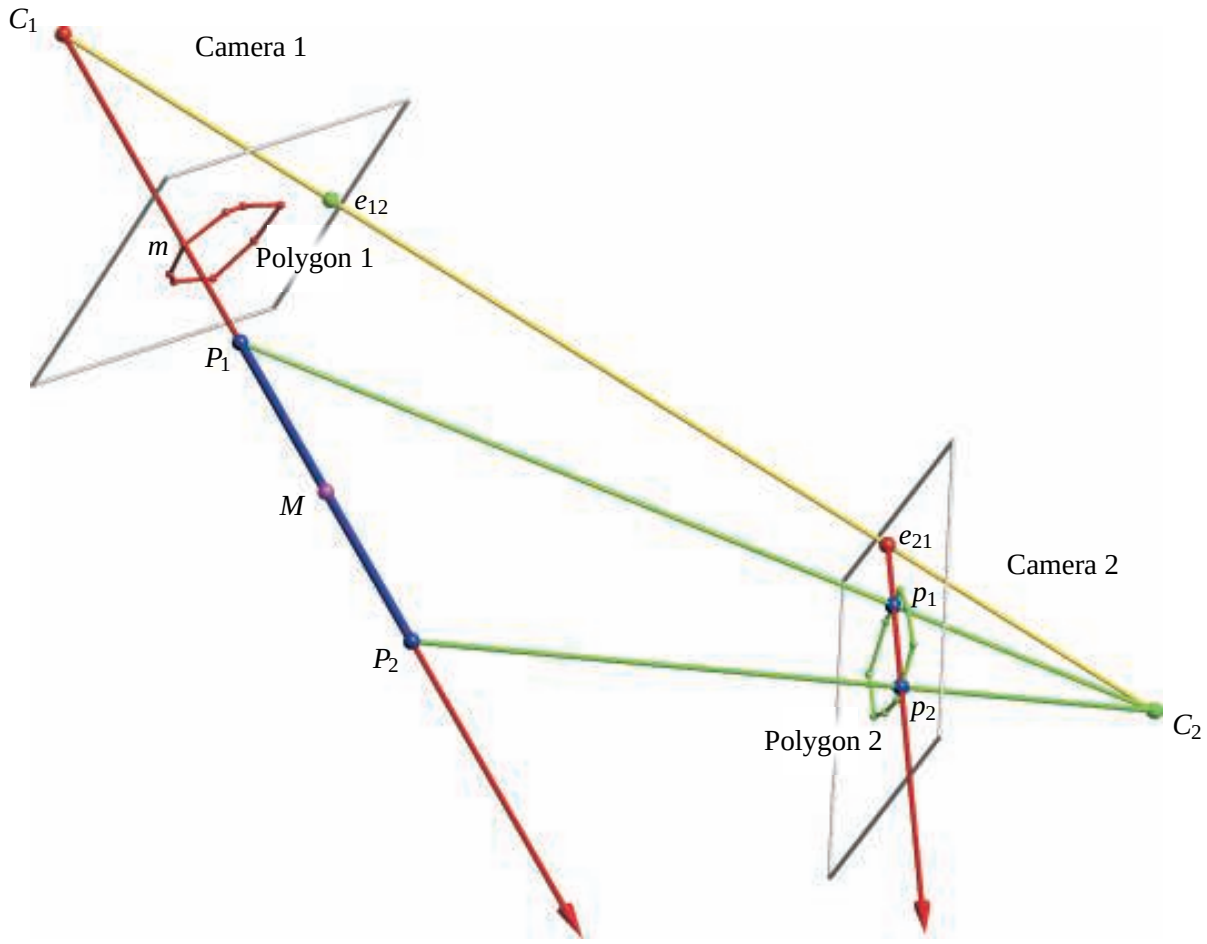


Figure 3.8: Computing a viewing edge from two convex silhouette views. Lines are represented as thin cylinders and points are represented as small spheres to aid 3D visualisation. The point *M* is the viewing edge midpoint.

image plane), p_1 and p_2. The points p_1 and p_2 are the points of intersection of Polygon 2 with the viewing line projection. Since the polygons are convex, there can be at most two such intersection points. Because of noise, there may be no intersection points in some cases.

Precomputed Edge-Bin Data Structures

An edge-bin data structure is precomputed for each silhouette-epipole pair. The edge-bin data structure allows the intersections of a viewing line projection through the epipole and the silhouette to be rapidly computed.

The edge-bin data structure is computed in a similar means to the one described by Matusik et al. [91]. However, since only convex silhouettes are used, it can be populated using a simpler algorithm.

First, bin boundaries are determined by sorting the polygon vertices according to the angle made with the epipole and an arbitrary fixed reference line. The implementation uses a line through the first vertex and the epipole as a reference line. An example is given in Figure 3.9: the six vertices of the polygon give rise to five bins.

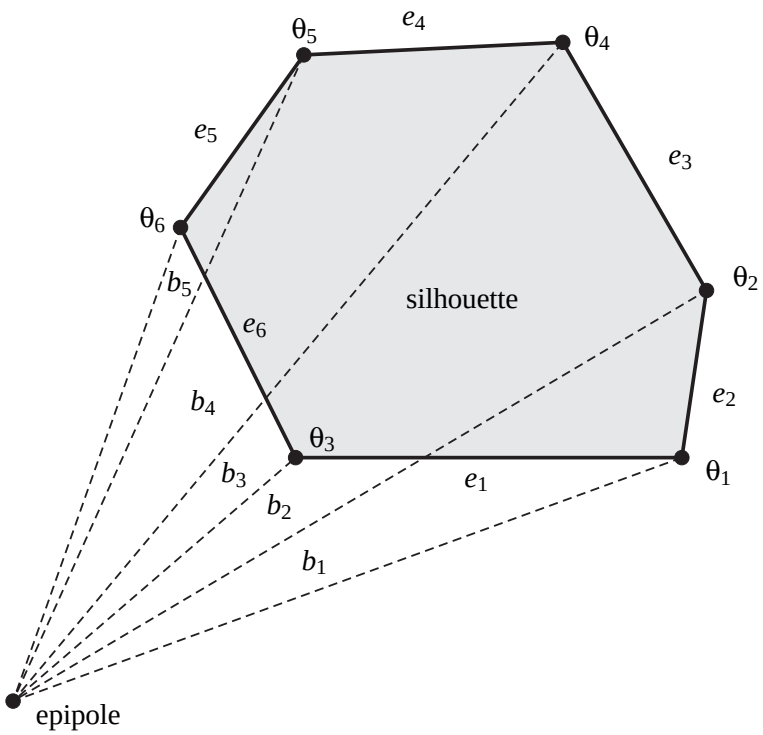


Figure 3.9: An example of a silhouette and epipole with edges and bins shown. The bin contents are listed in Table 3.1.

Each bin must store the edges that a line through the epipole whose angle falls into the bin's angular range will intersect. Since the polygon is convex, each bin will contain exactly two edges. This makes populating the bins easy.

The bins are traversed in order, and the polygon is traversed simultaneously, starting from the vertex with the smallest angle (θ_1 in the example). The current edge is added to the current bin if it falls within the bin's range, otherwise the current edge is updated by moving to the next edge (i.e., moving to the edge that shares a vertex with the current edge). Once the vertex with the largest angle is reached, each bin will contain one edge. The process is then reversed (the bins are traversed in reverse order) to add the second edge to each bin.

Table 3.1 shows the edges contained by each bin in the example.

Bin	Range	Edges
b_1	$\theta_1 - \theta_2$	e_1, e_2
b_2	$\theta_2 - \theta_3$	e_1, e_3
b_3	$\theta_3 - \theta_4$	e_6, e_3
b_4	$\theta_4 - \theta_5$	e_6, e_4
b_5	$\theta_5 - \theta_6$	e_6, e_5

Table 3.1: An example edge-bin data structure formed from the silhouette and epipole shown in Figure 3.9.

Intersections between the Viewing Line Projection and the Silhouette

Once the edge-bin data structure has been built, intersections can be computed efficiently. The angle of the viewing line projection is computed with respect to the reference line. This is used to determine the bin that contains the edges that intersect the viewing line projection. If the angle lies outside the range of all of the bins, then there is no intersection. Note that since the viewing line projections correspond to a polygon that is being traversed in another view, the appropriate bin is usually close to the most recently visited bin. This means that for B bins, lookup is of constant time complexity, rather than an $O(\log B)$ search, when the viewing line projections are processed in order.

It is possible that the following approach (not implemented) may further improve the simplicity and efficiency of the algorithm. Instead of forming edge-bin data structures, the intersection edges are determined by starting with the most recently intersected edge. Since the projected viewing lines are computed in order, the relevant edge will be found close to the most recently intersected edge, and the entire polygon need not be traversed. In other words, the silhouette polygon and the polygon that generates the viewing lines are traversed simultaneously.

Projecting Segments onto the Viewing Line

Once the intersection points p_1 and p_2 are known, they must be projected onto the viewing line to P_1 and P_2. To easily compute the intersection of line segments specified by different silhouettes, the points are specified as $P_1 = C_1 + d_1\hat{\mathbf{V}}$, where $\hat{\mathbf{V}} = (C_1 - m)/\|(C_1 - m)\|$ so that d_1 is the distance along the viewing line from C_1 to P_1.

The viewing edge is then computed as the intersection of all intervals as indicated by all silhouettes other than the silhouette corresponding to the viewing line. Because of noise, some interval intersections may be empty; in these cases the viewing line does not contribute a midpoint to the VEMH. (Figure 3.8 shows point M as the midpoint of the viewing edge specified by the two silhouettes for the viewing line under consideration.)

3.5 Measuring Silhouette Consistency

A consistent silhouette set is one that could have been produced as the silhouette projections of a 3D object. Geometrically, a silhouette set of an object is consistent if the intersection of the visual cones corresponding to each silhouette projects exactly onto the silhouettes. This is the cone intersection projection (CIP) constraint. It is a *sufficient* condition for consistency, since the cone intersection is an example of a 3D object that produces the silhouette set. It is also a *necessary* condition, since any portion of a silhouette that is not covered by the cone intersection projection provides contradictory information: the uncovered portion of the silhouette indicates that the corresponding viewing rays are occluded by an object, whereas the remaining silhouettes in the set indicate that the 3D region corresponding to these viewing rays is entirely empty.

Real silhouette sets are noisy: there will always be error associated with the camera parameters and the segmented silhouette boundaries. Real silhouette sets will therefore not, in general, be perfectly consistent. It is therefore useful to formulate a measure of the *degree* of inconsistency of a silhouette set.

The concept of a degree of inconsistency for a silhouette set is an important concept for this thesis:

1. By adjusting camera parameters to minimise the degree of inconsistency, cameras can be *self-calibrated*.
2. The degree of inconsistency can be used as a *diagnostic* to ensure that cameras have not been moved or adjusted since calibration. (Although this thesis does not analyse this diagnostic, it formed a useful tool during the data acquisition phase of the thesis project.)
3. It will also serve as a means for inferring whether two silhouette sets were produced by the same stone (a match): if a relative pose can be found to align the two silhouette sets so that the degree of inconsistency is sufficiently low, then the two silhouette sets are classified as a match.

This thesis makes use of a degree of inconsistency based on outer epipolar tangents and the epipolar tangency constraint. The use of epipolar tangencies for silhouette-based pose optimisation was first considered by Grattarola [54]. The method provides a computationally efficient means of obtaining pairs of point correspondences whose reprojection error provides a measure of inconsistency [138].

Other measures of silhouette consistency such as the *silhouette coherence* of Hernández et al. [39, 60] and the *silhouette calibration ratio* of Boyer [14, 15] use more information contained in the silhouettes, but are computationally inefficient. These measures are therefore not of primary importance for the methods developed in this thesis. However, they will be considered in the context of matching in Chapter 7.

3.5.1 The Epipolar Tangency Constraint

The epipolar tangency constraint is a geometrical constraint that applies to pairs of silhouette views (with associated pose and internal parameters): a line that is tangent to one silhouette and passes through the epipole

must project onto a silhouette tangent in the opposite image. With reference to an example (Figure 3.10), this section describes how the epipolar tangency constraint can be expressed in terms of the silhouettes and the camera pose and internal parameters.

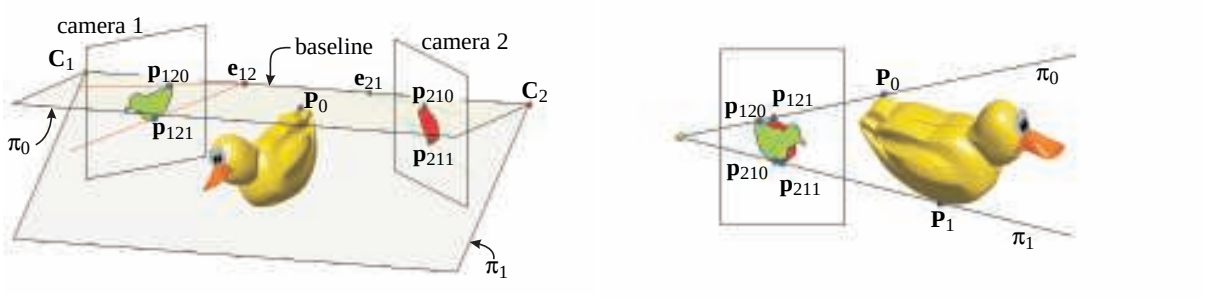


Figure 3.10: Two views of the epipolar geometry of a scene: (a) a front view, and (b) a side view looking onto the scene in a direction parallel to the baseline.

Figure 3.10 shows the same scene as shown in Figure 3.1, along with some additional points and planes. The line joining the two camera centres $\mathbf{C}_1$ and $\mathbf{C}_2$ is called the *baseline*. It pierces the image plane of Camera 1 at $\mathbf{e}_{12}$ and the image plane of Camera 2 at $\mathbf{e}_{21}$. The points $\mathbf{e}_{12}$ and $\mathbf{e}_{21}$ are epipoles. In the figure, the epipoles are represented as small circles (projections of spheres) on the image planes.

The two planes π_0 and π_1 that pass through the baseline and are tangent to the duck are shown. Provided that the baseline does not pass through the object, there will be two such planes for any object. The points $\mathbf{P}_0$ and $\mathbf{P}_1$, where the planes touch the object's surface, are frontier points. Since the planes pass through both camera centres and graze the surface of the object, the frontier points project onto the silhouette boundary in both views. The projection of a frontier point is the tangency point of a silhouette tangent that passes through the epipole. A projection of a frontier point is therefore termed an *epipolar tangency*. The epipolar tangencies $\mathbf{p}_{120}$ and $\mathbf{p}_{210}$ are projections of $\mathbf{P}_0$, and $\mathbf{p}_{121}$ and $\mathbf{p}_{211}$ are projections of $\mathbf{P}_1$. (The notation $\mathbf{p}_{ijk}$ is used so that i indicates the number of the camera whose image plane the point lies on, j indicates the number of the other camera of the silhouette pair, and k indicates to which of the two frontier points $\mathbf{p}_{ijk}$ corresponds.)

The intrinsic geometry between the views i and j is encapsulated by the 3×3 *fundamental matrix* F_{ji} [58]. If $\mathbf{x}_i$ represents the homogeneous coordinates of an image point from view i, and $\mathbf{x}_j$ represents the corresponding point in view j, then $\mathbf{x}_i$ is constrained to lie on the line $\mathsf{F}_{ji}\mathbf{x}_j$ in view i so that

$$\mathbf{x}_i^{\mathsf{T}} \mathsf{F}_{ji} \mathbf{x}_j = 0. \tag{3.1}$$

If the relative pose between view i and view j is described by a rotation represented by the matrix R followed by a translation represented by the vector $\mathbf{t}$ that transform points from the reference frame of camera j to the

reference frame of camera i, then an *essential matrix* can be computed using

$$\mathsf{E}_{ji} = [\mathbf{t}]_\times \mathsf{R}. \tag{3.2}$$

The antisymmetric matrix $[\mathbf{t}]_\times$ is computed from the translation vector $\mathbf{t} = [t_x, t_y, t_z]^\mathsf{T}$ using

$$[\mathbf{t}]_\times = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}. \tag{3.3}$$

The essential matrix can therefore easily be computed for a given known pose. The fundamental matrix can be computed from the essential matrix:

$$\mathsf{F}_{ji} = \mathsf{K}_i^{-\mathsf{T}} \mathsf{E}_{ji} \mathsf{K}_j^{-1} \tag{3.4}$$

where the K matrices store the internal parameters for cameras i and j so that

$$\mathsf{K} = \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.5}$$

for focal length f and principal point (u_0, v_0). This camera model assumes that pixels are square.

Figure 3.11 shows the epipolar tangents for each silhouette image of the duck example. Each line lies in

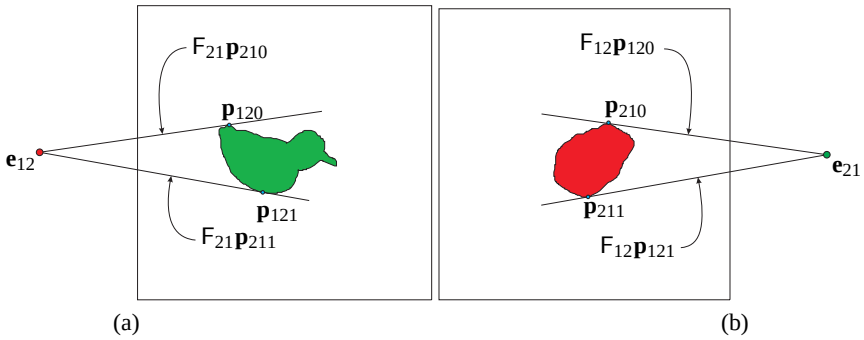


Figure 3.11: The epipolar tangency constraint: the epipolar tangent touches the silhouette at the projection of the frontier point, as shown in (a) and (b); the projection of this tangent onto the image plane of the opposite camera is constrained to coincide with the opposite epipolar tangent.

a tangent plane containing a frontier point, and therefore must project onto the corresponding line in the opposite image: this is the epipolar tangency constraint. In other words, in the noise-free case, the line passing through $\mathbf{e}_{ij}$ and $\mathbf{p}_{ijk}$ is the same line as $\mathsf{F}_{ji}\mathbf{p}_{jik}$.

3.5.2 A Measure of Inconsistency Based on Epipolar Tangents

If there are inaccuracies in the silhouettes or the pose, then the line passing through $\mathbf{e}_{ij}$ and $\mathbf{p}_{ijk}$ will not, in general, be the same line as $\mathsf{F}_{ji}\mathbf{p}_{jik}$. Figure 3.12 shows the noisy case in which there are inaccuracies in the assumed relative pose between the cameras. Note that the epipoles are positioned differently to Figure 3.11,

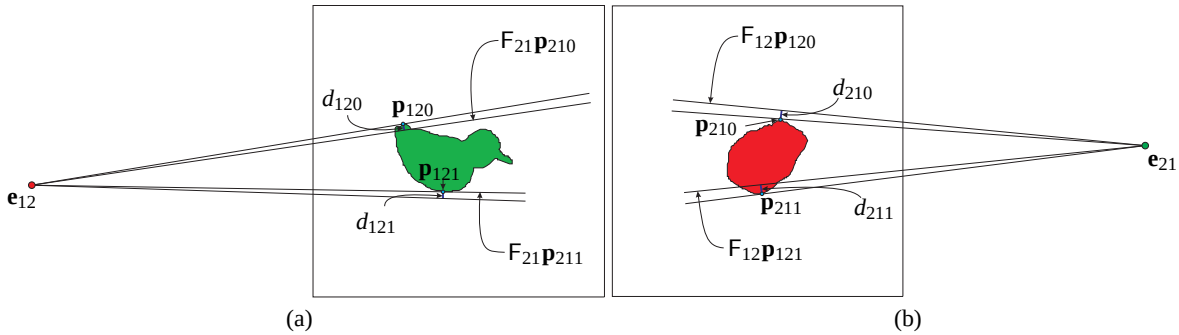


Figure 3.12: Epipolar tangents with the projection of the epipolar tangents of the opposite view and incorrect pose information: since the pose information is incorrect, the epipolar tangents do not project onto one another. The silhouettes are inconsistent with one another for the given viewpoints. The reprojection error is a measure of the degree of inconsistency.

since the pose is incorrect. The projection of the opposite camera's epipolar tangent is not exactly coincident with the epipolar tangent on the image plane. Reprojection errors can be computed as a measure of the inconsistency between a pair of silhouettes with an associated pose value. The reprojection error is the shortest distance from an epipolar tangency to the epipolar line of the corresponding point in the opposite image. The figure shows the reprojection errors d_{120}, d_{121}, d_{210} and d_{211}.

The distance d_{ijk} between an epipolar tangency $\mathbf{p}_{ijk}$ and the projection of the epipolar line from the opposite camera that passes through the tangency point $\mathbf{p}_{jik}$ can be computed using the fundamental matrix, as stated by Wong [138]:

$$d_{ijk} = \frac{\mathbf{p}_{ijk}^{\mathsf{T}}\mathsf{F}_{ij}\mathbf{p}_{jik}}{\sqrt{(\mathsf{F}_{ij}\mathbf{p}_{jik})_1^2 + (\mathsf{F}_{ij}\mathbf{p}_{jik})_2^2}}. \tag{3.6}$$

The expressions $(\mathsf{F}_{ij}\mathbf{p}_{jik})_1^2$ and $(\mathsf{F}_{ij}\mathbf{p}_{jik})_2^2$ denote the first and second elements of the vector $(\mathsf{F}_{ij}\mathbf{p}_{jik})^2$. Note that $\mathbf{p}_{ij0}$ and $\mathbf{p}_{ij1}$ are vertices of the polygon representing the silhouette.

Wong's definition of the reprojection error described by Equation 3.6 is related to the Sampson approximation [58] that provides an estimate of the locations of two projections of a point from two noisy observations. The Sampson approximation of the location of the frontier point projection is midway between the epipolar tangency and the projection of the opposite epipolar tangent. An alternative formulation that measures the distance to the Sampson approximation from the epipolar tangency (or from the projection of the opposite epipolar tangent) gives exactly half the value given by Equation 3.6. Yamazoe et al. [141] describe a method in which the locations of 3D points are explicitly modelled. However, results are not compared with the conventional method of using an error function based on Equation 3.6. Some initial experimentation indicated

that the method of Yamazoe et al. provides results that are almost identical to the those obtained with the conventional approach for the camera setups covered in this thesis. Hartley and Zisserman [58] recommend using the Sampson approximation, since it gives excellent results in practice and removes the requirement for $3n$ parameters to describe the locations of n frontier points.

A measure of inconsistency *within* a silhouette set is the RMS (root mean square) value of all reprojection errors (as specified by Equation 3.6) computed using all silhouette pairs within the silhouette set. This is referred to as the ET (epipolar tangency) error within the silhouette set, and is used for calibration (Chapters 4 and 5).

For two silhouette sets and an associated relative pose, a measure of inconsistency *across* the silhouette sets is the RMS value of all reprojection errors computed across all silhouette pairs in which one silhouette is from each silhouette set. This is referred to as the ET error across the silhouette sets, and will be used to optimise relative pose (Chapter 6).

This thesis makes use of the Levenberg-Marquardt [95] method to infer model parameters by adjusting the parameter values to minimise ET error. This approach is used in several contexts through the thesis.

3.5.3 Epipoles Inside Silhouettes

In cases in which the epipole falls within a silhouette, the ET error is not defined for the silhouette-epipole pair. Epipoles lie within silhouettes when the baseline passes through the object.

Interestingly, a configuration in which baselines connecting viewpoints all pass through the object allow consistent viewpoints to be specified for arbitrary single contour silhouettes. This is done by positioning all viewpoints on a line so that the line passes through all silhouettes. By ensuring that the viewpoints are sufficiently far apart, no silhouette will destroy the cone strip from any other silhouette; the visual cone intersection thus provides an object that exactly projects onto all silhouettes. Figure 3.13 shows an example using shapes considered by Bottino and Laurentini, who challenge readers to determine consistent orthographic views for three silhouettes: a square, a circle, and a triangle [11]. If the projection model is broadened from an orthographic model to a perspective model, then consistent viewpoints can be found for the three shapes (and indeed any number of single contour silhouettes), simply by selecting viewpoints sufficiently far apart on a common line that passes through the silhouettes.

The problem of the existence of trivial solutions, such as the one illustrated in Figure 3.13, is not an issue for the methods considered in this thesis, since silhouette consistency is never considered in cases in which the pose of individual silhouettes can be freely adjusted. In the cases in which poses are freely adjusted, either (1) multiple silhouettes correspond to each camera view, or (2) the pose of a silhouette set, rather than a single silhouette, is adjusted.

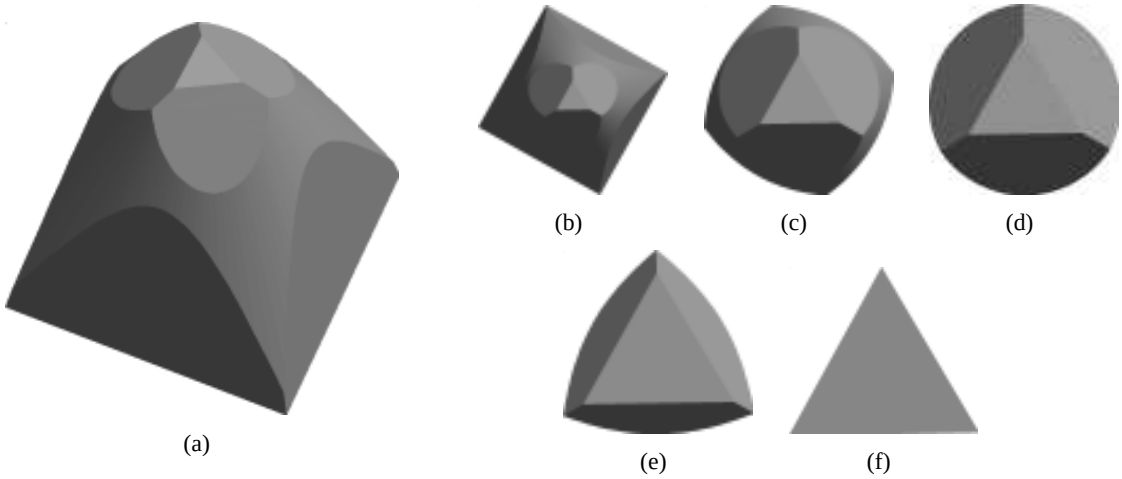


Figure 3.13: (a) A 3D shape with square, circular and triangular silhouettes from certain viewpoints. If viewed looking onto its triangular face from afar, the silhouette boundary is a square as shown in (b). As the viewpoint is moved towards the triangular face, the silhouette boundary begins to change to become more circular (c), until the boundary is a circle (d), and then becomes more triangular (e), until the viewpoint is sufficiently close to the triangular face that the silhouette is a triangle (f).

Nevertheless, it is still possible to obtain cases in which parameters are adjusted so that the epipole lies within a silhouette for certain view pairs. To prevent the number of residual values from changing within a Levenberg-Marquardt step, these cases are identified, and a residual value is chosen so that the mean square value over all epipole-outside pairs is the same as the mean square value of all residuals. To ensure efficiency, cases in which the epipole lies within the axis-aligned bounding rectangle of a silhouette are treated as if the epipole lies within the silhouette.

3.5.4 Efficiently Locating the Epipolar Tangencies

Computing the ET error requires the polygon vertices that are tangencies to be located. Since only outer tangencies are used, they are computed from the convex hulls of the polygonal silhouette boundaries. Convex hulls are efficiently computed from the boundaries using Melkman's algorithm [92] which has a time complexity of $O(n)$ for n-vertex polygons. It achieves its efficiency by assuming that input vertices lie on a non-self-intersecting polygon, rather than in general positions. Note that convex hulls need be computed only once for each silhouette, whereas tangencies need to be computed repeatedly when pose or camera parameters are adjusted within an iterative minimisation of ET error. This is why it is important to locate the tangencies efficiently.

A simple method for locating the two outer tangencies with respect to an epipole and a convex polygon is to visit each vertex and to check whether the edges arriving and leaving the current vertex are on the same side of the line through the current vertex and the epipole. If this is the case, then the current vertex is a tangency. Unfortunately, this simple method is computationally inefficient.

$x \leq 0$	$y > 0$	$v = g - 1$	$-\infty < v < -1$
$x \leq 0$	$y \leq 0$	$v = -1/(1+g)$	$-1 \leq v \leq 0$
$x > 0$	$y \leq 0$	$v = 1/(1-g)$	$0 < v \leq +1$
$x > 0$	$y > 0$	$v = g + 1$	$+1 < v \leq +\infty$

Table 3.2: Lookup value v as a function of x, y and $g = y/x$, where x and y are the vertex coordinates.

To speed up the location of tangencies, a method based on storing the edge angle associated with each vertex is proposed. (The edge angle is the angle of the vector from each vertex to its successor—the edges are *directed* and polygons are assumed to have vertices specified in anticlockwise order.) The method is applicable to ET error computed using both orthographic and perspective camera models, but the implementation is slightly different for the two camera models. Since the use of an orthographic camera model will be investigated in Chapter 7, locating tangencies with both orthographic and perspective models will be covered here.

Forming the Edge Angle Data Structure

A monotonic function of angle is computed and stored as this avoids calls to the relatively computationally expensive arctan function. A monotonic transform of the angle is sufficient as angle values are only used for ordering edges.

The monotonic function of angle is computed using the equations presented in Table 3.2. The same approach was used for efficiently computing viewing edges.

The lookup value v of the successor edge for each vertex is stored in a sorted associative container (the C++ `map` data structure was used). This allows angle values to be accessed in $O(\log n)$ complexity for an n-vertex polygon. Since n is small (the order of 100), a hashing approach which would allow $O(1)$ access was not used. (The $O(\log n)$ retrieval was found to make a negligible contribution to total running time in practice.)

Orthographic Imaging Model

In the orthographic case, the epipole is a direction, rather than a point. To determine the first tangency, a vertex must be found whose predecessor edge angle is less than the angle of the epipolar direction, and whose successor edge angle is greater than the angle of the epipolar direction. If the angle of the epipolar direction is greater than or smaller than all of the stored angles, then the relevant vertex is the vertex whose edge angles correspond to the greatest and smallest angles (this is caused by the discontinuity of v between $-\infty$ and $+\infty$). The located vertex is a tangency, since its two edges lie on the same side of the line specified by the vertex and the epipolar direction.

The second tangency is located by applying the same procedure to the direction opposite to the epipolar direction.

Perspective Imaging Model

The above method relies on the tangent direction being known in advance: it is the same as the epipolar direction. In the perspective case, the epipole is not a direction, and the tangent direction is therefore not known in advance. Instead, an approximate direction is computed as the direction from the epipole to the silhouette centroid. If the epipole is sufficiently far from the silhouette, this will lead to the correct vertex being located. However this is not guaranteed.

The located vertex must therefore be checked to determine whether it is a tangency. This is done by checking whether its two edges lie on the same side of the line passing through the vertex and the epipole.

If the vertex is not a tangency, then the direction from the epipole to the vertex is used to find the next candidate. The candidate direction therefore rotates clockwise until the tangency is found. Typically, the tangency is found on the first iteration, but in cases where the epipole is close to the silhouette more than one iteration may be required. An example is shown in Figure 3.14.

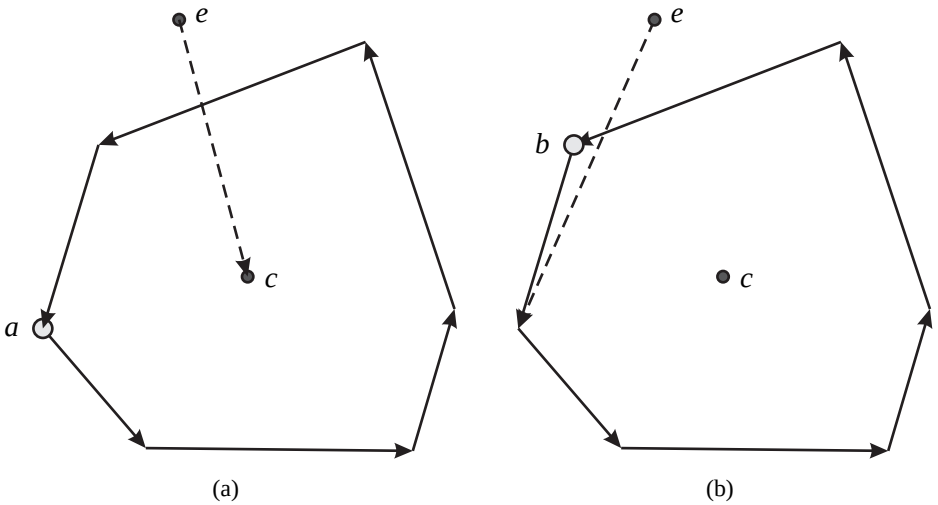


Figure 3.14: Clockwise rotation of the candidate direction for finding the tangency: (a) the initial candidate direction from the epipole *e* to the silhouette centroid *c*; the located vertex *a* is not a tangency, so it is used to define the candidate direction for the next iteration, (b) the located vertex *b* is a tangency, so the algorithm terminates.

Note that the tangency is always located, since the current candidate direction always locates a candidate vertex that is further clockwise than the vertex that specifies the direction. For any direction there are two tangencies: one to the left and one to the right. Since the polygon vertices are ordered anticlockwise, the rightmost tangency will always be selected. This is because the range of directions between the edges arriving and leaving the rightmost tangency vertex includes one that is parallel to the current direction, whereas the range corresponding to the leftmost tangency vertex includes one that is antiparallel. Since, from the point of view of the epipole, a step to the right is always a clockwise turn (since the epipole is outside the silhouette), the tangency will always be located.

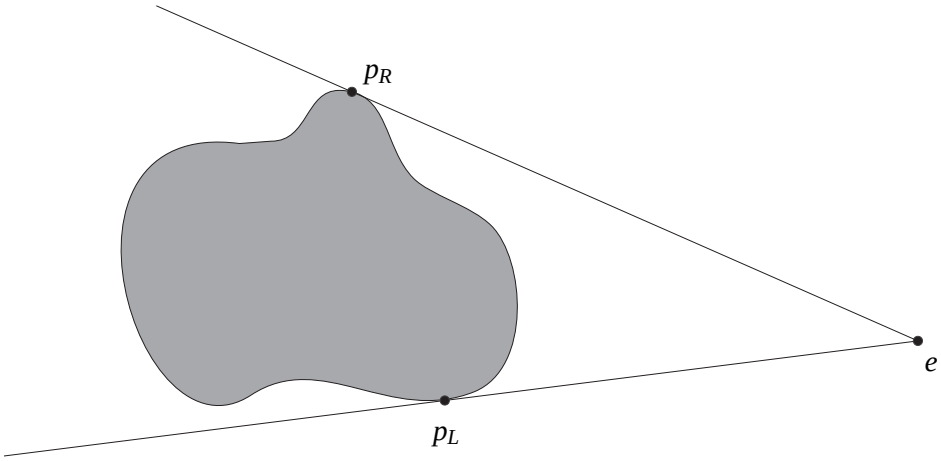


Figure 3.15: Two outer tangencies: p_L has a greater angular extent to the left (anticlockwise) and p_R has a greater angular extent to the right (clockwise) with respect to the epipole.

The second tangency is located by computing direction vectors from the silhouette to the epipole, rather than from the epipole to the silhouette as is used to locate the first tangency.

3.5.5 Determining Tangency Correspondences

To compute the ET error it is necessary to know which of the two outer tangencies in one image of a pair corresponds to which outer tangency in the other pair.

The literature mentions two approaches to solving the correspondences: (1) the correspondence that leads to the lowest ET error is selected [54], or (2) correspondences are determined by knowing that cameras are always upright: one pair will occur at the top of the image and another pair at the bottom [60].

Since there is no prior knowledge of what is upright for the camera views considered in this thesis, this constraint cannot be used to determine correspondences.

Instead of determining correspondences by selecting the pair with the lowest ET error, this section demonstrates that epipolar tangency correspondences can be determined by considering the camera poses alone. This provides a simpler algorithm.

A Method for Determining Correspondences

Figure 3.15 illustrates the two epipolar tangencies computed from a silhouette and an epipole. From the point of view of the epipole, one of the tangencies p_L will have a greater angular extent to the left (anticlockwise) and the other tangency p_R will have a greater angular extent to the right (clockwise). An alternative definition is that the epipole lies to the right of silhouette normal at p_R, and to the left of silhouette normal at p_L.

Let the tangencies from the first image be p_{LA} and p_{RA}, and those from the second be p_{LB} and p_{RB}. If exactly one camera is behind the other camera (i.e., if one camera's z coordinate specified in the other camera's reference frame is less than zero), then the correspondences are (p_{LA}, p_{LB}) and (p_{RA}, p_{RB}). If both cameras are facing each other or if both cameras are behind one another, the correspondences are (p_{LA}, p_{RB}) and (p_{RA}, p_{LB}). Examples illustrating the different configurations are shown in Figure 3.16.

Since orthographic cameras are at infinity, they are never behind one another, and the correspondences are therefore (p_{LA}, p_{RB}) and (p_{RA}, p_{LB}).

An Explanation of the Method

Consider a frontier point P that is generated by Cameras A and B. The surface normal of the frontier point is used to define the up direction, so that it can be specified whether a camera lies to the left or to the right of a line passing through the other camera and the frontier point. Figures 3.17a and 3.17b illustrate points in the plane containing A, B and P. In this case, A lies to the right of the line PB, as the normal at P is facing out of the page.

If Camera B is oriented so that A is in front of Camera B (as illustrated in Figure 3.17a), then the epipole e_{BA} (the image of A) is on the same side of p_B (the image of P) as A is of P. (In the example illustration, it is to the right). This is because both P and A are in front of Camera B. The point P is always in front of both cameras, since it is visible to both cameras.

If Camera B is oriented so that A is behind Camera B (as illustrated in Figure 3.17b), then the epipole e_{BA} (the image of A) is on the opposite side of p_B (the image of P) as A is of P. This is because the ray from A comes from behind Camera B and strikes the image plane from behind.

The handedness of an epipole with respect to the tangency is used to specify the handedness of the tangency. This is illustrated in Figures 3.17c and 3.17d. In the case of Figure 3.17c for example, the relevant epipole is $pB = pR$, the tangency for which p_B lies to the right of the epipole.

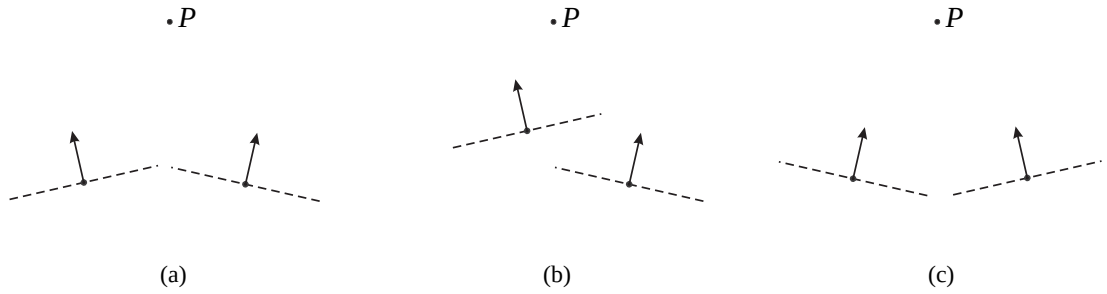


Figure 3.16: Examples illustrating (a) two cameras each behind the other, (b) one camera behind and one camera facing, (c) two cameras both facing the other. Dashed lines represent each camera's $z = 0$ plane and arrows specifying the camera directions depart from the camera centres and lie on the optical axes. In all cases, the external scene point P is visible to both cameras.

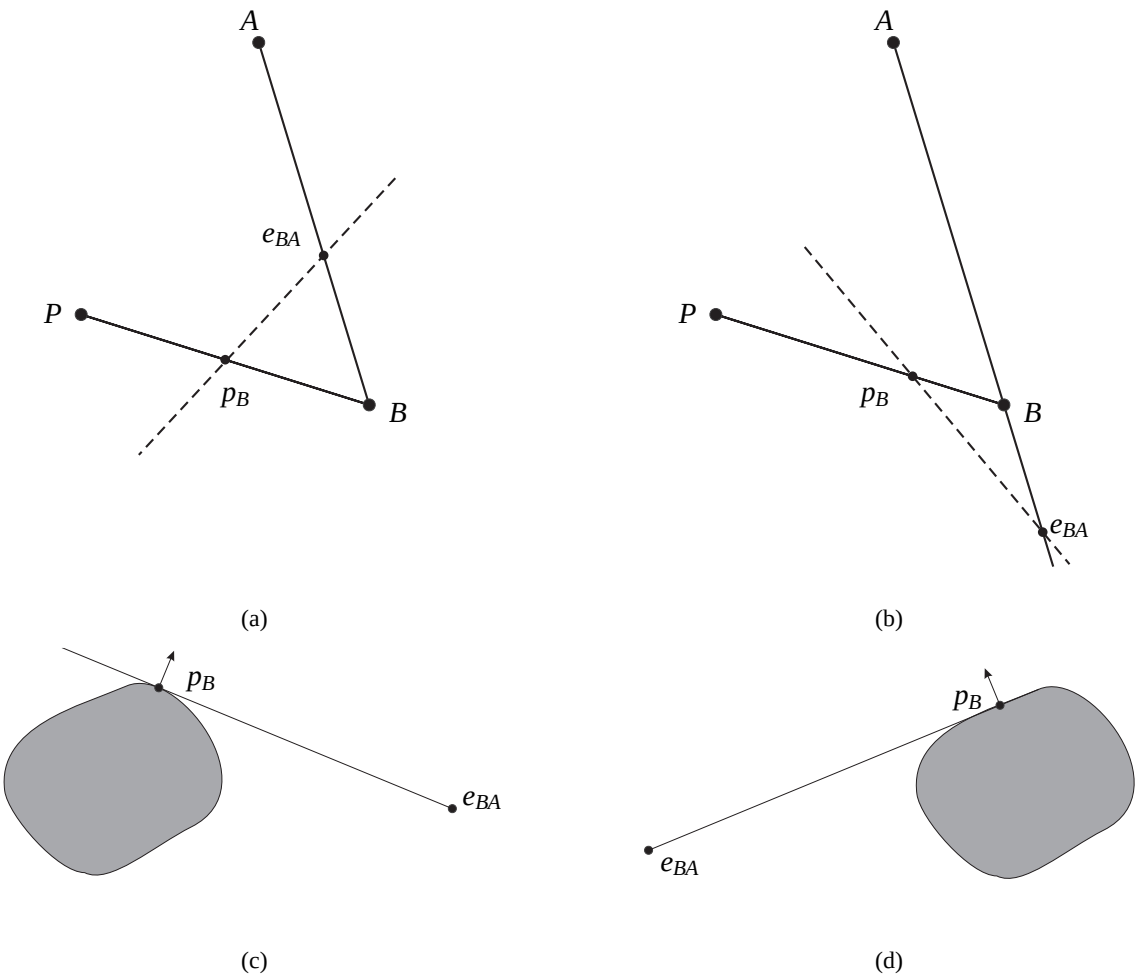


Figure 3.17: Camera *A* is to the right of *B* with respect to the line *PB* and with the normal to frontier point *P* facing upwards (out of the page): (a) Camera *B* is oriented so that *A* is in front of it; (b) Camera *B* is oriented so that *A* is behind it; (c) the image seen by *B* for the configuration in (a); (d) the image seen by *B* for the configuration in (b). The image plane of *B* is indicated with a dashed line. Note that cameras are modelled with the image plane in front of the camera centre.

Since the handedness of *A* with respect to *PB* is the opposite of *B* with respect to *PA*, the corresponding tangencies have opposite handedness if both cameras are in front of each other. However, if exactly one of the cameras is behind the other, the handedness of one of the tangencies flips, and the corresponding tangencies have the same handedness. If both of the cameras are behind the other, then the handedness of both of the tangencies flips, and the corresponding tangencies have opposite handedness (as for the case when both cameras are in front of one another).

3.6 Summary

This chapter has covered the main aspects of the geometry of silhouette sets that will be used throughout this thesis. First, the visual hull, a widely-used approximation of 3D shape computed from silhouette sets, was introduced. Next, it was shown that viewing edges impose constraints on the shape of the object that produced the silhouettes. Viewing edges were demonstrated to provide both a means for computing bounds on the longest and shortest diameters of a stone, and for computing an approximation to the convex hull of the 3D shape of the stone, the VEMH. In later chapters, the VEMH will be used for pose optimisation, approximating shape properties, and recognition tasks.

The ET error, a measure of silhouette inconsistency that is based on the epipolar tangency constraint was introduced. The ET error plays an important role in the calibration and recognition methods that are developed in the chapters that follow. Efficient algorithms for computing ET error that incorporate some new ideas have been described.

Chapter 4

Multiple Views from Mirrors

4.1 Introduction

Multiple silhouette images of particles for silhouette-based analysis are typically captured using a multi-camera setup [108]. Such equipment is often not readily available, and a simpler acquisition system may be beneficial for early investigations. For this reason, a simple setup using only two plane mirrors and a single digital camera was used for initial data acquisition.

In addition to providing a means for capturing calibrated silhouette sets of particles, the method can be used to create 3D visual hull models of objects for other applications such as 3D multimedia content creation. Other shape-from-silhouette methods [91, 97, 99] for 3D content creation typically make use of calibration objects, turntables, or synchronised multi-camera setups. The proposed setup provides a simple way of creating 3D multimedia content that does not rely on specialised equipment. The setup need not be accurately positioned, since self-calibration is used to determine all pose and internal parameters*.

Two mirrors are used to create five views of an object: a view directly onto the object, two reflections, and two reflections of reflections (see Figure 4.1). The image is segmented into foreground and background regions producing an image containing five silhouette sub-images.

The method presented in this chapter describes how the internal camera parameters and pose associated with each of the five silhouette views can be determined from the silhouette outlines alone. This means that self-calibration is possible: no calibration markers are required. The method therefore allows a 5-view visual hull model to be computed from a single snapshot of the scene.

By moving the camera, yet keeping the object and mirrors in the same positions, silhouettes from different viewpoints can be captured. The relative pose of the camera can be computed for the different shots, allowing

*Matlab software to perform the self-calibration is available from `http://www.dip.ee.uct.ac.za/~kforbes/`.

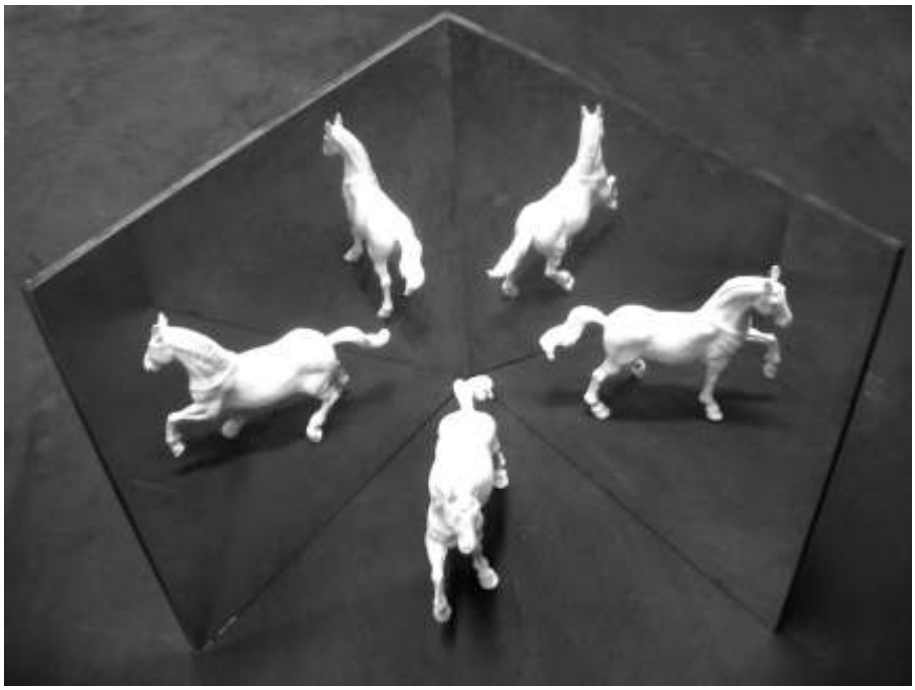


Figure 4.1: The two-mirror setup used to capture five views of an object in a single image.

silhouette sets with an arbitrary number of silhouettes to be captured. Figure 4.2 shows an example of two images of a scene captured from different viewpoints, allowing a 10-view silhouette set to be formed.

Another approach is to change the pose of the object between shots to capture different viewpoints. Chapter 6 explains how multiple 5-view sets can be merged into a single set.

Part of the work described in this chapter was presented as a conference paper [44]. This is an extension of earlier work that was presented as another conference paper [47]. The earlier work assumes an orthographic projection model and requires a dense search of parameter space to determine initial estimates. The method described in this chapter improves on this earlier work by providing closed form solutions for the initial parameter estimates using a perspective camera model.

4.2 Related Work

The computer vision literature describes various approaches for capturing silhouettes of an object from multiple viewpoints so that shape-from-silhouette reconstruction can be applied. Several approaches make use of self-calibration: the silhouettes themselves are used to estimate camera pose and internal parameters. Rather than assuming general poses for all silhouettes, these approaches typically make use of problem-specific constraints such as circular motion, known orientation, or coplanar viewing directions. The method described in

Figure 4.2: Two images of a two-mirror setup positioned so that five views of the object can be seen. Note that the camera has moved between shots, but the mirrors and object have not moved.

this chapter also makes use of problem specific constraints. The constraints in this case are imposed by the mirror configuration that is used to produce multiple views.

Wong and Cipolla [139] describe a system that is calibrated from silhouette views using the constraint of circular motion. Once an initial visual hull model is constructed from an approximately circular motion sequence, additional views from arbitrary viewpoints can be added to refine the model. The user must manually provide an approximate initial pose for each additional view which is then refined using an iterative optimisation. Their method of minimising the sum-of-square reprojection errors corresponding to all outer epipolar tangents is used in this chapter to provide a refined solution.

Okatani and Deguchi [101] use a camera with a gyro sensor so that the orientation component associated with each silhouette view is known. An iterative optimisation method is then used to estimate the positional component from the silhouettes by enforcing the epipolar tangency constraint.

Bottino and Laurentini [11] provide methods for determining viewpoints from silhouettes for the case of orthographic viewing directions parallel to the same plane. This type of situation applies to observing a vehicle on a planar surface, for instance.

Many works describe the use of mirrors for generating multiple views of a scene. For example, Gluckman and Nayar [53] discuss the geometry and calibration of a two-mirror system using point correspondences. Han and Perlin [55] use a kaleidoscope to simultaneously view a surface from many directions. This allows the bidirectional texture function to be computed without mechanical movement. Hu et al. [62] describe a setup similar to ours, however they use constraints imposed by both the silhouette outlines and point correspondences for calibration.

Huang and Lai [63] have also extended our original two-mirror setup [47] to use a full perspective camera model (as described in this chapter). However, their approach is different and was developed entirely independently of our work (and was published subsequent to both our original method and our full perspective method [44]). Their method of solving for the orientations is based on the equations involving the mirror

normals, and is similar to our original algorithm for an orthographic projection model. Huang and Lai's method requires a least squares cost function to be minimised to estimate the focal length. This means that an initial estimate of the focal length must be provided. The method described in this chapter provides a closed form solution for the focal length.

Moriya et al. [96] describe an idea that is related to the work described in this chapter. Epipoles are computed from the silhouette outlines of three shadows of a solid cast onto a plane, and are shown to be collinear. The authors do not, however, mention any applications that can be derived from their observed collinearity constraint.

4.3 Epipoles from Bitangents

This section deals with the case where a camera views an object and its reflection. It is shown how the epipole corresponding to the virtual camera (the reflection of the real camera) can be computed directly from the silhouette outlines of the real object and the virtual object in the image captured by the real camera. This result will be used to calculate the positions of epipoles for the two-mirror setup.

Figure 4.3 shows an example of a camera observing a real object and its reflection in a mirror. The virtual camera is also shown in the figure. Consider a plane Π_1 that passes through the camera centres C_R and C_V

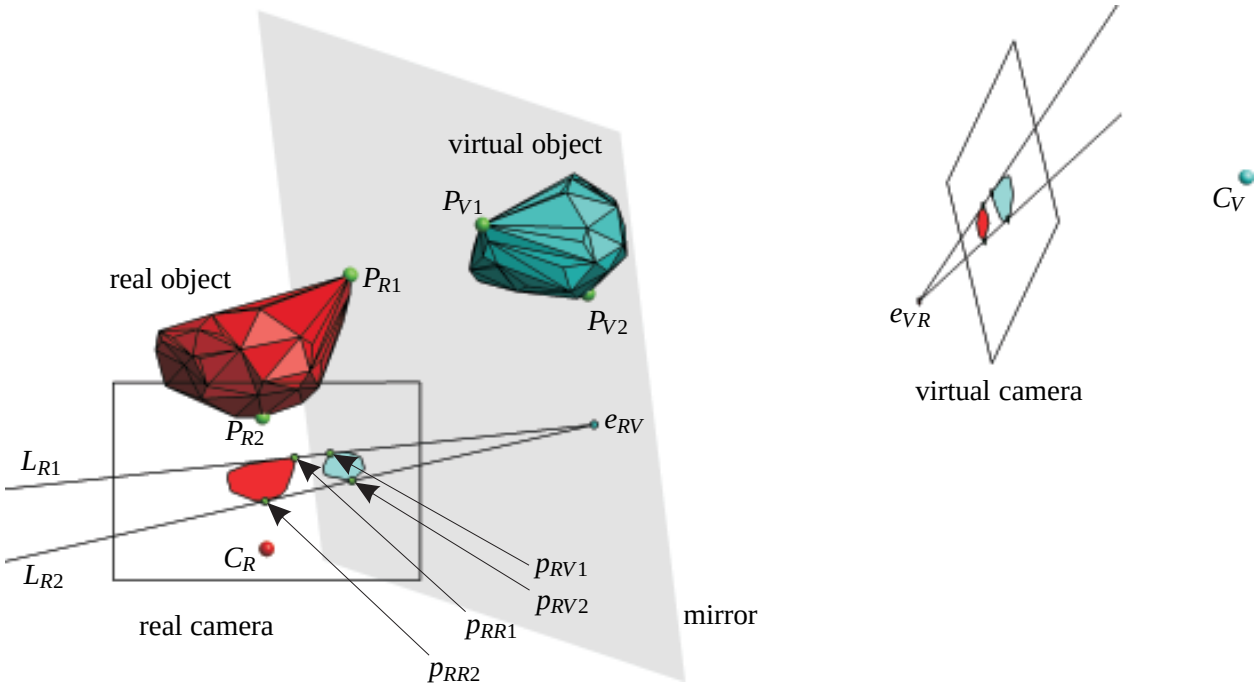


Figure 4.3: A camera viewing an object and its reflection. The epipole corresponding to the virtual camera can be computed from the silhouette bitangents L_{R1} and L_{R2}.

and touches the real object at the point P_{R1}. By symmetry, Π_1 will touch the virtual object at the point P_{V1} which is the reflection of P_{R1}. Since Π_1 is tangent to both objects and contains the camera centres C_R and C_V, P_{R1} and P_{V1} are frontier points. They project onto the silhouette outlines on the real image at points p_{RR1} and p_{RV1}. The points p_{RR1}, p_{RV1} and the epipole e_{RV} (the projection of C_R into the real image) are therefore collinear, since they lie in both Π_1 and the real image plane. Observe that the bitangent L_{R1} passing through these three points can be computed directly from the silhouette outlines: it is simply the line that is tangent to both silhouettes. Another bitangent L_{R2} passes through the epipole and touches the silhouettes on the opposite side to L_{R1}. These tangency points lie on a plane Π_2 that is tangent to the opposite side of the object and passes through both camera centres. Provided that the object does not intersect the line passing through both camera centres, there will always be two outer epipolar tangents L_{R1} and L_{R2} that touch the silhouettes on either side.

The position of the epipole e_{RV} can therefore be computed by determining L_{R1} and L_{R2} from the silhouette outlines: it is located at the intersection of L_{R1} and L_{R2}. Note that the epipole is computed without requiring knowledge of the camera pose and without requiring any point correspondences.

Also note that, by symmetry, the real camera's silhouette view of the virtual object is a mirror image of the virtual camera's silhouette view of the real object. The silhouette view observed by a reflection of a camera is therefore known if the camera's view of the reflection of the object is known.

4.4 Two-Mirror Setup

Figure 4.4 shows an example of a two-mirror setup that is used to capture five silhouette views of an object in a single image. The camera is centred at C_R and observes a real object O_R. The camera also captures the image of each of four virtual objects O_{V1}, O_{V2}, O_{V12}, and O_{V21}. Object O_{V1} is the reflection of O_R in Mirror 1; O_{V2} is the reflection of O_R in Mirror 2; O_{V12} is the reflection of O_{V1} in Mirror 2; and O_{V21} is the reflection of O_{V2} in Mirror 1.

The proposed method requires six virtual cameras to be considered. The virtual cameras are reflections of the real camera C_R. The virtual cameras C_{V1}, C_{V2}, C_{V12}, and C_{V21} are required, as their silhouette views of the real object are the same as the silhouettes observed by the real camera (or reflections thereof). Since silhouettes from the real camera are accessible, the silhouettes observed by the four virtual cameras can be determined. Each of the five cameras' silhouette views of the real object can then be used to compute the five-view visual hull of the object.

The virtual cameras C_{V121} (the reflection of C_{V12} in Mirror 1), and C_{V212} (the reflection of C_{V21} in Mirror 2) are to be considered too, since it turns out that their epipoles can be computed directly from the five silhouettes observed by the real camera. These epipoles, together with the epipoles from the virtual cameras C_{V1} and C_{V2} can then be used to calculate the focal length of the camera.

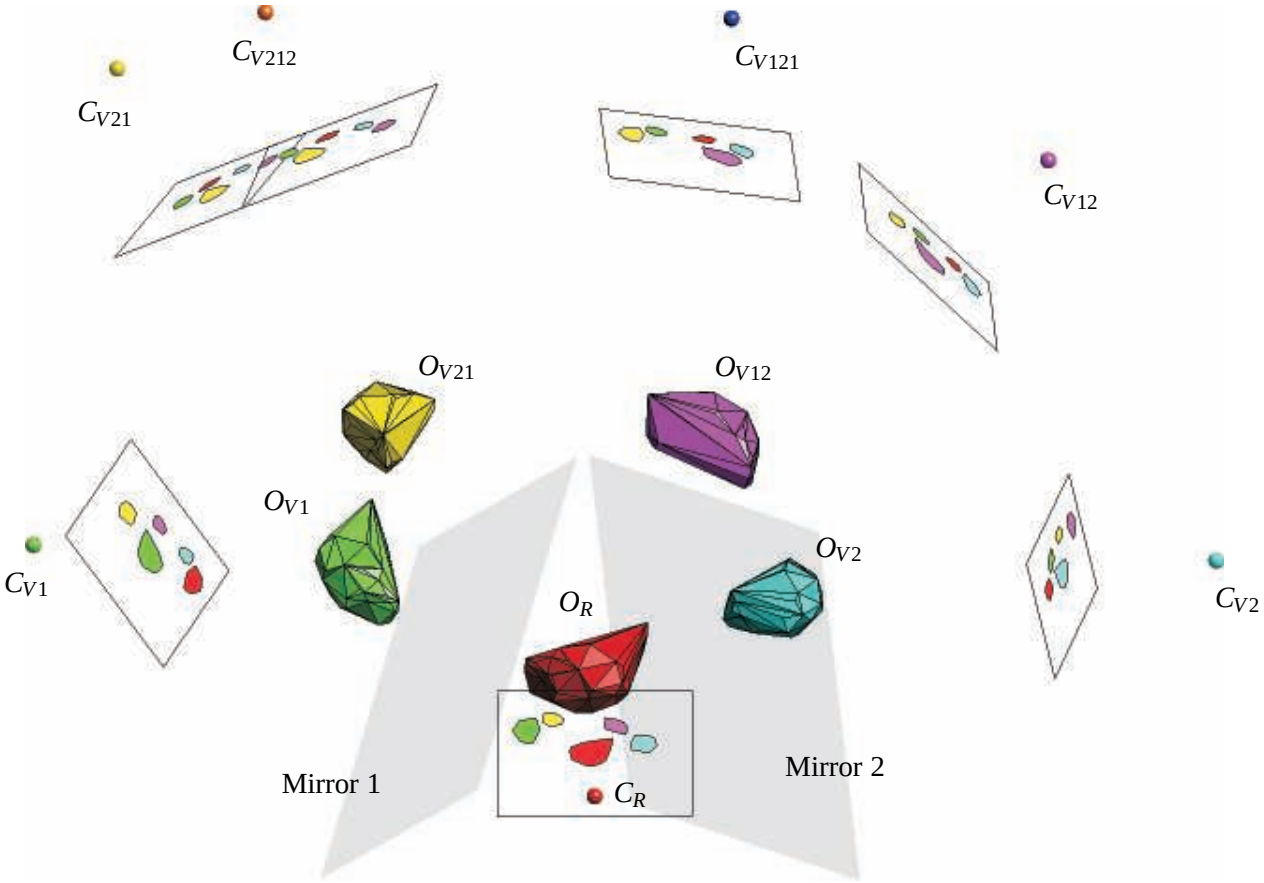


Figure 4.4: Mirror setup showing one real and four virtual objects, and one real and six virtual cameras.

4.5 Analytical Solution

This section presents a method to calculate the focal length and principal point of the camera and the poses of the virtual cameras relative to the pose of the real camera for the five camera views in an image. Next, a method for determining camera motion between snapshots is presented. This allows all silhouettes from all images to be specified in a common reference frame. Closed form solutions in which the required parameters are determined from the silhouette outlines alone are provided. Silhouette outlines are represented by polygons, and pixels are assumed to be square.

First, it is demonstrated how lines that are tangent to pairs of silhouettes can be used to calculate the position of four epipoles corresponding to four virtual cameras. The principal point is constrained by the epipoles to a line in each image; the intersection of the lines is the principal point. Next, it is shown how the focal length is a function of the relative positions of these four epipoles. Once the focal length is known, it is shown that mirror and camera orientation are easily determined from the positions of two epipoles. The positional component of the poses can be computed using the epipolar tangency constraint. Finally, it is shown how

the camera poses between shots are constrained by the constant positions of the mirrors with respect to the object.

4.5.1 Four Epipoles from Five Silhouettes

Here, it is shown how the epipoles are computed from pairs of silhouettes using the result explained in Section 4.3: the epipole corresponding to a camera's reflection can be computed from the camera's silhouette image of an object and its reflection by finding the intersection of the two outer bitangents. Figure 4.5 shows how the epipoles e_{V1}, e_{V2}, e_{V121}, and e_{V212} are computed from the outlines of the five silhouettes observed by the real camera. The distances a, b, and c between the epipoles will be used for computing the focal length. The outline γ_{RR} corresponds to the object O_R, and γ_{RV1} corresponds to O_{V1} which is the reflection of O_R in Mirror 1. The intersection of the pair of lines that are tangent to both γ_{RR} and γ_{RV1} is therefore the epipole e_{V1}, since C_{V1} is the reflection of C_R in Mirror 1. The two lines that are tangent to both γ_{RV2} and γ_{RV21} also meet at e_{V1}, since O_{V21} is the reflection of O_{V2} in Mirror 1. Similarly, the pairs of lines that are tangent to both γ_{RR} and γ_{RV2}, and to γ_{RV1} and γ_{RV12} meet at e_{V2}.

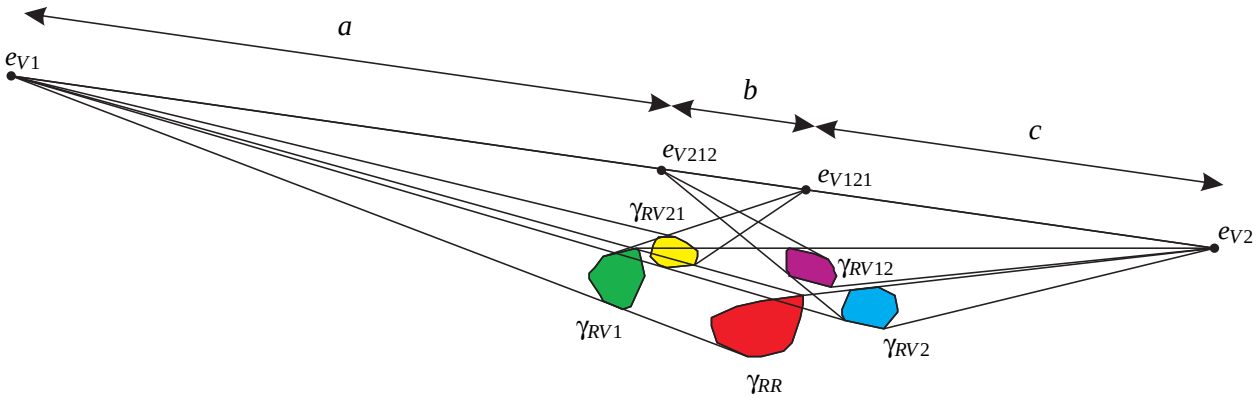


Figure 4.5: Computing epipoles e_{V1}, e_{V2}, e_{V121}, and e_{V212} from the silhouette outlines in an image.

Consider C_R observing O_{V1}. Object O_{V21} is related to O_{V1} through three reflections. Object O_{V1} must be reflected by Mirror 1 (to get O_R) and then Mirror 2 (to get O_{V2}) and then again by Mirror 1 to get O_{V21}. The effect of these three reflections can be considered to be a single reflection. Applying the triple reflection to C_R gives C_{V121}. The two lines that are tangent to both γ_{RV1} and γ_{RV21} therefore meet at e_{V121}. This is again because a camera (C_R) is observing silhouettes of an object (O_{V1}) and its reflection (O_{V12}), so the projection of the camera's reflection (C_{V121}) can be computed from the silhouette bitangents. Similarly, the two lines that are tangent to both γ_{RV2} and γ_{RV12} meet at e_{V212}.

Note that the epipoles e_{V1}, e_{V2}, e_{V121}, and e_{V212} are collinear, since they all lie in both the image plane of the real camera and in the plane Π_C in which all camera centres lie.

4.5.2 Focal Length and Principal Point from Epipoles

It is now shown how the focal length is computed from the positions of the four epipoles e_{V1}, e_{V2}, e_{V121}, and e_{V212}. This is done by considering the positions of the camera centres in the plane Π_C.

First, two new mirrors, Mirrors A and B, which do not correspond to physical mirrors in the scene, are introduced. This approach makes the problem of calculating the focal length tractable. Mirror A has the same orientation as Mirror 1, but is positioned so that it passes midway between e_{V1} and C_R (see Figure 4.6a in which the positions of points in Π_C are shown). The point e_{V1} is therefore the reflection of C_R in Mirror A.

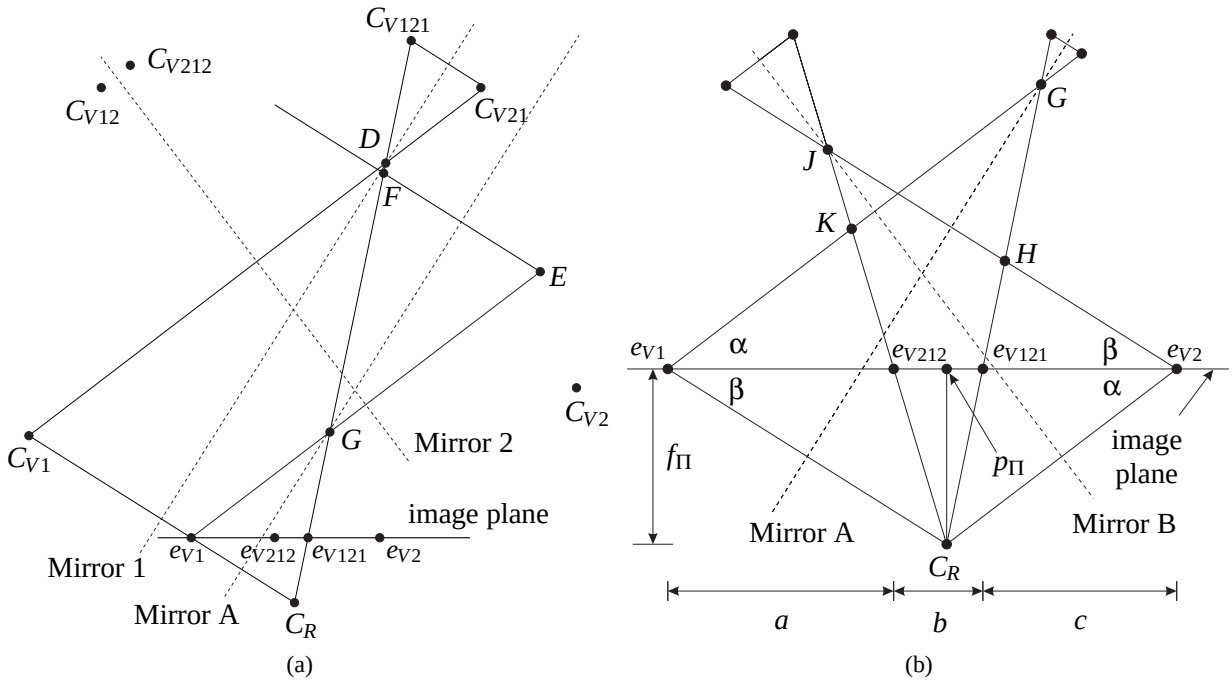


Figure 4.6: Diagrams showing (a) the intersections of Mirror 1, Mirror A and Mirror 2 with Π_C along with the positions of the cameras and epipoles, all of which lie in Π_C, and (b) computing f_π and p_π from the four epipoles e_{V1}, e_{V2}, e_{V121}, and e_{V212}

Point E is the reflection of e_{V1} in Mirror 2, and F is the reflection of E in Mirror A. Note that F lies on the ray passing through e_{V121} and C_R. Also note that F will stay on this line if the position (but not the orientation) of Mirror 2 changes. This is because triangles $\triangle C_R C_{V1} D$ and $\triangle C_R e_{V1} G$ are similar.

Figure 4.6b shows the positions of the epipoles and C_R in Π_C. The distances a, b, and c between the epipoles (as shown in the figure) are used to compute the distance f_Π between C_R and the image plane in the plane Π_C. The distance f_Π is then used to calculate the focal length. The figure also shows Mirror B which has the same orientation as Mirror 2, and is positioned midway between C_R and e_{V2}. The line joining e_{V2} to its reflection in Mirror A meets Mirror B at point J which projects onto e_{V212}.

The triangle $\triangle He_{V2}C_R$ is similar to $\triangle C_R e_{V1} G$, the line segment from e_{V121} to e_{V2} is of length c, and the line segment from e_{V1} to e_{V121} is of length $a+b$. This indicates that the ratio of the sides of $\triangle He_{V2}C_R$ to $\triangle C_R e_{V1} G$ is $c:(a+b)$. This means that $d(e_{V1},G) = d(C_R,e_{v2})(a+b)/c$. (The notation $d(x,y)$ indicates the distance between x and y.)

Similarly, the triangle $\triangle Ke_{V1}C_R$ is similar to $\triangle C_R e_{V2} J$, the line segment from e_{V1} to e_{V212} is of length a, and the line segment from e_{V212} to e_{V2} is of length $b+c$. This indicates that the ratio of the sides of $\triangle Ke_{V1}C_R$ to $\triangle C_R e_{V2} J$ is $a:(b+c)$. Therefore $d(e_{V2},J) = d(C_R,e_{V1})(b+c)/a$.

This allows $d(C_R,e_{V1})$ to be written in terms of $d(C_R,e_{V2})$, since $\triangle C_R e_{V2} J$ is similar to $\triangle C_R e_{V1} G$:

$$d(C_R,e_{V1}) = \frac{\sqrt{c\,(c+b)\,a\,(a+b)}}{c\,(c+b)} d(C_R,e_{V2}). \tag{4.1}$$

The sides of $\triangle C_R e_{V1} G$ are now known up to a scale factor.

The angle $\angle C_R e_{V1} G = \alpha+\beta$ can be computed using the cosine rule:

$$\cos(\alpha+\beta) = 1/2\,\frac{\sqrt{c\,(c+b)\,a\,(a+b)}}{(c+b)\,(a+b)}. \tag{4.2}$$

The cosine rule can be used to determine the sides of $\triangle e_{V1}C_R e_{V2}$. (The angle $\angle e_{V1}C_R e_{V2} = 180^\circ - \alpha - \beta$.)

The value of f_Π can now be stated in terms of a, b, and c (with the help of the Matlab Symbolic Toolbox for simplification):

$$f_\Pi = 1/2\,\frac{(a+b+c)\sqrt{ac(3ac+4ab+4bc+4b^2)}}{a^2+ab+c^2+bc+ac}. \tag{4.3}$$

The point closest to C_R on the line containing the epipoles, is

$$\mathbf{p}_\Pi = \mathbf{e}_{V1} + 1/2\,\frac{(2a+2b+c)\,a\,(a+b+c)}{a^2+ab+c^2+bc+ac}\frac{\mathbf{e}_{V2}-\mathbf{e}_{V1}}{||\mathbf{e}_{V2}-\mathbf{e}_{V1}||}. \tag{4.4}$$

The line passing through p_Π and perpendicular to the line containing the epipoles passes through the principal point p_0. The principal point can therefore be computed as the intersection of two such lines from two images of the scene. (If the principal point is assumed to lie at the image centre, then a single snapshot could be used.)

The focal length (the distance from C_R to the image plane) can now be calculated from $\mathbf{p}_\Pi$, the principal point $\mathbf{p}_0$ and f_Π (see Figure 4.7):

$$f = \sqrt{f_\Pi^2 - ||\mathbf{p}_0 - \mathbf{p}_\Pi||^2}. \tag{4.5}$$

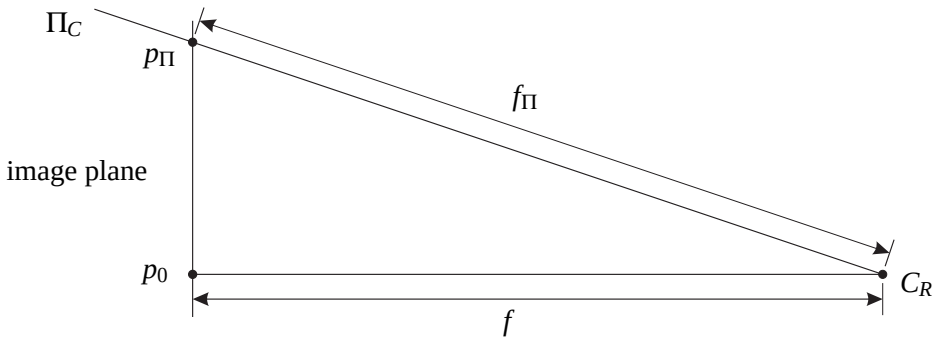


Figure 4.7: View of the camera perpendicular to both the image plane and Π_C.

4.5.3 View Orientations

Once the focal length of the camera has been calculated, the view orientation can be computed relatively easily. The mirror normal directions $\mathbf{m}_1$ and $\mathbf{m}_2$ are computed from the focal length, the principal point $\mathbf{p}_0$ and the epipoles $\mathbf{e}_{V1}$ and $\mathbf{e}_{V2}$:

$$\mathbf{m}_1 = -\begin{bmatrix} \mathbf{e}_{V1} - \mathbf{p}_0 \\ f \end{bmatrix}, \quad \mathbf{m}_2 = -\begin{bmatrix} \mathbf{e}_{V2} - \mathbf{p}_0 \\ f \end{bmatrix}. \tag{4.6}$$

A 3×3 matrix R that represents a reflection by a mirror with unit normal $\hat{\mathbf{m}} = [m_x, m_y, m_z]^T$ is used to calculate view orientation:

$$\mathsf{R} = \begin{pmatrix} -m_x^2 + m_y^2 + m_z^2 & -2m_x m_y & -2m_x m_z \\ -2m_x m_y & m_x^2 - m_y^2 + m_z^2 & -2m_y m_z \\ -2m_x m_z & -2m_y m_z & m_x^2 + m_y^2 - m_z^2 \end{pmatrix}. \tag{4.7}$$

4.5.4 View Positions

The point C_{V1} is constrained to lie on the line passing through e_{V1} and C_R. Similarly, the point C_{V2} is constrained to lie on the line passing through e_{V2} and C_R. Since absolute scale cannot be inferred from the image (if the scene were scaled, the image would not change), C_{V1} is fixed at unit distance from C_R. The only positional unknown across the entire setup is now the position of C_{V2} on the line passing through e_{V2} and C_R.

To solve for w, the distance from C_R to C_{V2}, the epipolar tangency constraint is used. This constraint requires that a tangent to a silhouette outline that passes through the epipole must be tangent to the corresponding point in its projection into the image plane of the other view. The relationship between the silhouette views of cameras C_{V1} and C_{V2} is used to enforce this constraint.

The poses of the cameras C_{V1} and C_{V2} are specified by 4×4 rigid transform matrices from the reference frame of the real camera:

$$\mathsf{M} = \begin{pmatrix} \mathsf{R} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{pmatrix}, \tag{4.8}$$

where the translational component $\mathbf{t}$ is given by $\mathbf{t} = 2(m_x p_x + m_y p_y + m_z p_z)(m_x, m_y, m_z)^T$ and $(p_x, p_y, p_z)^T$ is a point on the mirror.

The matrix $\mathsf{M}_1\mathsf{M}_2^{-1}$ represents the rigid transform from the reference frame of C_{V2} to that of C_{V1}.

The point p_{V2} is one of two outer epipolar tangencies formed by lines passing through e_{V2V1} (the projection of C_{V1} onto the image plane of camera C_{V2}) and tangent to the silhouette observed by the camera C_{V2}.

The point p_{V1V2} is the projection of p_{V2} into camera C_{V1}. It must correspond to p_{V1}, one of two outer epipolar tangencies formed by lines passing through e_{V1V2} (the projection of C_{V2} onto the image plane of camera C_{V1}).

The epipolar tangency constraint is expressed as

$$(\mathbf{p}_{V1V2} \times \mathbf{e}_{V1V2}) \cdot \mathbf{p}_{V1} = 0, \tag{4.9}$$

where $\mathbf{p}_{V1V2}$, $\mathbf{e}_{V1V2}$, and $\mathbf{p}_{V1}$ are represented by homogeneous coordinates. In other words, the line passing through $\mathbf{p}_{V1V2}$ and $\mathbf{e}_{V1V2}$ must also pass through $\mathbf{p}_{V1}$.

Equation 4.9 can be specified in terms of p_{V1}, p_{V2}, the computed orientation and camera internal parameters, and w. The Matlab Symbolic Toolbox was used to determine a solution for w (the equation is too large to reproduce here). Unfortunately, the values of both p_{V1} and p_{V2} are unknown, since the epipoles from which they may be computed are functions of the unknown w.

The values of p_{V1} and p_{V2} can be determined by exhaustive search, by finding the polygon vertex pair that fulfils the epipolar tangency constraint. Instead, the need for an exhaustive search is removed by using a parallel projection approximation to determine approximate correspondences. The tangencies are selected as the support points for outer tangent pairs that are parallel to the projected viewing direction. Unless the camera is very close to the object, this method selects either the same vertices, or vertices very close to the true tangencies under a perspective projection.

4.5.5 Combining Five-View Silhouette Sets

The calibration procedure described above allows five silhouette views from one image to be specified in a common reference frame. The pose and internal parameters of the four virtual cameras and one real camera

are known. The silhouettes observed by these cameras are also known: the silhouettes observed by the virtual cameras are those observed by the real camera of the corresponding virtual object.

The next step is to specify the silhouette sets from two or more images in a common reference frame. This is easily achieved, since the mirror poses are known with respect to the real camera for each image. The five-view silhouette sets are aligned by aligning the mirrors across sets. There are two additional degrees of freedom that the mirrors do not constrain: a translation along the join of the mirrors, and an overall scale factor. These are approximated using the epipolar tangency constraint and a parallel projection model (as for computing w): each five-view silhouette set is scaled and translated along the mirror join so that outer epipolar tangents coincide with the projected tangents from silhouettes in the other silhouette set. Each silhouette pair between silhouettes in different sets provides an estimate of translation and scale. The average result over all pairings is used.

4.6 The Refined Self-Calibration Procedure

The method described in Section 4.5 provides a means for computing all calibration parameters. However, better results are obtained if parameter estimates are refined in several steps. This is done by adjusting the parameters to minimise the sum-of-of square distances between epipolar tangencies and corresponding projected tangents using the Levenberg-Marquardt method. The geometry of the problem naturally allows for parameters to be decoupled from one another, allowing minimisation to be applied to small numbers of parameters at a time.

The first step of the procedure is to determine which silhouettes correspond to which camera views for each of the five silhouettes in the image. This is done by ordering the five silhouettes along their convex hulls, and then considering the five arrangements. The four epipoles e_{V1}, e_{V2}, e_{V121}, and e_{V212} are computed for each of the five possible arrangements. The lowest sum-of-square distances between silhouette tangents passing through the epipoles and tangents on the corresponding silhouettes is used to select the correct arrangement.

With noise, the tangent intersections used to calculate the four epipoles will, in general, produce epipoles that are not collinear. The epipoles e_{V1} and e_{V2} each lie at the intersection of four tangents. In the presence of noise, the four tangents will not intersect at a common point. For a refined estimate, the positions of the four epipoles are parameterised using only six degrees of freedom, so that the epipoles are constrained to be collinear. The sum-of-square distances from tangency points to the corresponding tangents generated by the opposite silhouette is minimised. The tangents pass through the appropriate epipole and touch the silhouette. To form a starting point (initial estimate) for the minimisation, the tangent intersections are computed, and the points closest to an orthogonal regression line through the intersection points are used.

Focal length and principal point values are then computed for each image, averaged, and adjusted to minimise reprojection error. The unknown positional component is computed next for each image. Parameters are then

adjusted by minimising reprojection error using all possible silhouette pairings between silhouettes within each set.

Finally, the five view sets are merged into a single large set as described in Section 4.5.5. A final minimisation adjusts all parameters simultaneously to minimise the sum-of-square distances across all silhouette pairings. There are $10k(5k-1)$ distance values for k input images.

4.7 Experiments

4.7.1 Qualitative Results from Real Data

Qualitative testing of the proposed self-calibration method was carried out using the two 2592×1944 images of a toy horse shown in Figure 4.2. The five silhouettes in each image were determined using an intensity threshold.

Figure 4.8 illustrates the bitangents and epipoles computed from one of the two input images. Poses and

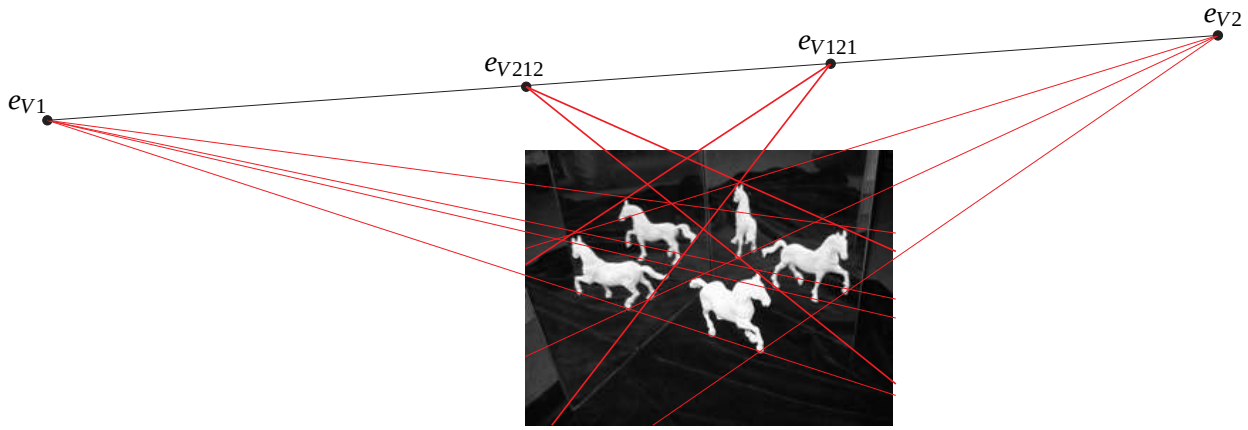


Figure 4.8: Computed bitangents and epipoles overlaid on one of the input images of a toy horse.

internal parameters were computed from the positions of the epipoles in the two input images using the methods described in this chapter. Visual hulls were computed from the silhouette to provide a qualitative assessment of the 3D shape reconstructions that one can obtain with the two-mirror setup.

The resultant visual hull model is shown in the third column of Figure 4.9. The figure also shows visual hull models created using only the five silhouettes from each of the images. This demonstrates the improvement in the quality of the model obtained by merging the silhouette sets. Note that both five-view visual hulls have regions of extra volume that are not present in the ten-view visual hull.

The angle between the mirrors was computed to be 73.1 degrees. The focal length was computed to be 2754 pixels and the principal point located at $(1306, 981)$. This compares with values of 2875 and $(1297, 958)$

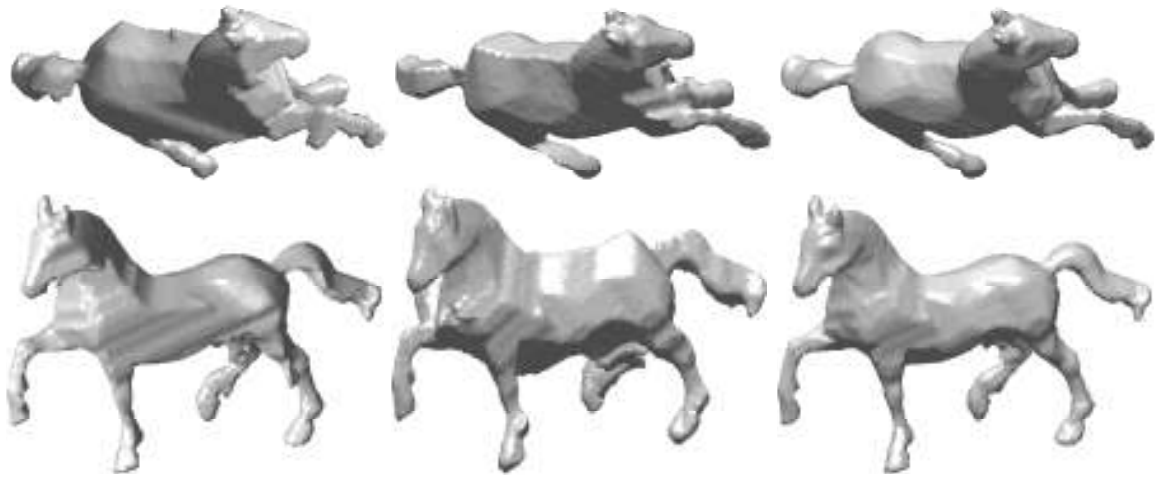


Figure 4.9: Two views of the visual hull of the horse formed from the silhouettes in image 1 (*first column*), the silhouettes in image 2 (*second column*), and all ten silhouettes (*third column*).

computed using a checkerboard calibration method (Jean-Yves Bouguet's Camera Calibration Toolbox for Matlab). Note, however, that a direct comparison of individual parameters does not necessarily provide a good indication of the accuracy of the calibration parameters. The calibration parameters should provide an accurate mapping from 2D image points to 3D rays *in the volume of interest*. The interplay between the different parameters can result in different parameter sets varying to some degree in magnitude, yet still providing a good mapping in the volume of interest. A difference in principal point location can largely be compensated for by a difference in translation parameters, for instance. A more meaningful measure of calibration parameter quality using the *silhouette calibration ratio* is described in Section 4.7.2.

Figure 4.10 provides further qualitative results, showing visual hulls of various objects computed using the proposed two-mirror setup.

4.7.2 Images Captured with a Moving Camera

Ball Images

To provide a quantitative evaluation of the viewpoint positions provided by the two-mirror setup, two images of a ball bearing were used. Since the imaged object is of known shape (it is spherical), it is possible to quantify the geometrical constraints that its silhouettes impose on its 3D shape.

Two images of a ball were captured from two viewpoints using the two-mirror setup (see Figures 4.11a and 4.11b). Self-calibration was applied to the two images using the method described in this chapter. The 3D position and diameter of the sphere were then estimated by iterative optimisation: the sum-of-square distances between the projected ball and the polygonal boundary vertices of the observed silhouettes was minimised using the Levenberg-Marquardt method. The inferred sphere parameters and calibration

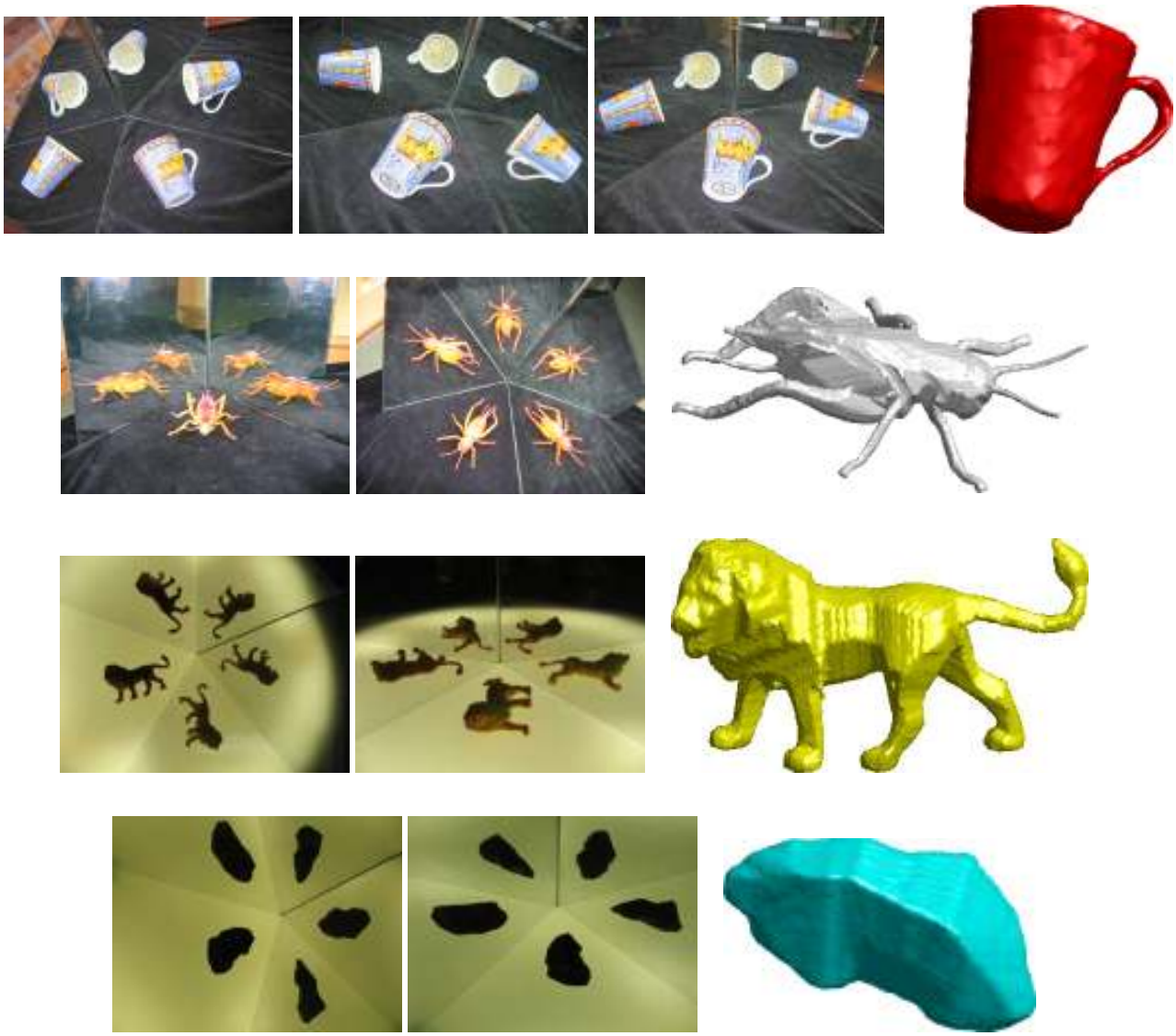


Figure 4.10: Visual hulls computed using the proposed two-mirror setup. Input images are shown in the to the left, and the resultant visual hulls are shown to the right. From top to bottom: a cup, a toy locust, a toy lion, and a piece of gravel. Black velvet was used as a background for the cup and the locust, whereas a backlight was used for the lion and the gravel.

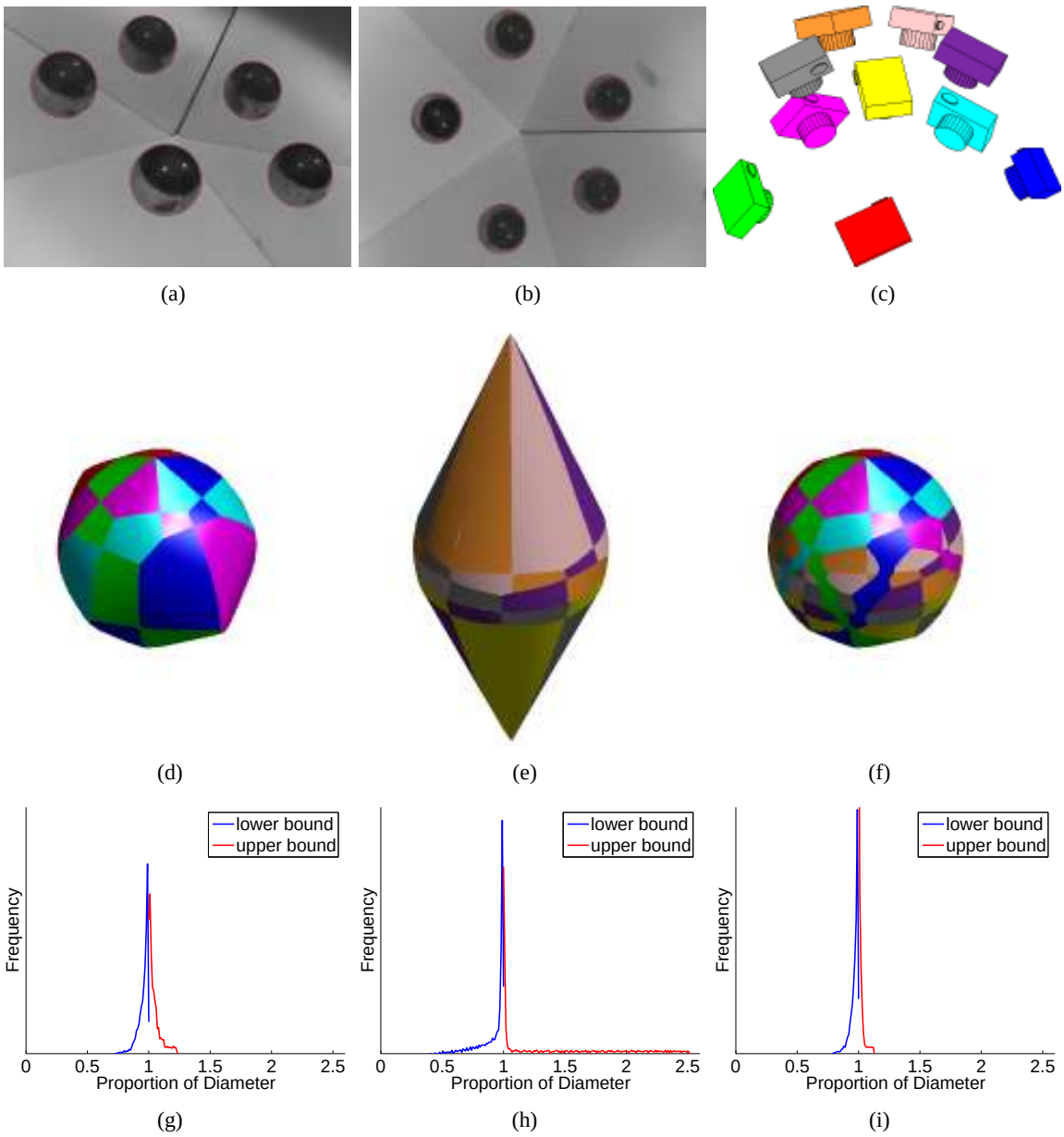


Figure 4.11: Shape inference from silhouettes of a ball bearing using the two-mirror setup: (a) first input image, (b) second input image, (c) ten viewpoints corresponding to the ten observed silhouettes, (d) synthetic 5-view visual hull corresponding to the first input image (e) synthetic 5-view visual hull corresponding to the second input image, (f) synthetic 10-view visual hull corresponding to both input images, (g)–(i) distributions of bounds of the diameter computed over all directions as a proportion of the true diameter.

parameters were then used to generate exact synthetic silhouette projections of the sphere that corresponds to the real data. This allows investigation of the inherent geometrical limitations of the extent to which 3D shape can be investigated from silhouettes using the two-mirror setup. In other words, the limitations that exist in the absence of noise can be investigated.

Results are presented in Figure 4.11. The ten silhouettes captured in the two images provide ten well-distributed viewpoints (Figure 4.11c).

The second row of the figure shows visual hulls computed from the 5-view silhouette sets from the images considered individually, and from the 10-view silhouette sets of both images considered together. The cone strip components are coloured according to the corresponding camera view. The 5-view visual hull from the first image is 105.3% of the sphere volume. The 5-view visual hull from the second image is clearly a poor approximation to the sphere, and is 149.8% of the sphere volume. Nevertheless, the 10-view visual hull is only 101.2% of the sphere volume, so both silhouette sets make significant contributions to carving away volume that is not part of the imaged object. (Since exact silhouette sets are used, the computed visual hull cannot be less than 100% of the sphere volume.)

The last row of Figure 4.11 quantifies the geometrical limitations that the three silhouette set impose on the diameter of the imaged object over all directions. (Coverage of all directions was approximated by considering directions specified by six icosahedron subdivisions.) The plots indicate that the 5-view silhouette set corresponding to the second image does not provide tight constraints on object shape. For instance, the upper bound on the diameter is 250% of its actual value in some directions. Since both the upper and lower bounds on the diameter in a given direction are closer to 100% for the 10-view silhouette set, it provides tighter constraints on the shape of the imaged object than either of the 5-view silhouette sets.

Gravel Images

Two images were captured for each of twenty pieces of gravel using the two-mirror setup. Figure 4.12 shows an example.

Although the primary purpose of capturing the data set was to generate synthetic data based on real data, the real data also allow the repeatability of the estimated calibration parameters to be quantified. Results that quantify repeatability are presented in Table 4.1. The mirrors and the camera internal parameters were

	mirror angle [degrees]	f [pixels]	u_0 [pixels]	v_0 [pixels]
mean	74.605	3862.7	653.55	467.6
standard deviation	0.017766	51.499	29.995	24.929

Table 4.1: Mean and standard deviation for parameter values computed using 20 different stones. Results are shown for the angle between the mirrors, the focal length f, and the principal point (u_0, v_0). Two images from different viewpoints were used for each stone.

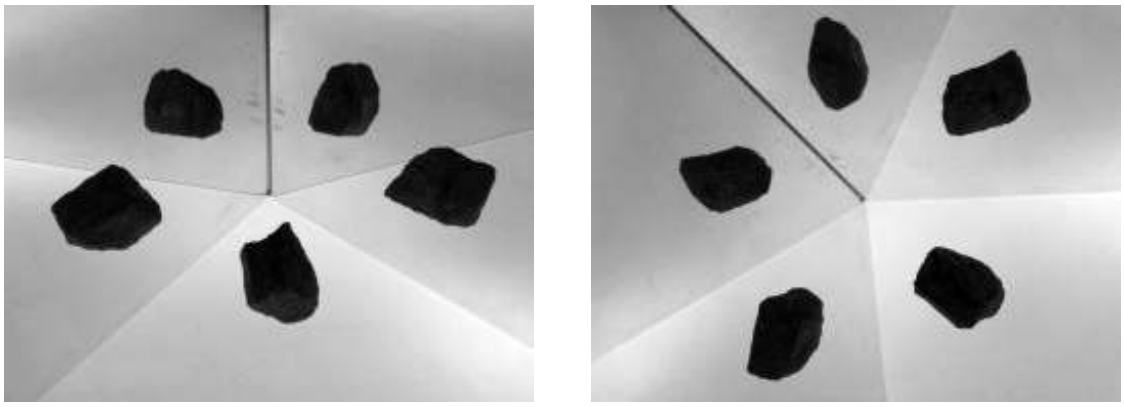


Figure 4.12: Example of two images of a piece of gravel.

fixed during the capture of 40 images of the 20 pieces of gravel, whereas the camera was held by hand and moved between shots. The estimated internal camera parameters tend to vary from set to set (for instance the coefficient of variation of the focal length is 1.3%). This occurs because small variations in these values can largely be absorbed by camera pose parameters while still maintaining an accurate image point to 3D ray mapping in the volume of interest. The angle between mirrors computed over twenty calibrations has a standard deviation of less than 1/50th of a degree.

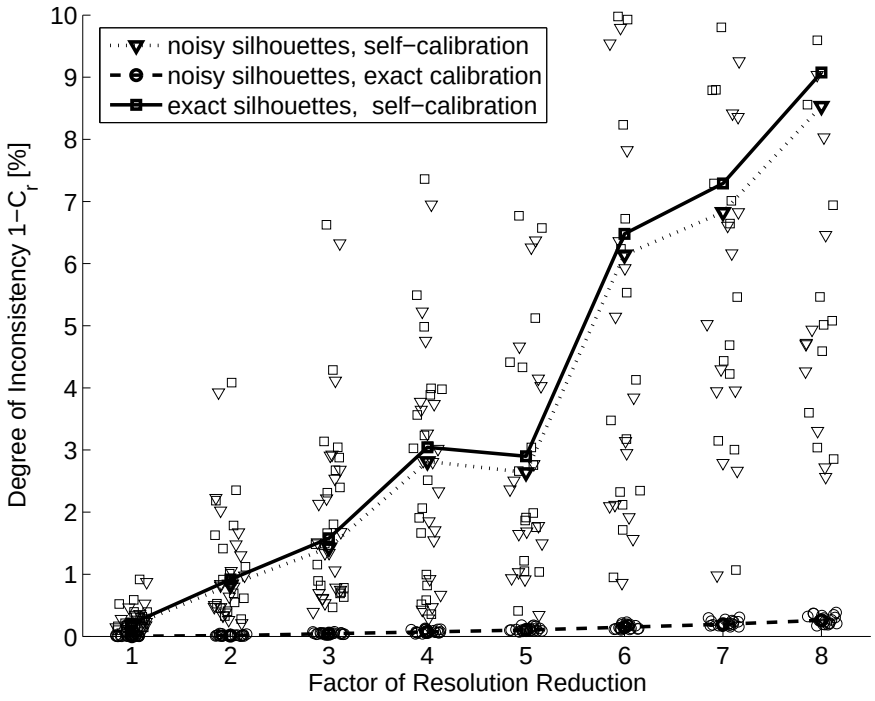
Synthetic Data

To investigate the sensitivity of the method to noise, synthetic images were used. This allows the exact values of calibration parameters to be known. To ensure that realistic parameter values were considered, the synthetic images were based on the real images of the gravel. Exact polygonal projections of the ten-view polyhedral visual hull of the gravel were generated using the output provided by the real images. This provides an exactly consistent set of silhouettes.

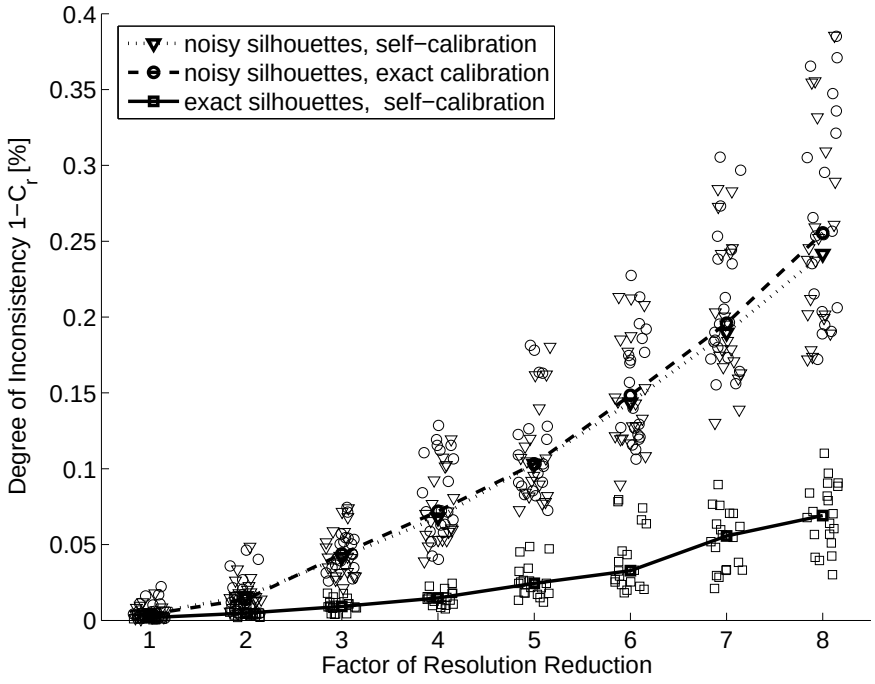
Quantisation noise was introduced by rendering the polygonal silhouettes, first at the original image resolution (2592×1944), and then at successively lower resolutions.

Boyer [14] introduced the silhouette calibration ratio C_r as a measure of the combined quality of silhouettes and camera parameters. Ideally, some point on any viewing ray in a silhouette must intersect all $n-1$ other visual cones of the n-view silhouette set. The ratio of the actual maximum number of intersections for points on the ray to $n-1$ is a measure of consistency; C_r is the mean value for all rays from all silhouettes. A measure of inconsistency is given by $1-C_r$.

Figure 4.13 shows plots of $1-C_r$ versus the degree of resolution reduction for the computed camera parameters and quantised silhouettes. Results are also shown with the computed camera parameters and exact silhouettes, as well as exact camera parameters and quantised silhouettes. The plots show that without refinement, the poor accuracy of the camera parameters is a greater contributor to inconsistency than the



(a) no refinement



(b) with refinement

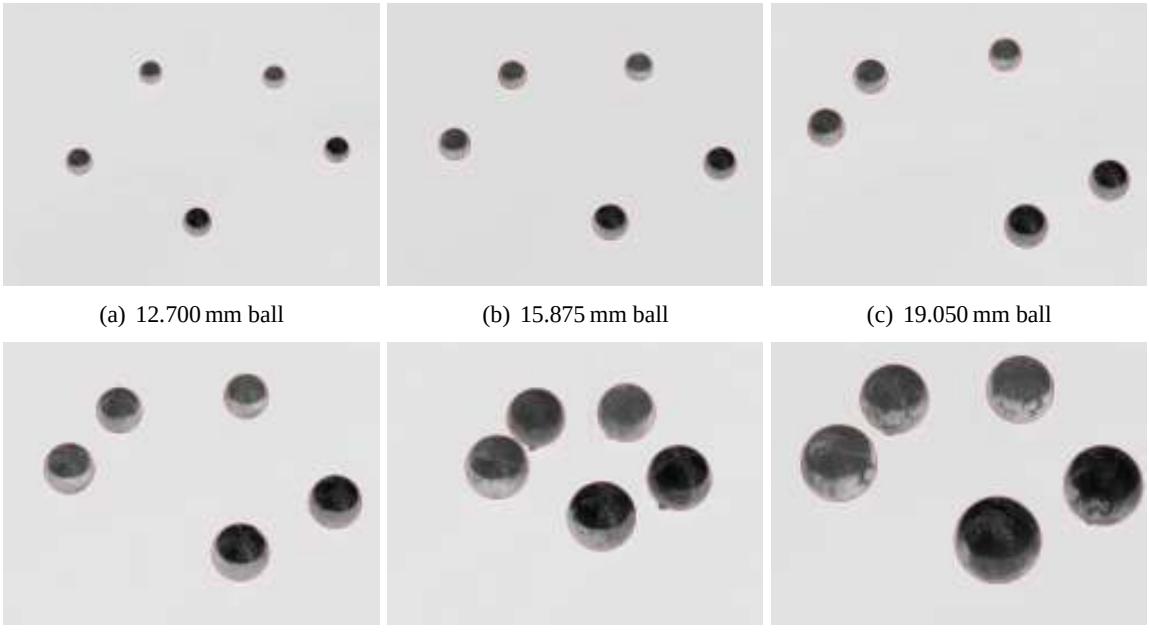
Figure 4.13: Plots of image resolution versus silhouette inconsistency measured using the silhouette calibration ratio for self-calibration (a) without, and (b) with refinement. The trend lines pass through the mean values of the data points. To aid visualisation, a small amount of jitter has been added to the horizontal components of the data points.

quantisation of the silhouettes alone. However, for the refined camera parameters, the quantised silhouettes and exact camera parameters are more inconsistent than the exact silhouettes and the computed camera parameters, demonstrating the accuracy of the refined calibration method. In other words, the quantisation of the silhouettes is a greater contributor to inconsistency than the camera parameters computed with refinement from the quantised silhouettes.

4.7.3 Images Captured with a Fixed Camera

Silhouette sets captured using a freely moving camera are calibrated up to an unknown scale factor. This means that dimensionless quantities such as the ratios used to specify particle elongation and flatness can be estimated from silhouette sets, but properties that require absolute scale such as particle volume can not. If the camera is kept in a fixed position with respect to the mirrors (using a tripod, for instance), then the relative scale for all silhouette sets will be the same. Absolute scale can be enforced by imaging an object of known size such as a ball bearing.

A data set of 220 pieces of gravel was captured using the mirror setup with the camera fixed to a tripod with a tilt angle of approximately 45°. Three images were captured of each stone, with the stones manually reoriented between shots. Polyhedral models of the stones are illustrated in Appendix C on page 220. The data set of 220 pieces of gravel is used to test shape and recognition algorithms in later chapters.



(a) 12.700 mm ball (b) 15.875 mm ball (c) 19.050 mm ball

(d) 25.400 mm ball (e) 31.750 mm ball (f) 38.100 mm ball

Figure 4.14: Six images of six ball bearings used for enforcing scale.

Scale was enforced for the gravel silhouette sets by fitting a sphere to silhouettes of an imaged ball bearing as described in Section 4.7.2. Stones were grouped in batches of 20 and calibration and scale enforcement was carried out separately for each batch.

To test the accuracy of scale enforcement, six images of ball bearings of different sizes were captured (see Figure 4.14). For each ball, the calibration and scale information estimated with another ball was used, and together with the ball's silhouettes the best fit sphere was computed. The diameter of the best fit sphere is compared with the ground truth diameter in Table 4.2. Results are shown for all pair combinations. The largest absolute percentage error for an estimated diameter is 0.281%.

	12.700 mm ball	15.875 mm ball	19.050 mm ball	25.400 mm ball	31.750 mm ball	38.100 mm ball
12.700 mm calibration		15.905 mm (+0.189%)	19.062 mm (+0.065%)	25.453 mm (+0.207%)	31.786 mm (+0.115%)	38.168 mm (+0.180%)
15.875 mm calibration	12.676 mm (-0.189%)		19.027 mm (-0.123%)	25.405 mm (+0.018%)	31.727 mm (-0.074%)	38.097 mm (-0.009%)
19.050 mm calibration	12.691 mm (-0.071%)	15.894 mm (+0.122%)		25.438 mm (+0.150%)	31.770 mm (+0.064%)	38.152 mm (+0.135%)
25.400 mm calibration	12.671 mm (-0.231%)	15.869 mm (-0.037%)	19.021 mm (-0.152%)		31.724 mm (-0.081%)	38.098 mm (-0.005%)
31.750 mm calibration	12.678 mm (-0.175%)	15.879 mm (+0.024%)	19.033 mm (-0.089%)	25.418 mm (+0.073%)		38.132 mm (+0.085%)
38.100 mm calibration	12.664 mm (-0.281%)	15.862 mm (-0.080%)	19.014 mm (-0.189%)	25.394 mm (-0.022%)	31.721 mm (-0.090%)	

Table 4.2: Ball diameters estimated from a 5-view image of a ball using calibration parameters determined by a 5-view image of a ball of another size. Estimated ball diameter and percentage error are shown.

4.8 Summary

A novel image capture setup that provides a simple means for capturing multiple silhouettes of an object from well-distributed viewpoints has been described. This chapter has demonstrated how silhouettes impose constraints that allow the pose and internal parameters associated with each view to be computed from the silhouettes alone. Since self-calibration is applied, there is no need for accurate positioning of the apparatus, and there is no need for a calibration object with control points whose coordinates must be known in advance.

Synthetic images have been used to demonstrate that the computed camera parameters have less effect on quality as measured by the silhouette calibration ratio than the noisy silhouettes from which they are computed.

The approach is limited to objects that can be segmented from the background to produce silhouettes. Objects are required to be positioned so that five non-overlapping views are visible to the camera.

The method provides the required input for multi-view silhouette-based particle analysis applications (such as recognition and shape analysis), and is also potentially a useful tool for 3D multimedia content creation.

Later chapters will quantify the performance that can be achieved for shape property estimation and matching applications using the two-mirror setup described in this chapter. This will be done using the data set of images of 220 pieces of gravel.

Chapter 5

Configuration and Calibration of a Multi-Camera Setup

5.1 Introduction

A multi-camera setup allows a much greater throughput rate than the two-mirror setup described in the previous chapter, but this comes at the cost of greater monetary expense.

The setup used in this thesis consists of six simultaneously triggered cameras. Particles are placed on a feeder above the cameras. The feeder causes the particles to fall past the cameras one by one at a rate of approximately ten particles per second. As each particle falls, it passes through a light curtain that triggers the cameras so that a 6-view image set of the particle is captured. The multi-camera setup used in this work was built by Anthon Voigt and his team at the premises of the company that commissioned part of the work described in thesis. The hardware aspects of the multi-camera setup lie outside the scope of this thesis.

In this chapter, the rationale behind the positioning of the cameras is discussed. A simple method for calibrating the cameras using images of a ball of known size is then presented.

5.2 Positioning Multiple Cameras

The multi-camera setup serves several purposes: matching stones, estimating various size and shape properties, and building 3D visual hull models of stones for visualisation purposes.

Accuracy can be improved for a given application by increasing the number of cameras in a multi-camera system. However, for a given number of cameras, it is not obvious how the cameras should be positioned

so as to obtain the best accuracy. The solution to the problem is somewhat dependent on the measure of accuracy, the specific application, and the sizes and shapes of the particles.

The multi-camera setup was built with six cameras positioned so that each camera looks onto one of the six parallel face pairs of a regular dodecahedron. Figure 5.1 illustrates the setup. Each camera is approximately

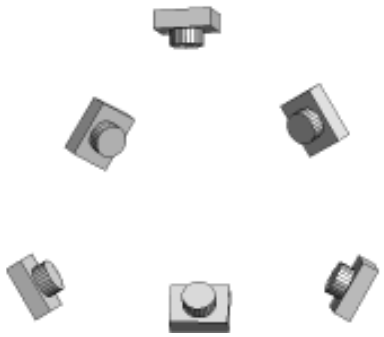


Figure 5.1: The configuration of the six-camera setup used in this thesis.

500 mm from the centre of the dodecahedron. Since particles tend to be imaged close to the centre of the dodecahedron, and are approximately 5 mm in diameter, the setup provides weak perspective imaging conditions: particles are close to the optical axes of all cameras, and particle depth variation is small with respect to the distances to the cameras.

The number of cameras was limited by monetary cost and hardware limitations. Six was also considered to be a more favourable number of cameras than five or seven, since a symmetrical configuration could be realised. The remainder of this section gives some justification to the choice of camera configuration.

5.2.1 Undesirability of Coplanar Cameras

Although the problem of determining the best next view for shape-from-silhouette modelling has been considered before [12, 114], the problem of optimally positioning a number of fixed cameras for shape-from-silhouette applications has received little attention in the computer vision literature.

Mundermann et al. [98] address the problem in the context of building visual hull models of humans. They find that cameras positioned in a geodesic dome configuration (i.e., well-distributed over a hemisphere), and cameras positioned in a circular coplanar configuration around the object produce the best results.

Other than Mundermann et al.'s findings that a circular coplanar configuration is desirable for at least certain applications, a coplanar configuration is worth considering since it simplifies the manufacture and assembly of the structure that houses the cameras.

Figure 5.2 illustrates why a circular configuration (cameras with coplanar optical axes) is undesirable for estimating 3D shapes of certain nonconvex particles from silhouette sets. In the first row of the figure, the

(a) particle and cameras (b) visual cones (c) visual hull

(d) particle and cameras (e) visual cones (f) visual hull

Figure 5.2: Six cameras observing a banana-shaped particle: (a)–(c) show cameras with coplanar optical axes, (d)–(f) show cameras based on a Platonic solid geometry. Note that the coplanar cameras yield a visual hull model that is much larger than the particle: the extra volume is due to the saddle-shaped region of the particle. The cameras based on a Platonic solid geometry yield a visual hull model that is a relatively close approximation to the particle.

cameras are positioned so that their optical axes are coplanar, with an even angular distribution about 180°. The visual hull model (Figure 5.2c) is a poor approximation to the banana-shaped particle (Figure 5.2a), since there is additional volume in the saddle-shaped region. This camera configuration would perform poorly at visual hull-based volume estimation, since for nonconvex particles the volume estimate would be highly dependent on the orientation of the particle with respect to the cameras. The circular camera configuration is desirable for building visual hull models of humans (the application of Mundermann et al.), since humans are not arbitrarily oriented with respect to the cameras.

Section 7.5.5 provides some further results that demonstrate the undesirability of coplanar cameras in the context of matching pairs of silhouette sets: mismatch pairs cannot be distinguished from match pairs for a range of particle orientations.

5.2.2 Positioning Cameras by Optimising Objective Functions

The camera configuration for the work presented in this thesis was determined by optimising an objective function. Two approaches were considered: (1) maximising the sum of distances between frontier points on a sphere, and (2) minimising the angle between the most isolated direction and its closest viewing direction. In

other words, two different criteria were considered for positioning the cameras. The *frontier point criterion* specifies that the cameras should be positioned so that the sum of distances between frontier points on a sphere is maximal. The *direction isolation criterion* specifies that the direction that is furthest (in terms of angle) from any of the viewing directions is minimal.

Representation for Cameras

The pose of a camera has six degrees of freedom. However, the orientation of the camera does not affect the information contained in a silhouette: rotating the camera about its centre does not alter the rays that pass through the centre. Furthermore, since the viewed particles are small with respect to the variation in particle position and the working distance of the cameras, it may be assumed that the distance from the cameras to the particles is many times greater than the size of the particles. This means that the positional component of each camera along its optical axis is almost inconsequential from an informational point of view. For these reasons, the positioning of each additional camera with respect to a fixed first camera can be considered to introduce only two additional degrees of freedom. For simplicity, cameras are considered to be directions specified by points on a sphere and an orthographic imaging model is used.

Camera positions are over-parameterised by using three coordinates to specify the position of each of the cameras on the viewing sphere (two degrees of freedom). The over-parameterisation prevents the occurrence of singularities and allows for a smooth function to aid the optimisation process.

The Frontier Point Criterion

The objective function to be maximised for the frontier point criterion is the sum of distances between frontier points on a sphere viewed by orthographic cameras. Each possible pairing of two cameras yields two frontier points on the viewed sphere: if n cameras are used, then there are $n(n-1)$ frontier points. A sphere is used instead of any other shape for reasons of symmetry and simplicity. In practice, particles being viewed by the cameras can be assumed to be arbitrarily oriented: a sphere does not introduce any directional bias. Maximisation of the objective function ensures that frontier points are well-distributed over the surface of the viewed object.

Since frontier points are well-distributed on a sphere, they are also well-distributed on the particle. (The frontier points that lie on the saddle-shaped region project to epipolar tangencies that are not *outer* epipolar tangencies, since the epipolar tangencies lie on concave boundaries of the silhouettes.) Frontier points lie both on the particle and on the visual hull, so regions close to the frontier points are accurately modelled by the visual hull (provided that there aren't sudden changes in the local surface geometry). A camera configuration that causes frontier points to be well-distributed over the object is therefore likely to provide a visual hull model that accurately approximates the particle over all regions of the particle's surface. This reduces the likelihood of certain regions being poorly modelled and ensures reasonable performance for

applications such as volume estimation and shape analysis in which the visual hull is used as an estimate of the shape of the particle.

The Direction Isolation Criterion

An alternative approach is to minimise the most isolated unused viewing direction. By limiting the maximum difference in direction between unobserved views of an object and the observed views, the probability of not observing a saddle-shaped region of the object's surface, for instance, is reduced.

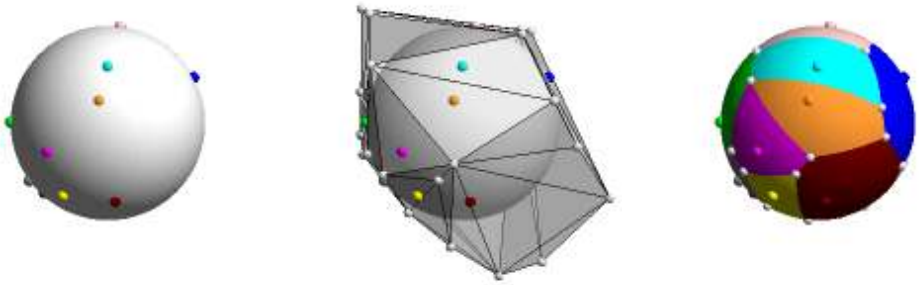
The most isolated direction is determined using a spherical Voronoi diagram. The Voronoi diagram is a division of the surface of a sphere into cells, based on the positions of a set of *site* points on the sphere. Each cell surrounds exactly one site point so that each point within the cell is closer to the site point contained in the cell than to any other site point. Each viewing direction is specified by a pair of antipodal points on the viewing sphere: these are the site points.

To determine the most isolated point from a set of site points on a sphere of any dimension, only the vertices of the Voronoi diagram need be considered, since for any non-vertex point there will be a Voronoi vertex point that is more isolated. The most isolated camera direction is therefore computed, for a given set of viewing directions, by finding the Voronoi vertex whose closest site point is further than for any other Voronoi vertex.

Spherical Voronoi diagrams can be easily computed for spheres in any dimension. The procedure is illustrated in Figure 5.3. Tangent planes at site points must be considered. The intersections of the halfspaces specified by the tangent planes is a convex polyhedron. The halfspace intersection can be formed by computing the convex hull in dual space, i.e., by treating the homogeneous representation of the tangent planes as points. The dual of the dual space polyhedron is the required polyhedron (Figure 5.3b). The Voronoi diagram is formed by projecting the polyhedron vertices onto the sphere (Figure 5.3c). The connectivity of the diagram is given by the connectivity of the polyhedron. This method of computing Voronoi diagrams using convex hulls in a higher dimension was introduced by Brown [19].

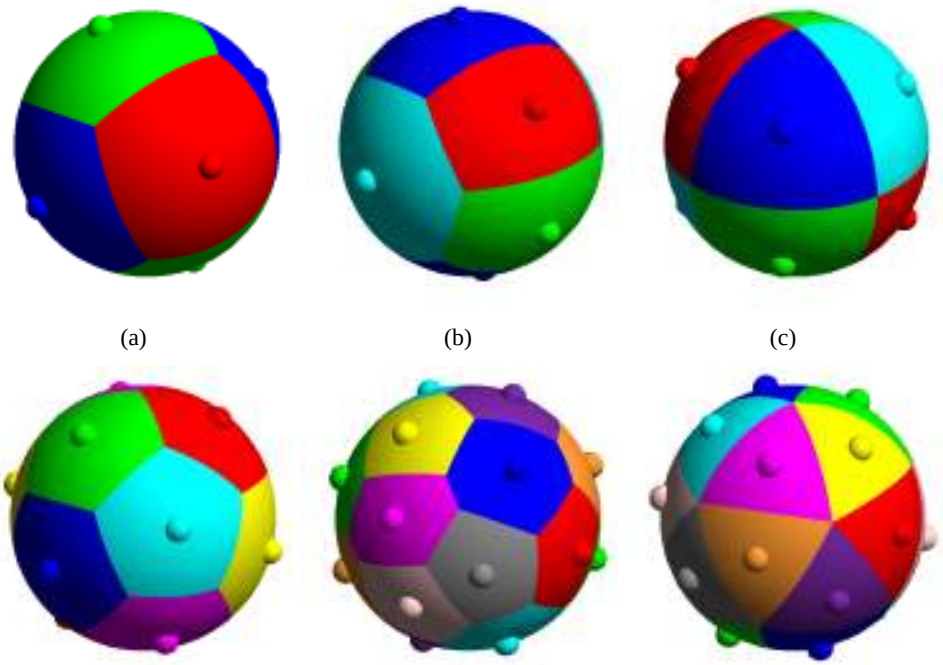
5.2.3 Configuration Optimisation Results

Both objective functions were optimised using Matlab's Nelder-Mead simplex method [28]. Starting points for camera positions on a sphere were chosen by randomly selecting points from a subdivided icosahedron. Four subdivisions of the icosahedron were performed to obtain 812 points that are well-distributed on a sphere. The objective function was evaluated for 1000 different randomly selected point sets, and the best of these point sets was used as a starting point for an optimisation. This procedure was repeated 1000 times, and the best result was selected. Multiple applications of this approach produced the same sets of relative camera positions. Figure 5.4 illustrates camera configurations optimised with the direction isolation criterion and with the frontier point criterion. Antipodal pairs of points on the unit sphere indicate camera directions,



(a) (b) (c)

Figure 5.3: Procedure for computing the Voronoi diagram on a sphere: (a) a sphere with some site points in colour, (b) a polyhedron formed by intersecting all halfspaces defined by tangent planes to the site points, (c) the Voronoi diagram formed by projecting the polyhedron vertices onto the sphere; the connectivity of the diagram is given by the connectivity of the polyhedron. Surface regions on the Voronoi diagram are coloured according to the nearest site point.



(a) (b) (c)

(d) (e) (f)

Figure 5.4: Camera configurations with camera directions represented by spheres of the same colour: (a) optimal 3-camera directions for both direction isolation and frontier point criteria, (b) optimal 4-camera directions for direction isolation criterion, (c) optimal 4-camera directions for the frontier point criterion, (d) optimal 6-camera directions for both direction isolation and frontier point criteria, (e) optimal 10-camera directions for direction isolation criterion, (f) optimal 10-camera directions for frontier point criterion.

and regions on the sphere are coloured to correspond to the closest camera direction. For certain numbers of cameras, the two camera positioning criteria produce different configurations, whereas for other numbers of cameras, one configuration is optimal for both criteria.

Notably, the direction isolation and frontier point criteria both produce the same configuration for six cameras (see Figure 5.4d). This configuration was therefore chosen for the six-camera setup used in this thesis.

In the case of the frontier point criterion, 3-, 4-, 6-, and 10-camera setups correspond to the directions specified by the face normals of the Platonic solids. (The regular tetrahedron and the regular octahedron both correspond to the same 4-camera setup.) In the case of the direction isolation criterion, the 3- and 6-camera setups correspond to Platonic solids, whereas the 4- and 10-camera setups do not.

Camera configurations optimised using the frontier point criterion are illustrated in Figure 5.5. The corresponding frontier points on a sphere are shown in Figure 5.6.

Although only the six-camera setup was physically realised, the camera configurations consisting of different numbers of cameras are used in this thesis for several experiments using synthetically generated data. This enables investigation of the performance of various algorithms with different camera configurations.

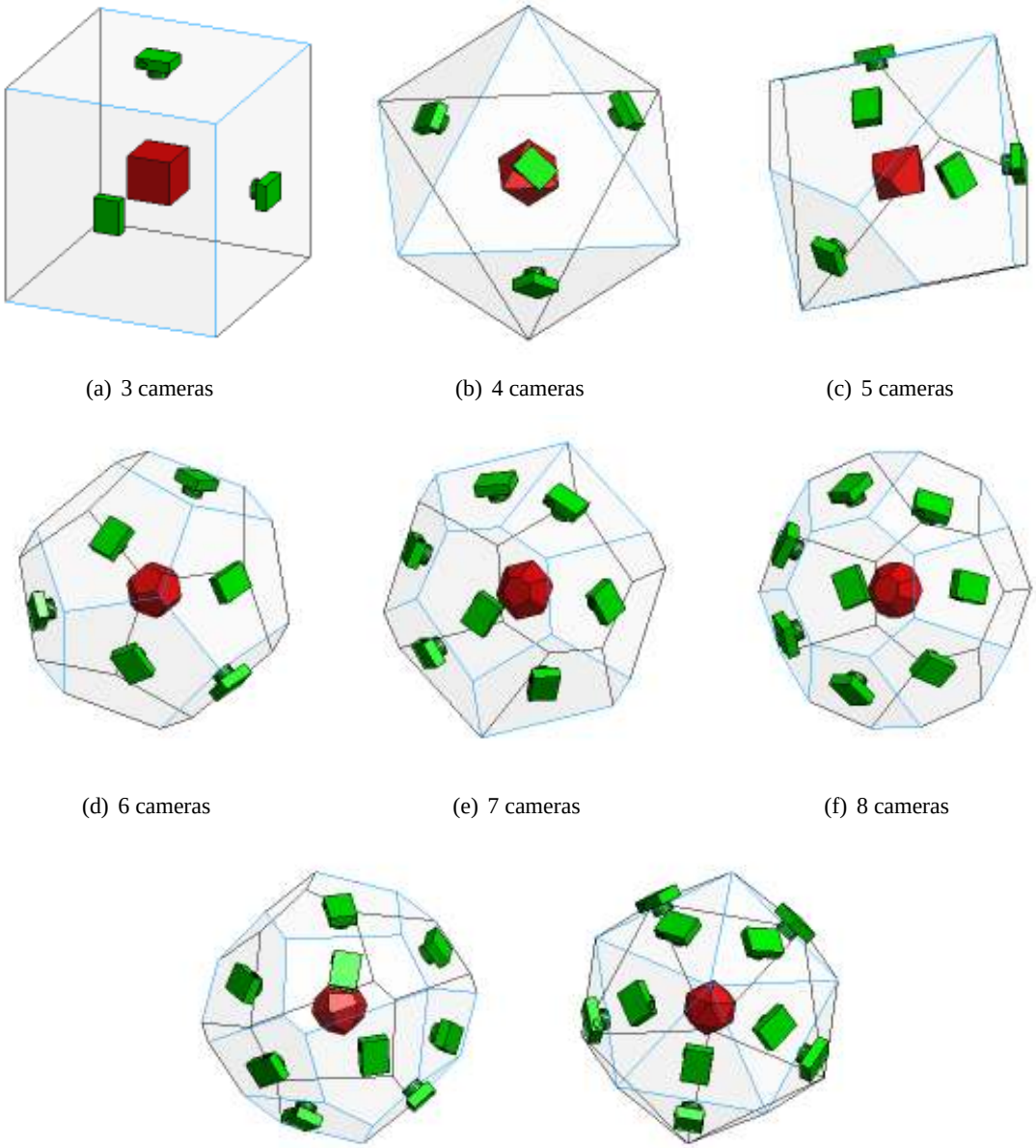
The best configuration of the two-mirror setup was also determined by the optimisation using the direction isolation criterion. It is a symmetrical setup with 72° between the mirrors and the camera tilted at 42.0°. This produces a most isolated direction that is 48.0° from the closest viewing direction. This is only 2.1° larger than the optimal most isolated direction that can be achieved from any five viewing directions.

5.3 Camera Calibration

Multi-view, silhouette-based particle analysis applications such as particle size and shape analysis, and individual particle recognition require accurate camera calibration. The internal and pose parameters of each camera in a multi-camera setup must be estimated so that the 3D ray corresponding to any 2D image location is known in a common reference frame.

In earlier work [45], a calibration method was developed using a calibration object with coded marker patterns. Figure 5.7 shows two examples of the calibration objects with coded marker patterns. The circular markers are identified by their code bands, and the camera internal and pose parameters are inferred from the positions of the imaged markers across multiple images.

Here, a different approach to calibration is described. A sphere (typically a ball bearing) is passed through the multi-camera setup several times, and several image sets are captured. Pose and internal parameters are then inferred from the images of the ball. This approach of using ball bearings to calibrate the multi-camera setup has several advantages over using a calibration object with coded targets:

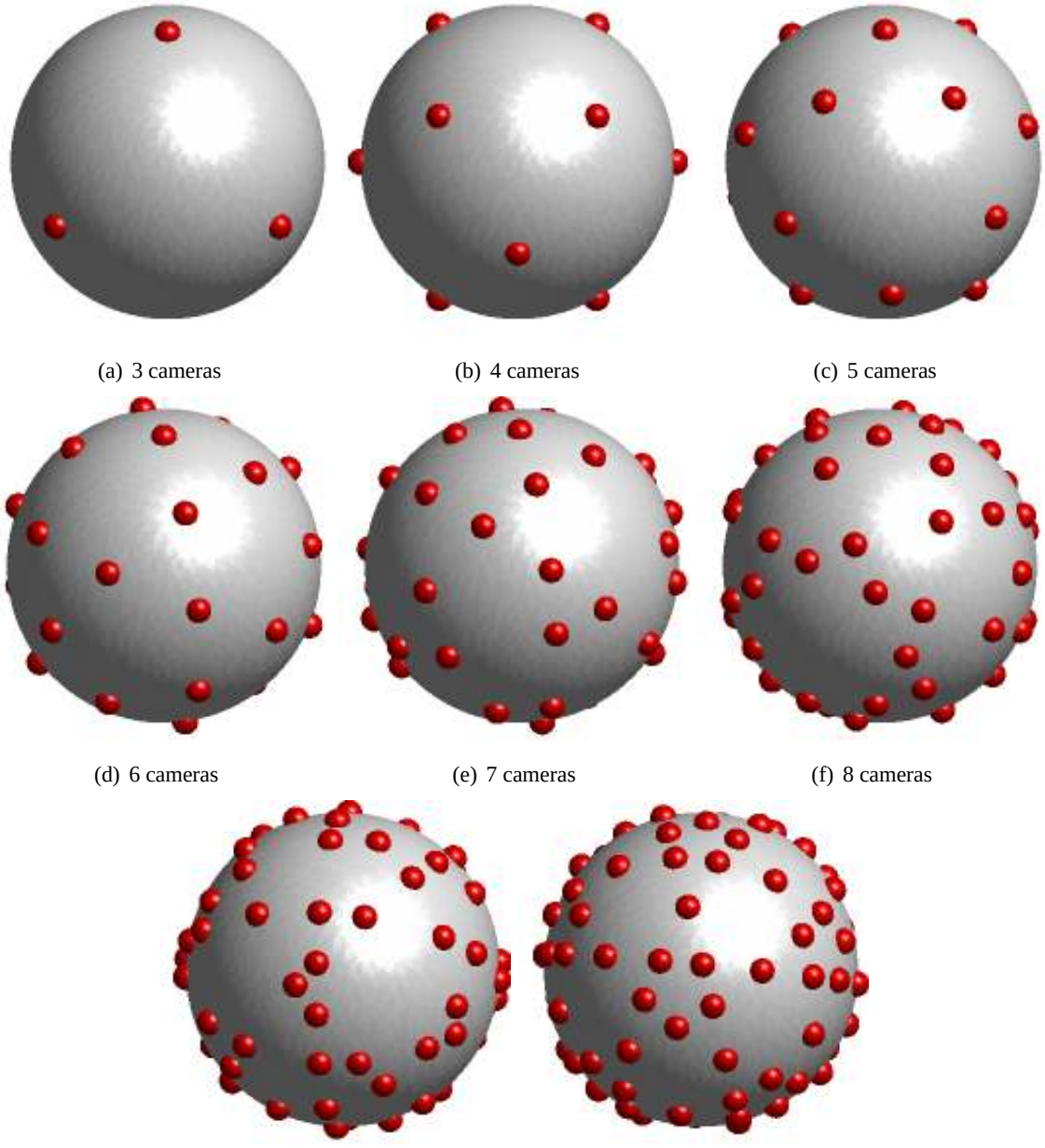


(a) 3 cameras (b) 4 cameras (c) 5 cameras

(d) 6 cameras (e) 7 cameras (f) 8 cameras

(g) 9 cameras (h) 10 cameras

Figure 5.5: Camera setups optimised with the frontier point criterion. The n cameras (green) are shown together with the $2n$-faced polyhedra representing each camera configuration. The polyhedra are shown as a casing on which the cameras are mounted and as a positioning aid at the centre of the casing, with the cameras looking onto the parallel face pairs. The setup in (a) is based on the geometry of a cube; this is the configuration used by the University of Illinois Aggregate Image Analyser [108]. The six-camera setup used in this thesis is configured as in (d). Note that (a), (b), (d), and (h) show Platonic solids.



(a) 3 cameras (b) 4 cameras (c) 5 cameras

(d) 6 cameras (e) 7 cameras (f) 8 cameras

(g) 9 cameras (h) 10 cameras

Figure 5.6: Positions of frontier points on a sphere for camera setups optimised with the frontier point criterion. The images correspond to the images shown in Figure 5.5.

(a) (b)

Figure 5.7: Two examples of calibration objects with coded targets: (a) a cube with 54 targets and 9-bit code bands, (b) an icosahedron with 60 targets and 10-bit code bands.

1. Unlike calibration using coded marker patterns, calibration using balls makes use of silhouette images. This means that there is no need for front lights to illuminate object surfaces. Ball calibration therefore has the potential to reduce the complexity of a multi-camera setup by removing the need for two sets of lights; only the back lights that are already required for creating silhouette images of particles are needed.

2. By using objects that fall off the feeder (balls) instead of moving a calibration object in front of the cameras, the appropriate 3D region is calibrated. Since the calibration parameters are to be used with objects that fall off the feeder, appropriate coverage is achieved.

3. Since the shape of the balls is known in advance, the silhouettes boundaries can be robustly detected from within images: the image of a sphere is a conic section, and can be closely approximated by a circle in many practical imaging configurations.

4. Unlike calibration objects with coded marker patterns, ball bearings of many sizes are inexpensive and readily available.

5. Balls can be used to calibrate common fields of view that are too small for coded marker patterns to be used. It is impractical to create a calibration object with coded marker patterns that is much smaller than an inch in diameter. However, small ball bearings can be used with relative ease.

Camera calibration is carried out by adjusting all camera parameters simultaneously to minimise the ET error across all observed silhouette sets using the Levenberg-Marquardt method. Although this approach can be carried out using silhouette sets of stones rather than silhouette sets of a ball, using a ball instead of stones provides two advantages:

1. The ball images provide an effective means for computing initial parameter estimates. Without good initial parameter estimates, Levenberg-Marquardt optimisation may converge to a local minimum that is far from the global minimum.

2. Minimisation of ET error across all silhouette sets does not determine absolute scale. A ball of known size provides a convenient means for enforcing absolute scale.

5.3.1 Related Work

Early works on camera calibration (within the field of computer vision), such as Tsai's method [130], rely on control points with accurately known 3D coordinates. In the 1990s, self-calibration methods were developed for computing camera parameters from correspondences of points with unknown 3D coordinates. One of the original self-calibration methods was developed by Tomasi and Kanade [129] for orthographic cameras. Although the method has been extended in various ways to a perspective camera model [56, 115], the method described in this chapter uses the Tomasi-Kanade method to establish initial camera parameters. This is because the perspective modelling methods are unstable if the degree of perspective distortion in a scene is small. By using images of a ball, it is easy to closely approximate multiple point correspondences that would be observed by orthographic cameras with the same viewing directions as the actual cameras.

Practical methods for calibrating multi-camera setups based on self-calibration point correspondences have been described in the computer vision literature. For instance, Svoboda et al. [122] calibrate a multi-camera smart room. Their system consists of four cameras that share a large common field of view. Point correspondences across multiple views are obtained by having a person move a laser pointer around the common field of view.

Following the analysis of the generalisation of the epipolar constraint to include silhouettes [3], there has been interest in calibrating multi-camera setups using silhouettes. Sinha and Pollefeys [120] make use of outer epipolar tangents to calibrate a network of cameras using silhouettes. Random sampling is used to identify consistent corresponding epipolar tangencies to use for computing initial parameter estimates. Since the six-camera setup considered in this chapter is a highly controlled environment, it is not necessary to resort to random sampling to estimate initial parameters, since multiple point correspondences can easily be generated using a ball.

The computer vision literature describes several approaches to calibrating cameras using spheres. Shivaram and Seetharaman [118] point out that the major axis of an elliptical projection of a sphere always passes through the principal point. Using this observation, they derive equations for camera poses and internal parameters, and test their method with synthetic images.

Xu et al. [128, 140] show how internal and pose parameters can be estimated separately using linear methods. The solution is then globally refined using the Levenberg-Marquardt method.

Agrawal and Davis [1] describe a method for multi-camera calibration using spheres. They use a dual space approach and solve the camera parameters using semi-definite programming (an extension of linear programming where positive semi-definiteness constraints are used on matrix variables). The method appears to solve the same problem addressed in this chapter.

These approaches provide alternatives to the approach that was used, which was chosen for its relative simplicity. A possible problem with the above approaches is that the perspective distortion in individual images is low: imaged balls appear as circles. This makes it difficult to resolve the relationship between depth and focal length from individual images. The method that was used is able to resolve these factors by considering multiple ball images in which the ball position varies somewhat. To a good approximation, the ball projections appear as circles of varying size, allowing depth and focal length to be estimated.

5.3.2 Preprocessing

The calibration routine requires the same ball to be passed through the six-camera setup several times. Usually approximately 20 image sets are captured. A background image is also captured for each camera.

The first step of the calibration procedure is to compute threshold values to use for threshold-based segmentation. This is done using Otsu's method which minimises the intra-class variance of pixel intensity values [105]. Polygonal ball boundaries are extracted from each image using the same threshold-based segmentation routine that is to be used for subsequently extracting stone silhouette boundaries. The routine is described in Appendix A.

A circle is fitted to each ball boundary. First, a linear least squares method is used to form an initial solution. This solution is then refined by minimising the sum-of-squared distances from the polygon vertices to the circle. The fitted circles are used for determining initial parameter estimates; the original polygonal boundaries are used for refining the solution.

5.3.3 Initial Parameter Estimate

The initial pose estimates are computed using the Tomasi-Kanade factorisation method. The method determines 3D point locations and camera poses from orthographic projections. To estimate the orthographic projections of the ball centres from the same viewing directions as the cameras, the radii of circles representing the imaged ball boundaries are used. By scaling the circles with the image centres as the origins (i.e., assuming that principal points are at image centres) the scaled circle centres provide a close approximation to the orthographic projection that would be obtained from the viewing direction. The Tomasi-Kanade method provides the 3D positions of the ball centres and the camera poses (although camera depths are not given, since an orthographic projection is unchanged by a change in depth). However, there are always two consistent solutions. To resolve the ambiguity, the circle diameters are again used. The solution that results

in the circle diameter decreasing with ball depth is chosen. Orthogonal regression lines are fitted to the ball depth and circle diameter values to compute the camera depth and focal length values.

Computing Approximate Orthographic Projection Coordinates

A good approximation of the orthographic projection of the ball's centre is obtained from the camera's projection of the ball. Since the distance from the ball to the camera is large with respect to the ball diameter, and since wide angle lenses are not used, the ball boundary is a close approximation to a circle. The coordinates of the orthographic projection of the ball centre (x_c, y_c) are estimated from the extracted circle centre coordinates (u_c, v_c) as follows:

$$x_c = \frac{u_c - u_0}{r_i} \tag{5.1}$$

$$y_c = \frac{v_c - v_0}{r_i} \tag{5.2}$$

where (u_0, v_0) is the estimate of the principal point (the image centre is used) and r_i is the radius of the extracted circle in pixels. These equations produce coordinates that are in units of the ball radius.

Tomasi-Kanade Factorisation

This section briefly describes the Tomasi-Kanade factorisation method. Further details are given by Tomasi and Kanade [129].

The first step is to move the origin to the centroid of the projected points. This removes the translational component of the pose, since the projection of the 3D centroid of the 3D points is the 2D centroid of the 2D point projections.

Next, a *measurement matrix* $\hat{W}$ is formed from the translated coordinates:

$$\hat{W} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \\ y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix} \tag{5.3}$$

The $2n$ rows of $\hat{W}$ correspond to the n cameras, and the m columns correspond to the m image sets of different 3D ball positions.

Singular value decomposition is applied to $\hat{W}$ to give

$$U\Sigma V^T = \hat{W}. \tag{5.4}$$

The first three columns of U are used to form a *motion matrix* $\hat{M}$. A *shape matrix*

$$\hat{S} = \Sigma_3 V_3^T \tag{5.5}$$

is formed from Σ_3, the first three rows and columns of Σ, and V_3, the first three columns of V. This results in the factorisation

$$\hat{W} = \hat{M}\hat{S}. \tag{5.6}$$

The shape matrix and the motion matrix represent the 3D structure and camera poses up to an arbitrary affine transformation. In other words, any arbitrary affine transform of the 3D structure yields a consistent solution.

The true motion matrix M has rows that are unit vectors, and the corresponding rows in the upper and lower halves of the matrix are orthogonal. To enforce these constraints, a matrix A is sought such that

$$M = \hat{M}A \tag{5.7}$$

$$S = A^{-1}\hat{S}, \tag{5.8}$$

and A enforces the metric constraints

$$\mathbf{i}_r^T AA^T \mathbf{i}_r = 1 \tag{5.9}$$

$$\mathbf{j}_r^T AA^T \mathbf{j}_r = 1 \tag{5.10}$$

$$\mathbf{i}_r^T AA^T \mathbf{j}_r = 0, \tag{5.11}$$

where $\mathbf{i}_r^T$ is the rth row of M and $\mathbf{j}_r^T$ is the $(r+n)$th row of M. These constraints are imposed using linear least squares to determine Q, where

$$Q = AA^T. \tag{5.12}$$

Once Q is determined, Cholesky factorisation is used to determine A. (Tomasi and Kanade use nonlinear optimisation to determine A directly; the approach of using Cholesky decomposition is described by Weinshall and Tomasi [134].) If the matrix Q is not positive definite, then Cholesky decomposition cannot be applied. This will occur if the system becomes completely overwhelmed by noise.

The rotation matrices associated with each camera have rows $\mathbf{i}_r^T$, $\mathbf{j}_r^T$ and $\mathbf{k}_r^T$ where

$$\mathbf{k}_r^T = \mathbf{i}_r^T \times \mathbf{j}_r^T. \tag{5.13}$$

In the presence of noise, these matrices will not in general be orthonormal. The singular value decomposition is used to enforce orthogonality: the diagonal matrix in the decomposition is replaced by the identity matrix.

Resolving the Reflection Ambiguity

There is an inherent ambiguity in the solution to the camera poses and 3D point positions: two solutions are consistent with the observed orthographic projections. The two solutions correspond to A and $-A$ both providing consistent solutions. Figure 5.8 shows an example of two scenes in which both sets of cameras capture the same orthographic projections. The ambiguity arises because a positive rotation of a point in front of a centre of rotation cannot be distinguished from a negative rotation of a point behind the centre of rotation [18].

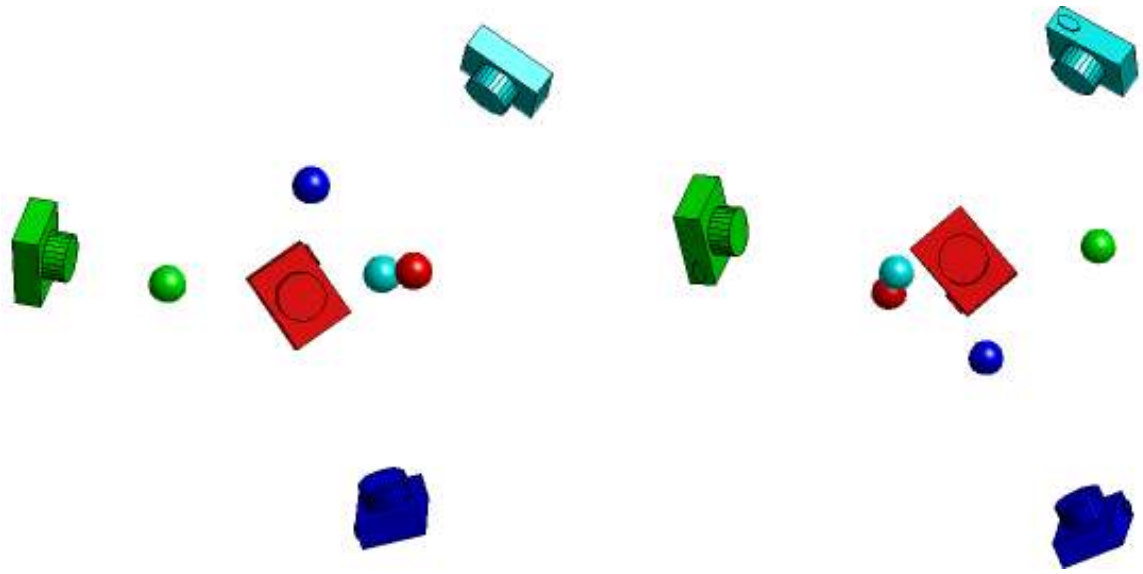


Figure 5.8: Two consistent setups for a set of observed orthographic point projections. Note that the camera icons represent viewing directions; the position of the camera parallel to the viewing direction is inconsequential.

To resolve the ambiguity, each of the two possible solutions is considered in turn. For each camera, the imaged circle radius should be inversely proportional to the associated depth, since for a weak perspective projection

$$r_i = \frac{f r_w}{z}, \tag{5.14}$$

where f is the focal length, r_w is the ball radius, and z is the depth. World coordinates are measured in terms of r_w, therefore $r_w = 1$. Camera depths are unknown at this stage and are set to zero.

The correlation coefficient of the radius inverses and the depths are computed for each camera. The solution that produces the largest positive correlation coefficient is selected. (In the noise-free case, the true solution will produce correlation coefficients of $+1$ and the incorrect solution will produce correlation coefficients of -1.)

Estimating Focal Length and Depth Values

The depth of the cameras and the focal lengths are computed by fitting an orthogonal regression line to the radii inverses and depth values (with all cameras initially positioned at $z = 0$).

The slope of the regression line gives the focal length and the negative of the intercept is the camera depth.

5.3.4 Parameter Refinement

The initial parameter estimate is refined by using the Levenberg-Marquardt method to adjust all calibration parameters to minimise the sum of residual ET errors across all silhouette sets.

Since each camera pair generates two outer frontier points, each of which is imaged by each camera, each camera pair generates four residual ET error values. The 15 camera pairings from six cameras therefore generate 60 residual ET error values per image sets; there are $60k$ residual values for k image sets. The calibration parameters for each camera consist of 10 parameters per camera: three for the internal parameters (f, u_0, and v_0), and seven pose parameters (a quaternion to represent orientation, and a three element vector to represent position). (The four element quaternion overparameterises the orientation which has only three degrees of freedom.) In total, 60 calibration parameters are therefore adjusted to minimise the sum of square residual error over $60k$ residual values. Note that further parameters that model, for instance, radial or tangential lens distortion could be added at this stage (with initial values of zero). However, the lenses used did not exhibit significant distortion, and initial experimentation showed no benefit in including lens distortion terms.

Since six cameras are used and pixels are modelled as squares, there are sufficient constraints to calibrate up to only a single unknown scale factor [58]. (Fewer cameras or unknown pixel skew and aspect ratios can lead to cases in which calibration can only be carried out to a projective transform.)

Scale is enforced subsequent to the Levenberg-Marquardt minimisation using the prior knowledge of the ball diameter. Linear Euclidean triangulation [58] is used to determine the 3D position of the ball centre from the circle centres of the images in each set. The ball diameter d_{world} implied by the model is then estimated from each image using

$$d_{\text{world}} = \frac{z}{f} d_{\text{image}}, \tag{5.15}$$

where d_{image} is the diameter of the circle in the image, z is the z-coordinate of the ball position in the camera's reference frame (i.e., the depth) and f is the camera focal length in pixels. This is a weak perspective approximation that assumes that the rim (i.e., the contour generator that projects to the ball boundary in the image) is at the same depth as the ball centre. This is a good approximation since the ball diameter is small with respect to the distance to the camera centre. Camera positions are scaled to enforce absolute scale so that the mean computed ball diameter is equal to the known value.

5.3.5 Experiments

The calibration method was tested using 20 image sets of a 5.54 mm ball bearing. Figure 5.9 shows a six image set of a garnet, with projected epipolar tangents derived using the computed calibration parameters. The accuracy of the computed calibration parameters affects how close the projected tangents are to being tangential to the silhouettes: in the noise-free case the epipolar tangency constraint specifies that the projected tangents are tangent to the silhouettes.

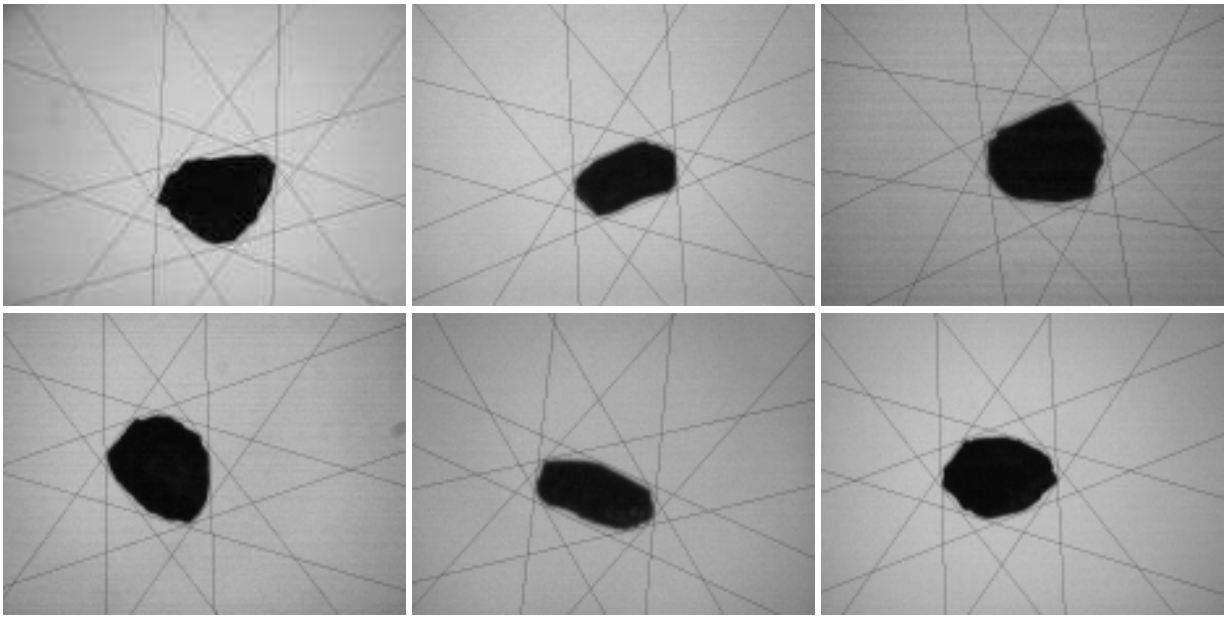


Figure 5.9: An example of a six-image set of a garnet. The epipolar tangents from each image are projected onto the remaining five images. The projected epipolar tangents are ideally tangent to the silhouettes; for real data that is not noise-free they are almost tangential.

To quantify the accuracy of the proposed calibration routine and to investigate how calibration accuracy varies with the number of ball image sets used, calibration was applied using randomly selected subsets of the ball image sets. The accuracy of the calibration was then quantified by using the computed calibration parameters to calculate the RMS ET error computed over 100 silhouette sets of garnets. Results are presented in a plot in Figure 5.10. The plot indicates that RMS ET errors of less than 0.4 pixels can be achieved if a sufficient number of ball image sets is used for calibration. The results also demonstrate that the parameter refinement by minimising ET error improves the accuracy of calibration parameters. The improvement is largest when a small number of balls is used, but is still significant when 15 ball image sets are used.

Calibration based on six ball sets was compared with calibration using 30 image sets of a calibration object [45]. The calibration object is illustrated in Figure 5.7b. During the calibration procedure, 2996 control points were located across the $30 \times 6 = 180$ input images. Silhouette sets of 98 uncut gemstones were used as a test set. ET errors for the 98 silhouette sets are plotted in Figure 5.11. Similar accuracy is observed for the ball-based calibration parameters and the calibration object parameters with RMS ET error values over

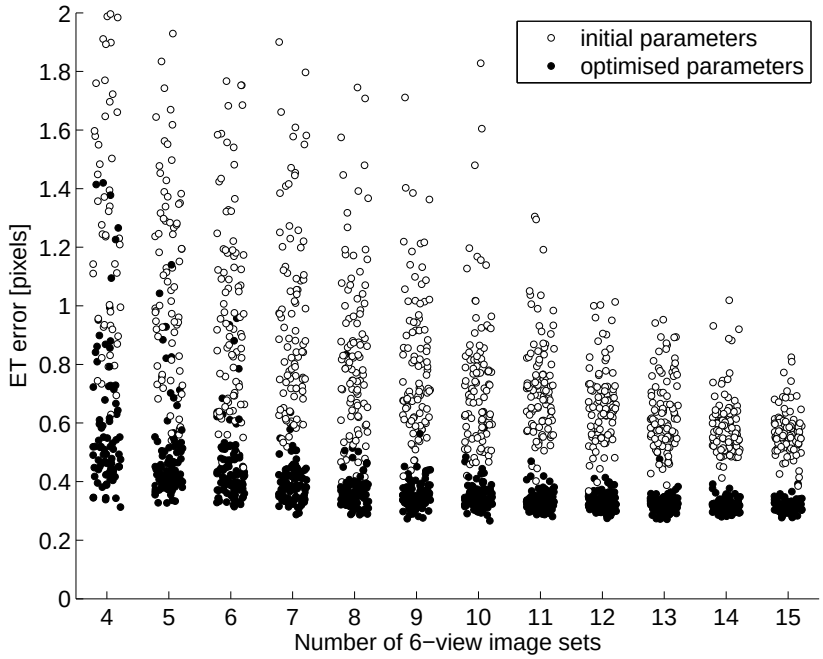


Figure 5.10: RMS ET error computed over a test set of 100 six-view silhouette sets of garnets using different calibration parameters. Each data point corresponds to the RMS ET error over 100 silhouette sets. Calibration parameters were determined from ball image sets randomly selected from 20 available image sets.

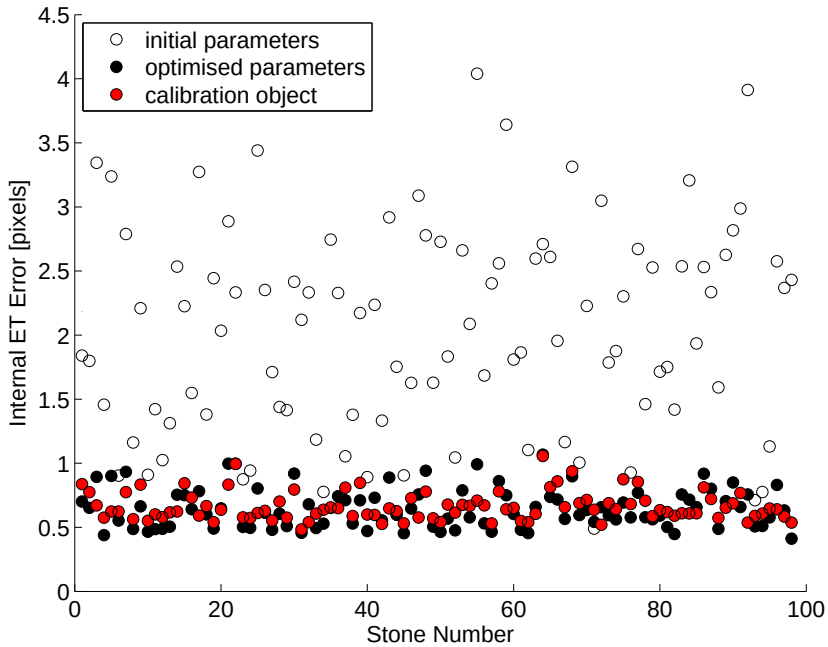


Figure 5.11: Internal ET error for 98 uncut gemstones computed using initial parameter estimates, optimised parameter estimates, and parameters computed using a calibration object.

all 98 silhouette sets of 0.667 pixels and 0.671 respectively. The initial parameter estimates produced an RMS ET error of 2.18 pixels over the 98 silhouette sets.

The ET error computed on a test set of stones provides no indication of the accuracy of scale enforcement, because the ET error is invariant to the absolute scale enforced. To quantify the performance of scale enforcement, it is necessary to image objects of known size. Image sets of three different sized balls were used to determine the accuracy of scale enforcement.

Table 5.1 shows the results of using the image sets of one ball for calibration, and then estimating the ball diameters of all three balls using the computed calibration parameters with scale enforced using the known diameter of the calibration ball. The diagonal of the table shows ball diameters that are exact, as the same ball image sets are used for calibration and for testing in these cases. The table indicates that the typical difference between estimated ball diameters the ground truth values is approximately 10 microns.

	5.54 mm ball	8.73 mm ball	10.50 mm ball
5.54 mm calibration parameters	5.540 mm (0.007 mm) **0%**	8.741 mm (0.021 mm) **+0.126%**	10.502 mm (0.013 mm) **+0.019%**
8.73 mm calibration parameters	5.527 mm (0.011 mm) **-0.237%**	8.730 mm (0.028 mm) **0%**	10.488 mm (0.018 mm) **+0.114%**
10.50 mm calibration parameters	5.531 mm (0.011 mm) **-0.162%**	8.738 mm (0.029 mm) **+0.091%**	10.500 mm (0.017 mm) **0%**

Table 5.1: Mean estimated ball diameters (with standard deviation over all image sets considered in brackets) for balls computed with calibration parameters determined from different sized balls. Percentage errors are shown in bold face. Nine image sets were used for the 5.54 mm ball; nine image sets were used for the 8.73 mm ball; and seven image sets were used for the 10.50 mm ball.

5.4 Summary

This chapter has described the geometric configuration of the multi-camera setup used for much of the work described in this thesis, and has presented the ball-based method used to calibrate the cameras.

Although some justification has been given for the choice of objective functions used for optimising the camera configurations, the objective functions are essentially ad hoc. This is the case because the multi-camera setup is to be used for several different applications whose performance can be measured in different ways, so the goal is to find a configuration that will be desirable for all applications. Two different approaches

(the frontier point criterion and the direction isolation criterion) yield the same configuration for six cameras. This configuration corresponds to viewing directions that are parallel to the face normals of a regular dodecahedron (one of the five Platonic solids).

Ball-based calibration produces ET errors of less than a pixel for image sets of garnets and gemstones. Approximately the same ET errors are obtained using a calibration object with coded targets.

The following chapters will quantify the performance that can be achieved for shape property estimation and matching applications using the camera configuration and calibration method described in this chapter.

Chapter 6

Merging Silhouette Sets

6.1 Introduction

This chapter describes a simple but effective method for merging two silhouette sets of the same rigid object into a single large silhouette set where all silhouette poses are specified in a common reference frame. The single large silhouette set allows a more accurate estimate of the 3D shape of the object to be made than either of the original silhouette sets.

The same method can be used to merge further silhouette sets of the object in different poses with the merged silhouette set. This allows an arbitrary number of silhouette sets of an object to be merged into a single large silhouette set.

The problem addressed in this chapter is another silhouette-based self-calibration problem. Here, it is the external camera parameters (i.e., pose parameters) rather than internal camera parameters that must be estimated. The approach taken here is the same as for the self-calibration problems addressed in Chapters 4 and 5: use the problem-specific constraints to obtain initial parameter estimates, and then refine the parameter estimates by minimising the ET error across silhouette pairs. The unknown parameters that are to be inferred from the two silhouette sets describe the relative pose between the two silhouette sets.

To obtain an initial estimate of the relative pose, the approximate 3D shape of the corresponding stone is estimated separately from each silhouette set. This can be done using the visual hull, or the VEMH as an estimate of 3D stone shape. The moments of the 3D shape are then used to estimate the components of relative pose between the silhouette sets. Centroids are used to estimate relative translation, principal directions are used to estimate relative orientation, and third order moments are used to resolve the four-way alignment ambiguity (since pairs of principal axes can be aligned in four ways).

This approach will be shown to work in most (but not all) cases for the silhouette sets of stones considered in this work. The method fails in cases in which third order moments do not resolve the four-way alignment

ambiguity of the principal axes, and in cases in which the principal axes of 3D approximations of the stone provide a poor estimate of the relative orientation between silhouette set pairs. In these failed cases, the initial parameter estimate does not lie within the basin of convergence of the optimal alignment parameters, and a local minimum that lies far from the optimal solution is located by Levenberg-Marquardt minimisation.

To address this issue, pose optimisation is attempted from successive different starting points based on different initial pose estimates, and the pose estimate corresponding to the lowest ET error (i.e., the smallest degree of silhouette inconsistency across the two silhouette sets) is selected. Initial pose estimates may be based on all four alignments of pairs of principal axes, and on random sampling of orientation space.

A version of the work described in this chapter was presented as a conference paper [46].

6.2 Related Work

One of the earlier methods to create refined visual hull models by making use of two or more silhouette sets of an object is described by Wingbermühle et al. [137]. The relative pose between silhouette sets is determined by means of an optimisation procedure. The cost function is the mean squared distance between surface points of the first visual hull and the closest surface points of the second visual hull. A starting point for the optimisation is determined from the principal axes and centres of gravity (centroids) of the two visual hulls. If the cost associated with the starting point is too high, then a heuristic approach is used: the relative rotation is adjusted incrementally about each of the principal axes in steps of 15° until an adequate starting point is found. Since the cost function is based on the visual hull rather than the observed silhouettes, there is no reason to expect that the correct alignment should correspond to a cost function minimum, even with exact silhouette sets.

Cheung et al. [26, 27] describe a method for determining rigid transforms for aligning image sets of the same object in different poses. Although their goal is the same as for the method described in this chapter, they make use of colour stereo matching in addition to silhouette information, whereas in this thesis only silhouette images are considered. Their method involves using silhouettes to constrain the search for corresponding points along viewing edges (which they term *bounding edges*). Pose parameters are iteratively adjusted to minimise a cost function based on colour consistency across image sets. Their setup therefore requires objects and lighting such that (1) both silhouettes and foreground texture can be reliably measured from images, and (2) colour and intensity varies as little as possible with viewpoint (i.e., a Lambertian model must be a good approximation). This thesis takes a different approach, and lighting is set up to obtain the best possible silhouettes at the cost of discarding foreground texture.

Cheung's motivation for using colour information in addition to silhouettes is that alignment using silhouettes is 'inherently ambiguous' [26]. To demonstrate the ambiguity, it is shown that more than one alignment of certain specific noise-free silhouette set pairs is exactly consistent.

Despite Cheung's illustration of certain specific ambiguous cases, the view taken here is that there is no need to discard the possibility of alignment based on silhouettes alone. Although certain specific cases are inherently ambiguous, they are unlikely to occur in practice. This is especially so with arbitrarily oriented natural objects such as stones, for which an ambiguous pair of silhouette sets arising by coincidental alignment appears to be close to impossible. Of course, real silhouette sets are noisy and are therefore inexact; it is certainly plausible that attempting to align silhouette sets consisting of too few views or too much noise may fail. This chapter will demonstrate that merged silhouette sets captured using the imaging setups considered in this work are sufficiently accurate to provide measurable improvement in estimates of size and shape properties that are of interest to particle shape analysts.

Subsequent to our initial publication [46] of the method described in this chapter, Hernández [39] describes a solution to the same problem in the context of creating refined visual hull models of museum pieces such as ornamental pitchers. Calibrated sequences of silhouettes are captured using a turntable; this provides a silhouette set of the object. The object is then reoriented on the turntable and another silhouette set is captured. The method of merging the silhouette sets is essentially the same as the approach described here: pose and scale parameters are adjusted to optimise a measure of silhouette consistency. Instead of using ET error, Hernández proposes an alternative measure of silhouette consistency that he terms *silhouette coherence*. Silhouette coherence measures the extent to which visual hull projections match the corresponding silhouettes. This has the advantage of using more information contained within the silhouettes than the ET error, but comes at the cost of having a discretised nature, and requires selecting the value of a tunable distance offset parameter. Results demonstrate that the visual hull model formed from the merged silhouette set is a better approximation to the shape of the object than visual hulls formed from either of the original silhouette sets.

Wong [138] describes merging individual silhouettes with silhouette sets. Since individual silhouettes are used, approximate 3D models cannot be used to provide initial pose estimates, and initial pose estimates must be provided by the user. The pose estimate is then refined by minimising ET error.

6.3 Moments for Initial Parameter Estimates

A triangular mesh model that approximates the 3D shape of the corresponding stone is computed for each silhouette set. This is done using the visual hull or VEMH described in Chapter 3. The moments of the mesh models are used to form initial parameter estimates for aligning silhouette sets of the same object.

6.3.1 Computing Moments from Triangular Meshes

The moments of the solid enclosed by a triangular mesh can be elegantly computed by visiting each triangle and forming a polynomial function of the vertex coordinate values. The basis for the method is described by Lien and Kajiya [81], and Zhang and Chen [143] derive explicit equations for third order moments. Mirtich

[94] describes an alternative approach in which the Divergence Theorem is used to reduce volume integrals to surface integrals. Zhang and Chen's equations are presented in this section.

The moments of the solid enclosed by a mesh are defined as

$$M_{pqr} = \iiint x^p y^q z^r \rho(x,y,z)\,dx\,dy\,dz, \tag{6.1}$$

where $\rho(x,y,z) = 1$ for points inside the mesh, and $\rho(x,y,z) = 0$ for points outside the mesh.

The moment equations depend on a determinant T that must be computed for each triangular face:

$$T = x_1(y_2z_3 - y_3z_2) + y_1(x_3z_2 - x_2z_3) + z_1(x_2y_3 - x_3y_2). \tag{6.2}$$

For convenience, the equations derived by Zhang and Chen are restated here (using a slightly different format for clarity):

$$M_{000} = 1/6 \sum T \tag{6.3}$$
$$M_{100} = 1/24 \sum T(x_1 + x_2 + x_3) \tag{6.4}$$
$$M_{110} = 1/120 \sum T(2x_1y_1 + 2x_2y_2 + 2x_3y_3 + x_1y_2 + x_2y_1 + x_2y_3 + x_3y_2 + x_3y_1 + x_1y_3) \tag{6.5}$$
$$M_{200} = 1/60 \sum T(x_1^2 + x_2^2 + x_3^2 + x_1x_2 + x_2x_3 + x_1x_3) \tag{6.6}$$
$$M_{300} = 1/120 \sum T(x_1^3 + x_2^3 + x_3^3 + x_1^2(x_2 + x_3) + x_2^2(x_1 + x_3) + x_3^2(x_1 + x_2) + x_1x_2x_3). \tag{6.7}$$

The summation sign indicates summation over all triangles that make up the mesh. The triangle vertices are (x_1,y_1,z_1), (x_2,y_2,z_2) and (x_3,y_3,z_3). Since triangles share vertices with other triangles, vertices will be visited on multiple occasions. The equations for the other relevant moments can be inferred from the equations given above.

To determine an initial estimate of the relative pose between two silhouette sets A and B, the centroid and principal axes are computed for each of the two meshes that are 3D approximations to the stone computed from each silhouette set. For each mesh, a 4×4 rigid transform matrix M that aligns the principal axes of the mesh with the x-, y-, and z-axes is computed:

$$\mathsf{M} = \begin{pmatrix} \mathsf{R}_3\mathsf{R}_2 & -\mathbf{c} \\ \mathbf{0}^\mathsf{T} & 1 \end{pmatrix}, \tag{6.8}$$

where $\mathbf{c}$ is the centroid of the solid enclosed by the mesh, R_2 is a rotation matrix that aligns the principal axes, and R_3 is a rotation matrix that is used to resolve the four-way alignment ambiguity.

Once rigid transform matrices M_A and M_B have been computed for the two silhouette sets A and B, the initial

pose estimate M_{init} to transform from B's world reference frame to A's world reference frame is computed:

$$\mathsf{M}_{\text{init}} = \mathsf{M}_A^{-1}\mathsf{M}_B. \tag{6.9}$$

To compute M, the following steps are applied to each mesh. First, the mesh is translated so that its centroid $\mathbf{c}$ lies on the origin. The centroid is calculated as

$$\mathbf{c} = \begin{pmatrix} M_{100} \\ M_{010} \\ M_{001} \end{pmatrix} / M_{000}, \tag{6.10}$$

where M_{000} is the volume bounded by the mesh.

Next, a 3×3 matrix of second order moments (a covariance matrix) is constructed:

$$S = \begin{pmatrix} M_{200} & M_{110} & M_{101} \\ M_{110} & M_{020} & M_{011} \\ M_{101} & M_{011} & M_{002} \end{pmatrix}. \tag{6.11}$$

The columnwise eigenvectors $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ of this matrix are used to form a rotation matrix $\mathsf{R}_2 = [\mathbf{e}_1\, \mathbf{e}_2\, \mathbf{e}_3]$. The mesh vertices are then multiplied by R_2^{-1} to align the principal axes of the mesh with the *x*-, *y*- and *z*-axes. This is done so that the third order moments can be computed.

The two third order moments M_{003} and M_{030} are computed to resolve the four-way alignment ambiguity. (This arises because $\mathbf{e}$ and $-\mathbf{e}$ are both valid eigenvectors.) The value of R_3 is determined from the signs of M_{003} and M_{030} as indicated in Table 6.1.

$M_{003} > 0$	$M_{030} > 0$	R_3
no	no	180° rotation about *x*-axis
no	yes	180° rotation about *y*-axis
yes	no	180° rotation about *z*-axis
yes	yes	3×3 identity matrix

Table 6.1: Selecting R_3 based on the signs of the third order moments M_{003} and M_{030}.

This ensures that the composite rotation $\mathsf{R}_3\mathsf{R}_2$ aligns the original mesh so that $M_{003} > 0$ and $M_{030} > 0$.

In certain cases, the M_{003} and M_{030} values of the visual hull or the VEMH may not match the sign of the M_{003} and M_{030} values of the stone. This is particularly likely to occur when the skewness of the volume distribution along a particular principal axis is close to zero. These cases may result in silhouette set pairs being out of alignment by 180°. In order to find the next most likely alignments, a rotation of 180° about the *z*- or *y*-axes can be used. These produce alignments in which the signs of either M_{003} or M_{030} will differ

for a pair of visual hulls or VEMHs (though the values for the stone may share the same sign). To obtain the fourth alignment in which the values of neither M_{003} nor M_{030} share the same sign across the two mesh approximations, a 180° rotation about the x-axis is used (i.e., a 180° rotation about the y-axis followed by a 180° rotation about the z-axis).

6.3.2 Experiments Using Moment-Based Initial Estimates

Synthetic Data

Experiments on synthetic data were carried out to investigate the performance of moment-based estimates of initial pose. Synthetic data has the advantage of having exactly known ground truth values for pose.

Refined visual hull models formed from a data set of garnets were used to create synthetic silhouette images. The data set is illustrated on page 221 of Appendix C.

Exact polygonal silhouettes that were generated from projections of the mesh models were rasterised to create synthetic digital images, and polygonal boundaries were extracted using a subpixel segmentation method that is described in Appendix A. The resultant digital images were downsampled to create sets of images at different resolution levels (see Figure 6.1). Synthetic data were generated for different configurations of

(a) 1/4 resolution (b) 1/8 resolution (c) 1/16 resolution (d) 1/32 resolution

Figure 6.1: An example of a synthetic silhouette shown at four different resolution levels.

different numbers of cameras: 2-, 3-, 4-, 6-, and 10-camera configurations were investigated. The configurations are based on the Platonic solids as illustrated in Figure 5.5. The synthetic 6-camera setup corresponds to the configuration of the real 6-camera setup. Two runs of silhouette sets were synthesised for each case. In each case the stone models were oriented using a uniform random rotation, and were positioned with their centroids at the intersection of optical axes. Camera depths were based on the depths of the six real cameras from the stones.

Pose optimisation was carried out using the Levenberg-Marquardt method. The orientational component of pose was parameterised using quaternions. This eliminates potential gimbal lock problems at the cost of

an extra parameter: the relative pose is parameterised with seven parameters, but has only six degrees of freedom.

The ET error is computed *across* the two silhouette sets. This means that each silhouette pair in the first set is paired with each silhouette in the second set.

Figure 6.2 shows empirical CDFs (cumulative distribution functions) for the angle between the computed relative pose and the ground truth relative pose. The angle provides a useful single-number measurement of the dissimilarity between two poses. This approach comes at the cost of discarding the positional component of pose. At this stage, it is useful to consider the angle for investigating the behaviour of the proposed pose optimisation method. Later in this chapter, practical application-based methods of accuracy will be considered too.

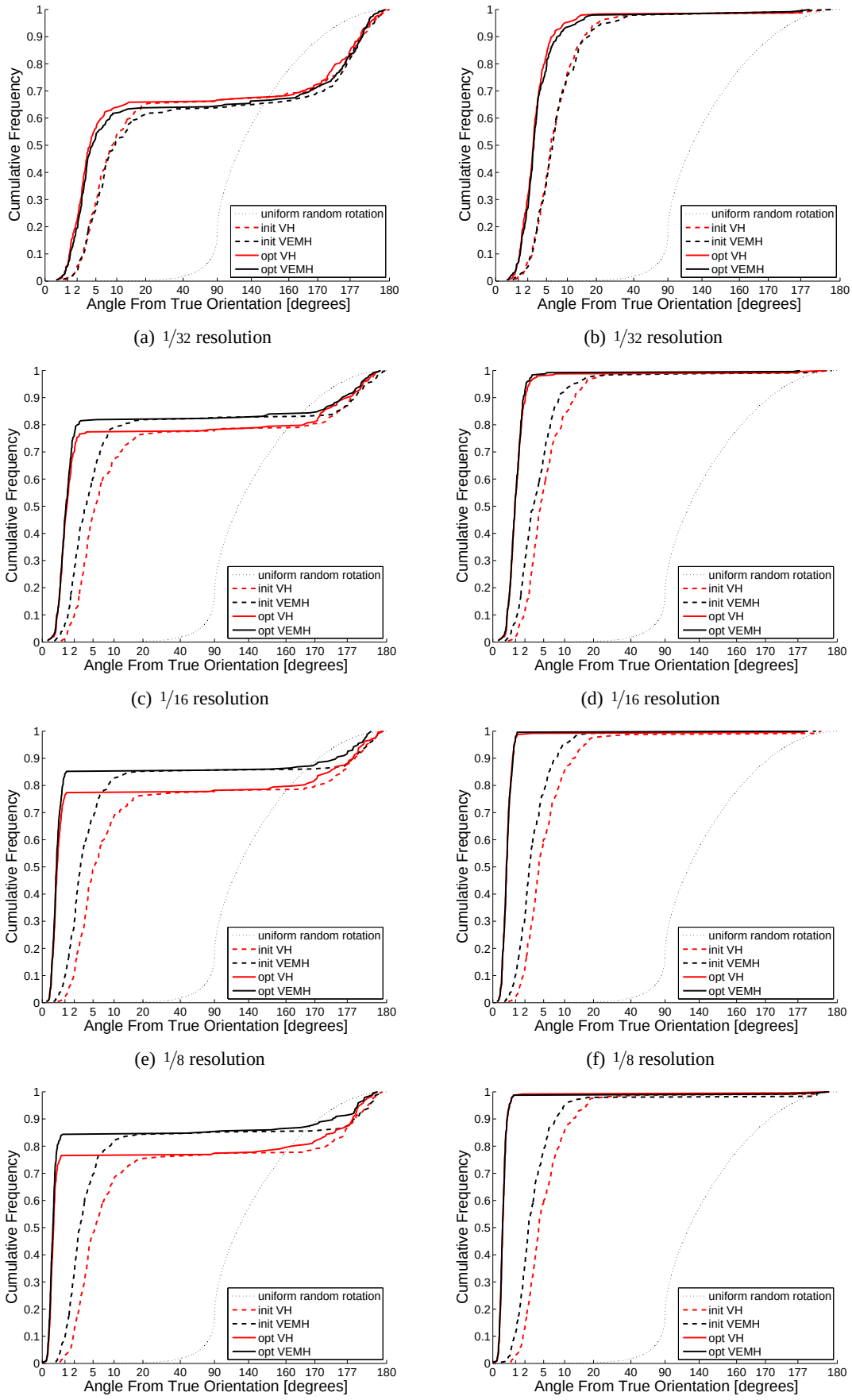
The figure shows CDFs for the initial pose estimates ('init') as well as optimised pose estimates ('opt') for initial estimates based on the moments of both the visual hull ('VH') and the VEMH. Results are shown for a 6-camera setup and experiments are repeated at different levels of image resolution. A nonlinear scale (based on a sinusoidal transformation) is used for the horizontal axis. This aids visualisation, because interesting portions of the CDFs occur near 0° and 180°, whereas the CDFs tend to have almost constant value between 45° and 135°. Also shown on each plot is the CDF corresponding to a uniform random orientation. The plots on left side show the results of optimisations based on a single initial pose estimate in which third order moments are used to resolve the four-way alignment ambiguity of the principal axes. The plots on the right side show the results of pose optimisation in which four initial pose estimates based on the four alignments of the principal axes are considered. The computed pose with the lowest ET error is selected.

The CDFs allow one to read off the proportion of cases where estimated poses are within a certain angular displacement from the true pose. The plots on the left suggest that in approximately 80% of cases, optimisation based on a single initial pose estimate leads to a pose within two degrees of the true pose. The closeness to the true pose improves with higher resolution silhouettes. The plots on the right show that approximately 98% of cases lead to a pose within two degrees of the correct pose when all four alignments of the principal axes are considered. Although a threshold of two degrees is arbitrary, the horizontal sections of the CDFs suggest that there is a large range of threshold angles for which these proportions are insensitive.

The plots indicate that the VEMH slightly outperforms the visual hull for alignments based on third order moments, but performance is approximately the same when considering four initial estimates per case. This suggests that the VEMH provides a better estimate of the skewness of the volume distribution of the corresponding stone than the visual hull.

The results of experiments repeated with different numbers of cameras is shown in Figure 6.3. The plots indicate that as the number of cameras is increased, the proportion of estimated poses that are close to the true pose increases.

Figures 6.4 and 6.5 show plots of normalised ET error versus angle from the true alignment for the experiments whose results are displayed in Figures 6.2 and 6.3 respectively. Normalised ET error is the RMS ET



(a) $1/32$ resolution

(b) $1/32$ resolution

(c) $1/16$ resolution

(d) $1/16$ resolution

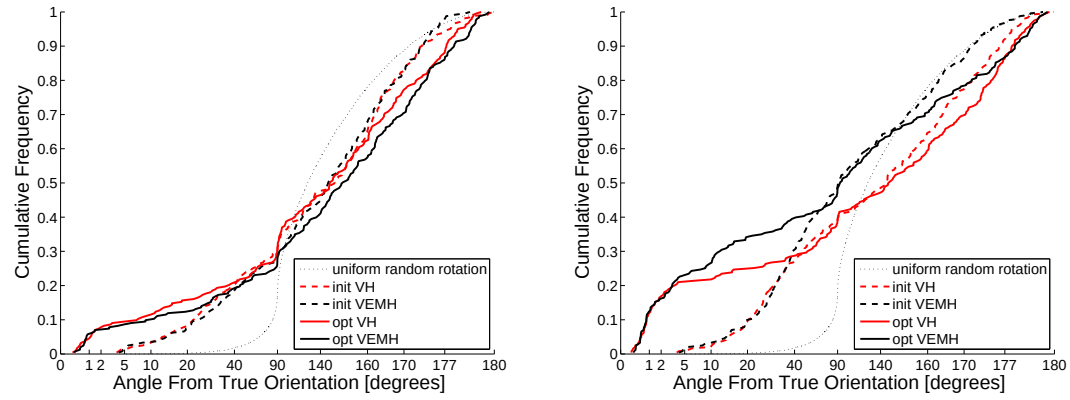
(e) $1/8$ resolution

(f) $1/8$ resolution

(g) $1/4$ resolution

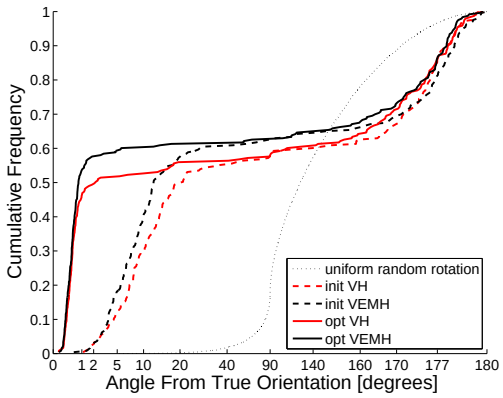
(h) $1/4$ resolution

Figure 6.2: CDFs of angle between computed pose and ground truth pose for merging 6-view silhouette sets.The left column shows results based on a single initial pose estimate based on moments up to order three. The right column shows results based on the best (lowest ET error) of four initial pose estimates from the four possible alignments of principal axes.

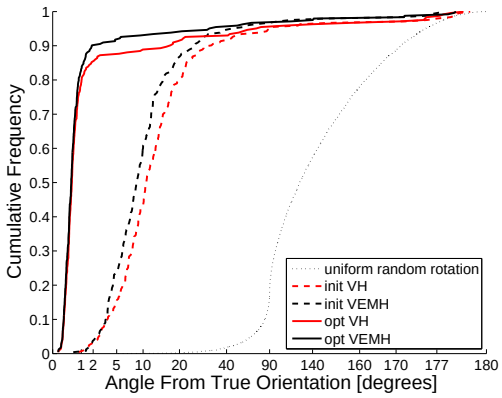


(a) 2 cameras

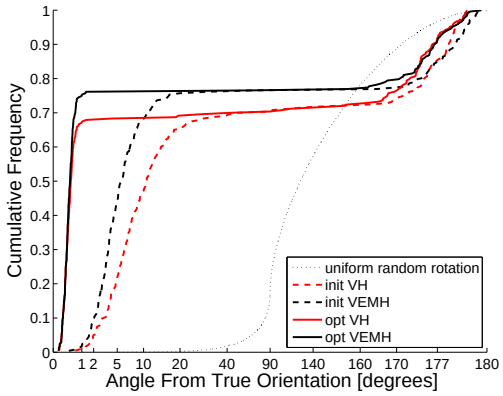
(b) 2 cameras



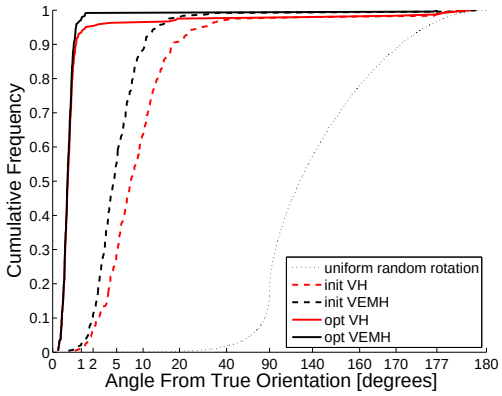
(c) 3 cameras



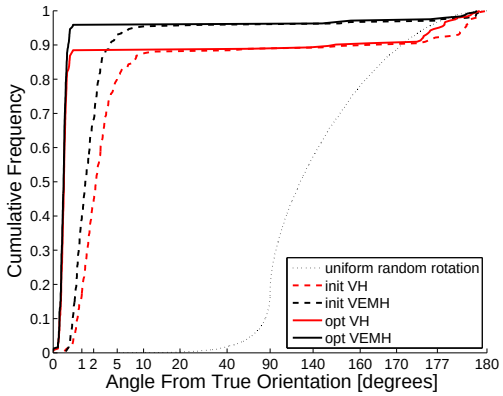
(d) 3 cameras



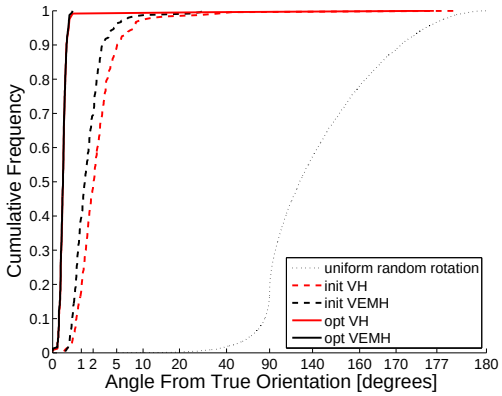
(e) 4 cameras



(f) 4 cameras

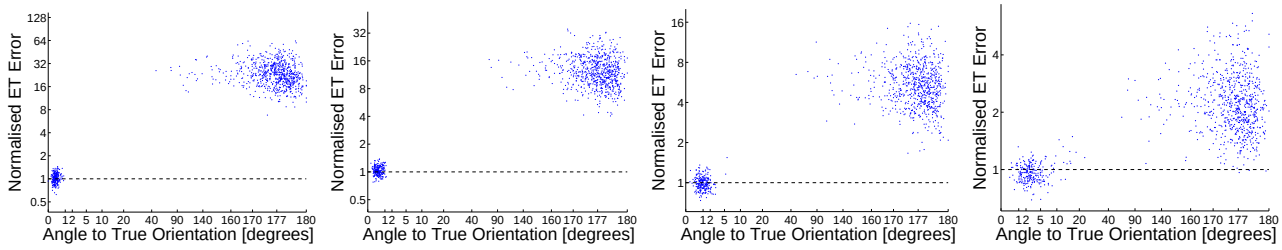


(g) 10 cameras



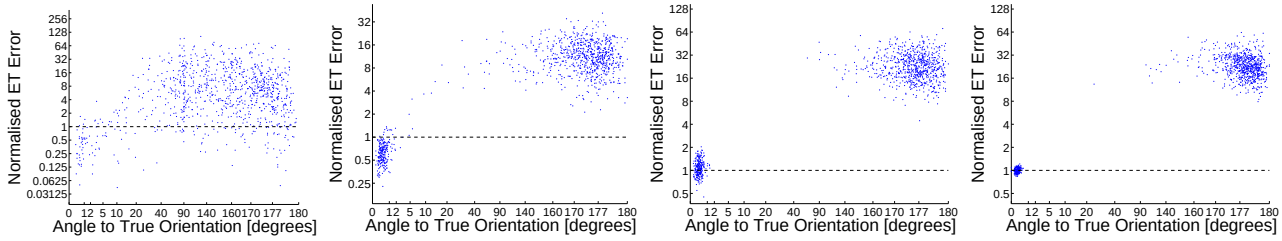
(h) 10 cameras

Figure 6.3: CDFs of angle between computed pose and ground truth pose for silhouettes sets formed with different numbers of cameras at 1/4 resolution.The left column shows results based on a single initial pose estimate based on moments up to order three. The right column shows results based on the best (lowest ET error) of four initial pose estimates from the four possible alignments of principal axes.



(a) 1/4 resolution (b) 1/8 resolution (c) 1/16 resolution (d) 1/32 resolution

Figure 6.4: Plots of normalised ET error versus angle between computed pose and ground truth pose for the 6-camera setups with different resolution levels as considered in Figure 6.2.



(a) 2 cameras (b) 3 cameras (c) 4 cameras (d) 10 cameras

Figure 6.5: Plots of normalised ET error versus angle between computed pose and ground truth pose for the different camera setups considered in Figure 6.3. The 1/4 resolution level is used.

residual error computed across the merged silhouette set pair divided by the RMS ET residual error computed within each of the two silhouette sets. Since the six degrees of freedom of pose optimisation is small with respect to the number of different outer epipolar tangent planes ($2n^2$ for n cameras) that generate the residual errors, the normalised ET should be close to one for correctly aligned silhouette set pairs. The plots show two distinct clusters that correspond to correct alignment (low ET error and small angle to the true pose) and incorrect alignment (high ET error and large angle to the true pose). Note that both axes are nonlinear: this aids visualising the clusters on the bottom left that are substantially more compact than the clusters on the top right.

In the case of two cameras (Figure 6.5a), two clusters are less distinct than for larger numbers of cameras. In the case of three cameras (Figure 6.5b), two clusters are clearly visible, yet the bottom left cluster consists of normalised ET errors less than one. This is evidence of overfitting: the $2 \times 3^2 = 18$ outer tangent planes that generate ET errors across silhouette set pairs is not much larger than the six degrees of freedom of the pose optimisation. For larger numbers of cameras, the normalised ET errors tend to cluster around a value of one for the lower left cluster. The lower left cluster tends to become more compact and move towards an error angle of zero, as image resolution is increased (Figure 6.4), and as the number of cameras is increased (Figure 6.5).

The plots in Figures 6.4 and 6.5 indicate that ET errors tend to form two clusters, one of whose alignments are substantially closer to the true alignment than the other. This supports the use of ET error to investigate the behaviour of real data.

Real Stone Images

Pose optimisation using moments for determining initial parameter estimates was applied to the data set of images of 220 pieces of gravel, and 246 garnet stones. The pieces of gravel were imaged using the two-mirror setup described in Chapter 4, and the garnets were imaged using the 6-camera setup described in Chapter 5.

Figure 6.6 illustrates the results of pose optimisation applied to two 5-view silhouette sets. Only one of the two silhouette sets is shown. The five images are cropped out of the original image, since all five silhouettes were captured in a single image using the two-mirror setup. (There is some overlap present in the second and fourth images.) In this case, the computed pose appears to be close to the correct pose, since all the projected tangents are approximately tangent to the silhouettes. The figure also shows projections of the 10-view visual hull onto the original silhouettes. The visual hull projections come close to covering the original silhouettes. This is consistent with a pose that is close to the true relative pose.

Figure 6.7 illustrates the results of pose optimisation applied to the same pair of silhouette sets, but from a different starting point. The initial pose estimate used here causes the principal axes of the two VEMHs to be aligned, but the third order moments do not have the same signs. In this case, pose optimisation appears to have found a pose that is far from the true pose. The projected epipolar tangents are not approximately tangent to the silhouettes (as indicated by red line segments), and the visual hull projections leave large portions of the silhouettes uncovered. The silhouettes in the bottom row have been coloured using a distance transform, so that the distance of uncovered portions from the silhouette boundary is apparent.

Figure 6.8 shows CDFs of ET error for the garnet and gravel data sets. Similar behaviour to the experiments with synthetic data is observed. In approximately 80% of cases, for both the garnet and the gravel data, the normalised ET error is below 2.0 when optimising pose from a single starting point based on third order moments. The VEMH curves lie above the visual hull curves for both data sets, indicating that the VEMH provides a better starting point. However, the two curves are similar in shape when using four starting points based on four alignments of principal axes. The plots also show results computed using the CDRH to approximate 3D stone shape. (The CDRH is defined in Section 3.4.2 on page 39.) The poor performance of the CDRH demonstrates the importance of using varying rim depths as for the VEMH, rather than constant depth rims. The additional complexity of computing the VEMH rather than the CDRH is therefore justified in this context.

Qualitative Results for 3D Multimedia Content Creation

The proposed method of merging silhouette sets is useful not only for characterising stone shape, but also for reconstructing the 3D shape of arbitrary objects for 3D multimedia content creation. Easily recognisable shapes help to provide a qualitative demonstration of the effectiveness of the proposed method for creating more accurate 3D reconstructions than can be made from any of the original silhouette sets.

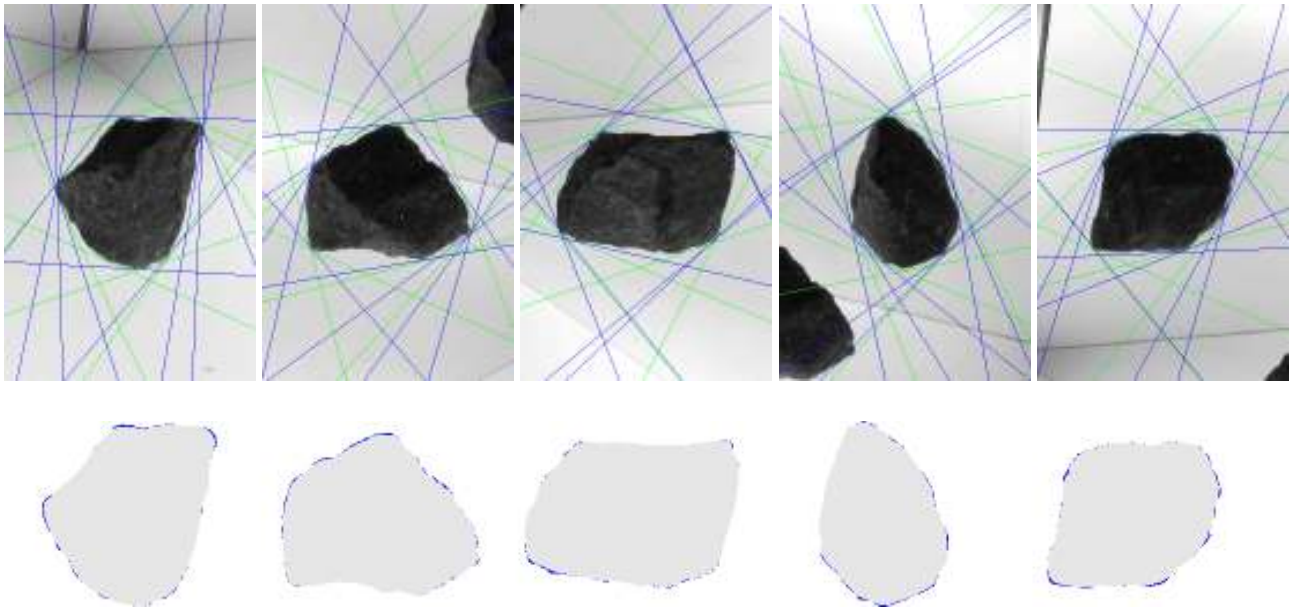


Figure 6.6: Correct alignment computed using moments up to order three for an initial pose estimate. The top row shows projected epipolar tangents within the silhouette set in green, and across silhouette sets in blue. The bottom row shows silhouettes in colour with 10-view visual hull projections in grey.

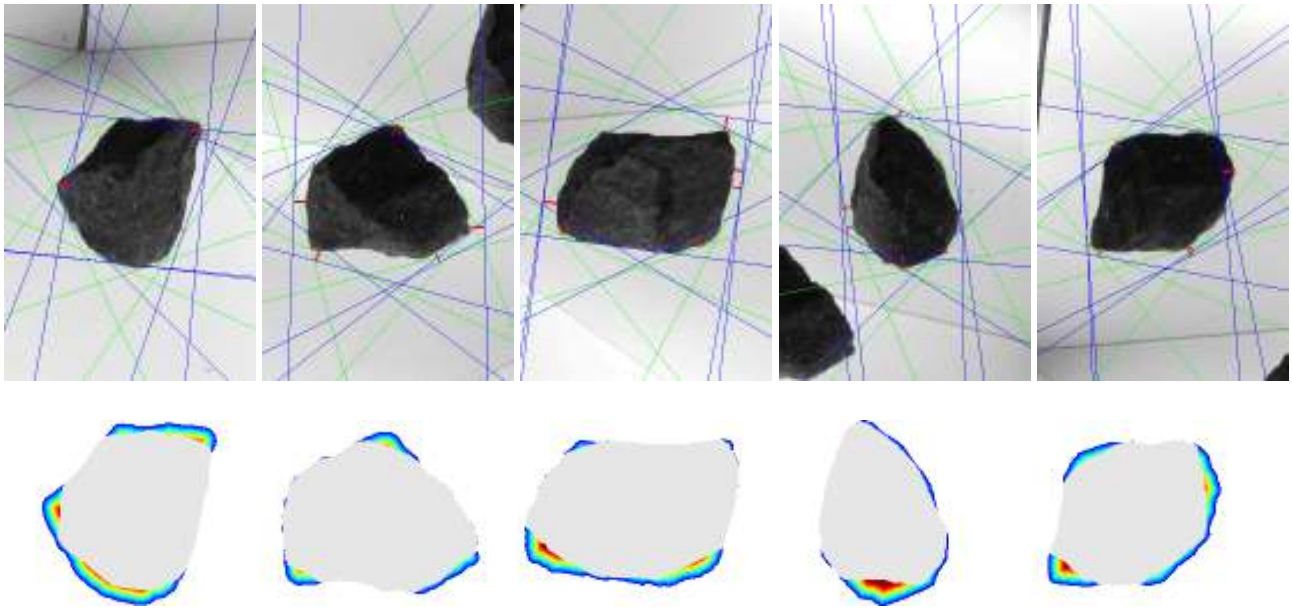
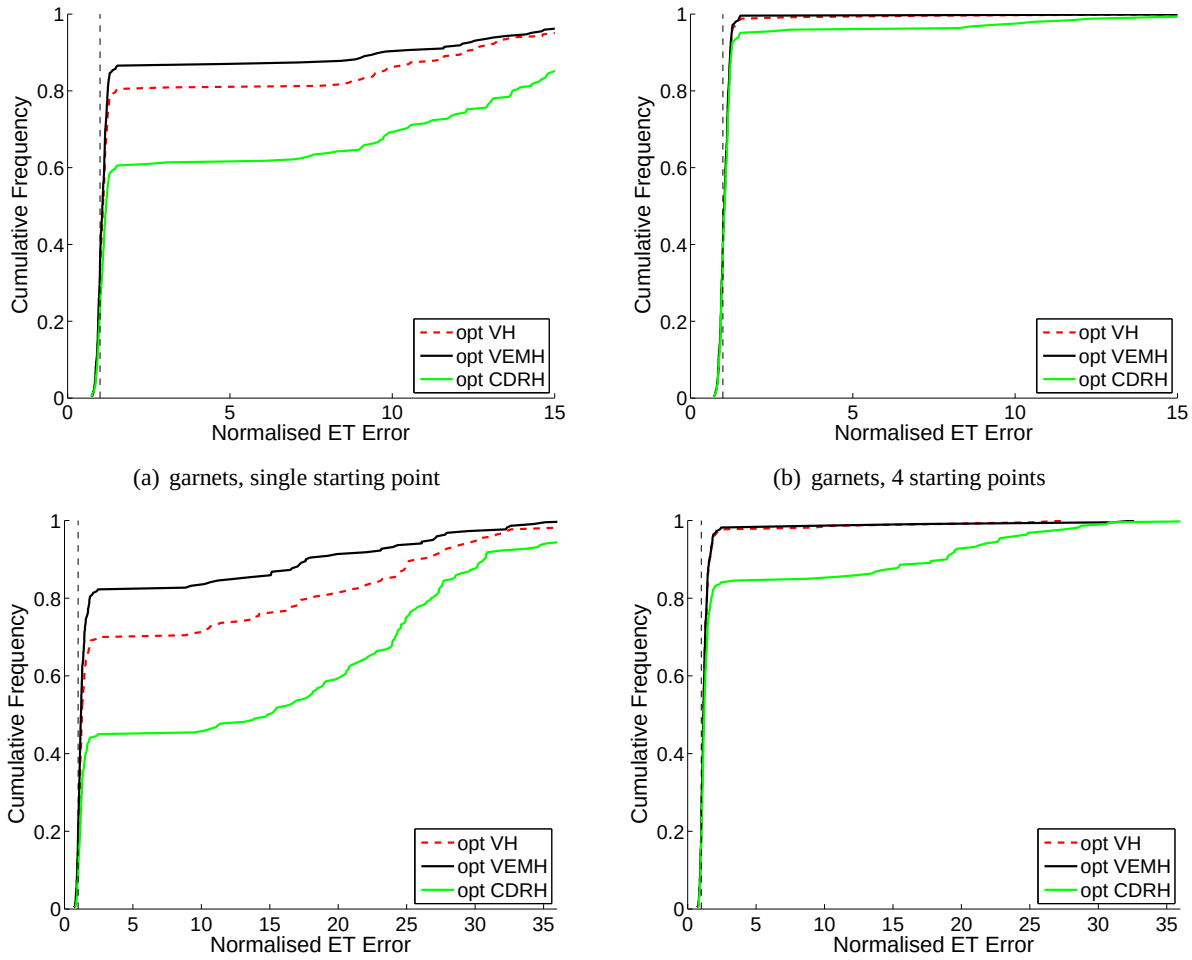


Figure 6.7: Incorrect alignment. The top row shows projected epipolar tangents within the silhouette set in green, and across silhouette sets in blue. Distances from the tangents to the silhouette are in red. The bottom row shows silhouettes in colour with 10-view visual hull projections in grey.



(a) garnets, single starting point

(b) garnets, 4 starting points

(c) gravel, single starting point

(d) gravel, 4 starting points

Figure 6.8: CDFs of normalised ET error computed using real image data. Dashed vertical lines indicate normalised ET error of 1.0.

Figure 6.9 shows an example of visual hulls formed from four 5-view silhouette sets. The images were captured using a 5-camera setup that was a predecessor to the 6-camera setup described in Chapter 5. The visual hull formed from the merged 20-view silhouette set is also shown. The merged silhouette set was obtained by merging the silhouette sets one at a time. A final parameter adjustment of all pose parameters using ET error computed across all silhouette pairs was found to result in negligible further reduction in ET error. Notice that the 3D reconstruction of the wingnut from the merged silhouette set appears to be more accurate than any of the original 5-view visual hulls.

Figure 6.10 shows another example, a toy cat, using images captured with a 5-camera setup. Again, the 20-view visual hull formed from the merged silhouette set appears to be a better 3D reconstruction than any of the original 5-view visual hulls, each of which have substantial regions of extra volume. The figure also shows the computed positions of the 20 silhouette views as well as the corresponding visual cones. Note how the viewpoints provide a good coverage of the viewing sphere.

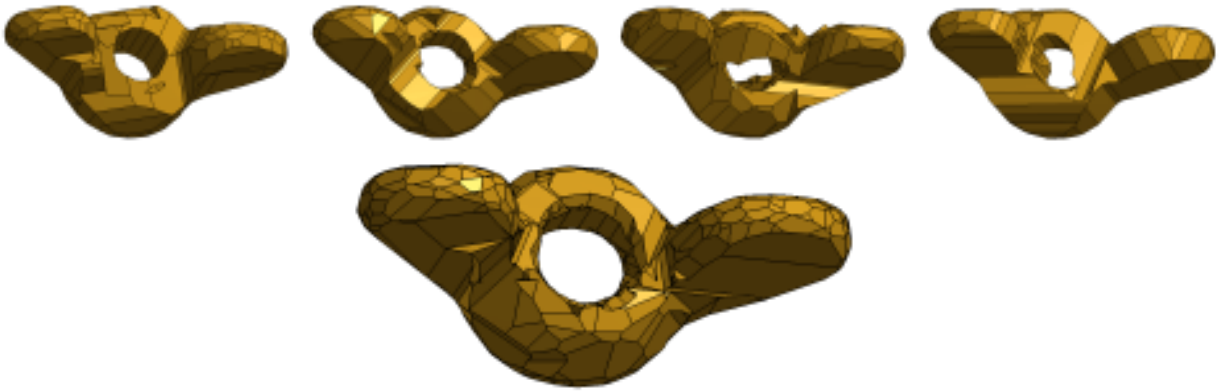


Figure 6.9: Visual hulls of a wing nut. The top row shows four 5-view visual hulls. The bottommost illustration shows the refined 20-view visual hull obtained by merging the four 5-view silhouette set into a single large set containing 20 silhouettes.

The proposed method provides an alternative means for capturing silhouettes from many well-distributed viewpoints using the two-mirror setup. In Chapter 4, a method was described in which the camera is moved with respect to the mirror and object so that a good coverage of the viewing hemisphere can be obtained. The proposed method provides another approach: the object is moved and the camera and mirrors stay fixed. This requires a tripod or some other method of fixing the camera with respect to the mirrors. Figure 6.11 shows an example in which three images of a toy moose are captured using the two-mirror setup. The figure illustrates once more that a refined visual hull model formed from a merged silhouette set is a better reconstruction than can be formed from any of the original silhouette sets.

An advantage of using the proposed method with the two-mirror setup is that images can be captured over the entire viewing sphere (as opposed to a viewing hemisphere). This allows 3D reconstructions to incorporate texture, and also allows foreground information to be incorporated for estimating 3D shape. Figure 6.12 shows an example in which a toy cheetah is modelled. For each object pose, two images are captured: one with the backlight switched on to facilitate silhouette extraction, and another with no backlight to capture the foreground texture of the object.

6.4 Estimating Shape Properties

This section describes several experiments that quantify the repeatability and accuracy with which shape properties can be estimated using the proposed merging method. Readers who are not specifically interested in shape property estimation may wish to skip this section, and continue reading Section 6.5 on page 128.

To address the problem of initial pose estimates that do not lead to sufficiently low ET error, the best optimisation based on 100 starting points formed with uniform random sampling of orientation space was used. The large number of starting points ensures that a pose close to the true pose is likely found, but this comes

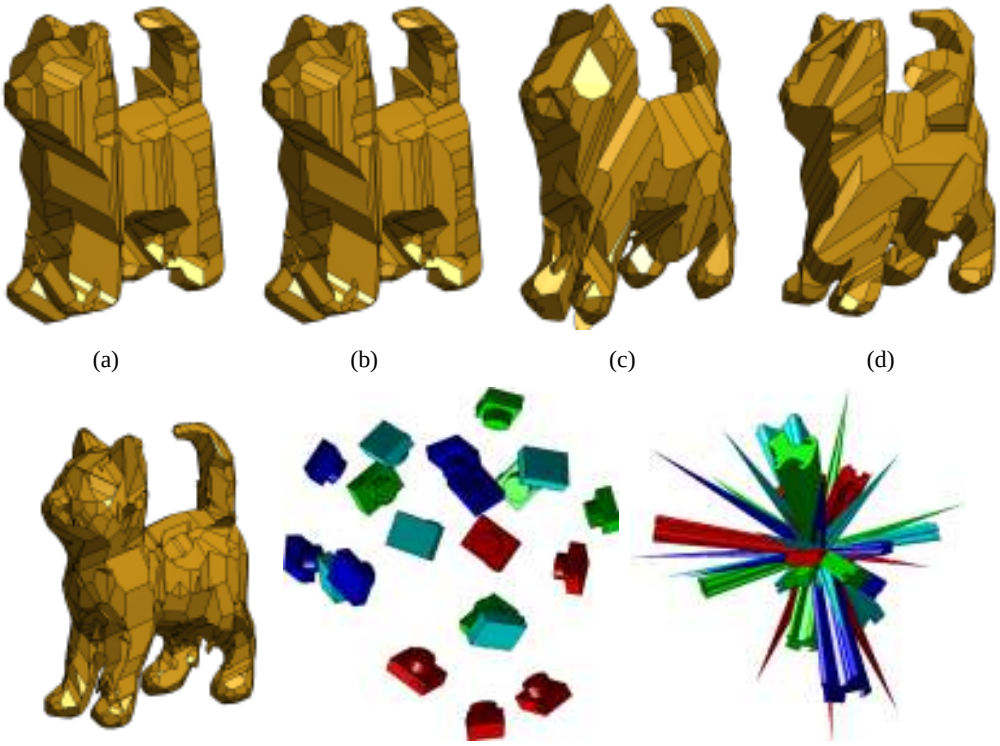


Figure 6.10: Visual hull models of a toy cat: (a)–(d) four models each built from five silhouettes, (e) the model built from the 20 silhouettes used in (a)–(d) after the poses of all silhouettes have been determined in a common reference frame. The camera poses corresponding to the twenty views are shown in (f), and the visual cones are shown in (g).

Figure 6.11: Reconstructing the 3D shape of a toy moose. The top row shows the three input images, and the bottom row shows the corresponding 5-view visual hulls. The rightmost visual hull is formed from the merged 15-view set.

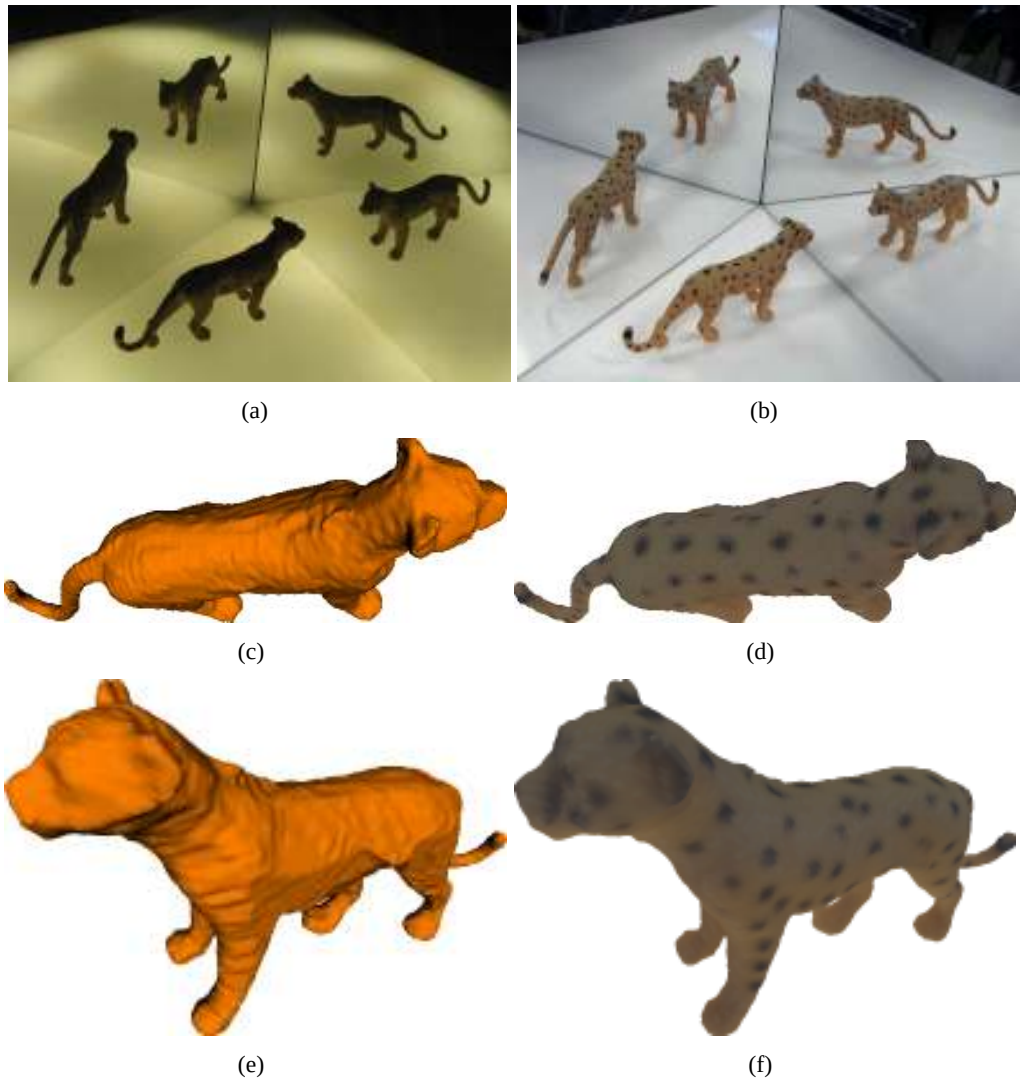


Figure 6.12: Reconstruction of a toy cheetah by merging five 5-view silhouette sets. Input images were captured using two mirrors and a backlight: (a) shows an example of a backlit image and (b) shows an example of the corresponding frontlit image. After calibration and pose optimisation using silhouettes extracted from the backlit images, the frontlit images were used to build a photo-consistent three-dimensional model. This was done with software created by Mathew Price (University of Cape Town) that is based on the work of Vogiatzis et al. [132]. The software uses optimisation based on graph-cuts to compute a textured photo-consistent mesh. Two novel views of the three-dimensional model with and without texture are shown in (c)–(f).

at the cost of increased running time. In Chapter 9, where the pose optimisation is used for matching, a framework is introduced that removes the need for specifying the number of starting points in advance.

To be useful in the context of particle shape analysis, the proposed method must produce silhouette sets for which shape features can be estimated more accurately from the merged silhouette set than from any of the original silhouette sets. The shape features measured from the merged silhouette set should also be more accurate than the mean value computed from both original silhouette sets, otherwise the merging method is not providing any benefit.

6.4.1 Volume Estimation with Synthetic Data

A set of volume estimation experiments was carried out using the synthetic garnet data. Synthetic data provide two important advantages over real data: (1) exact ground truth is known for the stone volumes, (2) the exact ground truth is known for the relative poses between silhouette set pairs. Knowing the ground truth relative poses allows one to compare the accuracy of volume estimates based on inferred pose with those computed using the actual pose. This provides an indication of how well the proposed method performs compared with the optimal (i.e., exact) alignment.

Table 6.2 presents the results of the volume estimation experiments for a synthetic six-camera setup. The table shows the mean percentage error of volume estimation using volumes of the visual hull or VEMH as estimates of stone volume. (The VEMH can be used because the synthetic stones are convex.) The mean percentage error gives an indication of the systematic error associated with a volume estimate. Since the visual hull is an upper bound for the volume of the stone that produced the silhouettes, the volume estimates tend to be overestimates and the percentage errors are therefore positive. However, when computed using noisy data, cone intersections will erroneously carve away extra volume, yet cannot add extra volume. This means that with sufficient noise, the visual hull-based volume estimates become underestimates. This is the case with the rightmost column in which the images have the greatest degree of downsampling.

The table shows RMS percentage errors for volume estimates computed using the equation

$$V_{\text{est}} = kV_{\text{shape}}, \tag{6.12}$$

where V_{est} is the volume estimate, V_{shape} is the volume of the 3D approximation to the stone (either the visual hull or the VEMH), and k is a constant selected such that the mean percentage error is zero. The constant k is used to remove the systematic component of error. (Since visual hulls will consistently overestimate volume, it makes sense to correct for this bias.) The value of k is estimated from the data. This biases the computed error downwards, but since the one degree of freedom is small with respect to the number of samples (246), this bias is negligible. The approach of bias removal using multiplication by a constant determined from the data will be used for further shape property estimation in this chapter.

Confidence intervals given in the table are computed using Efron's bias-corrected and accelerated bootstrap method [36]. (This method is used for all the confidence intervals presented in this thesis.)

The table indicates that the proposed method provides volume estimates close to those obtained using the exact alignment. For the higher resolution images, the merged silhouette sets provide better volume estimates than the original silhouette sets from which they are formed. The mean of the volume estimates from the pairs of original silhouette sets provides a better volume estimate than the original silhouette sets, but at sufficiently high resolution it is not as accurate as the estimates from merged silhouette sets.

VEMHs provide more accurate volume estimates than visual hulls for higher resolution images, but not for lower resolution images. This is because at lower resolution, substantial portions of cone strips are destroyed, resulting in large regions in which there are no midpoints. This reduces the volumes of the computed VEMHs and increases the volume variance, since cone strip regions are destroyed at random.

The table also shows volume estimates based on the geometric and arithmetic means of the silhouette areas. (All of the silhouettes for the merged silhouette sets are used, i.e., 12 silhouettes per stone for the results shown in Table 6.2.) To remove the effect of depth on silhouette size, the depth z of the visual hull centroid is used. Silhouettes are specified in normalised image coordinates and then multiplied by the depth factor z. This closely approximates an orthographic projection since the depth of the stone is large with respect to the depth variation of points on the rim and the visual hull centroid.

The volume estimate V_Σ based on the arithmetic mean is computed as follows:

$$V_\Sigma = k_\Sigma \sum_{i=1}^{n} A_i^{3/2}, \tag{6.13}$$

where A_i is the area of the ith silhouette, and k_Σ is an empirically determined constant.

The volume estimate V_Π based on the geometric mean is computed as follows:

$$V_\Pi = k_\Pi \prod_{i=1}^{n} A_i^{3/2n} = k_\Pi \exp\left({}^{3}\!/_{2n} \sum_{i=1}^{n} \ln A_i \right), \tag{6.14}$$

where k_Π is an empirically determined constant.

The factors of ${}^{3}\!/_{2}$ in Equations 6.13 and 6.14 ensure a linear relationship with volume for parallel projections of a set of objects with the same shape and orientation, but varying size. In practice, variation in object shape and orientation is the main source of error.

Table 6.2 indicates that volume estimates based on silhouette area are less affected by image resolution reduction than the visual hull- and VEMH-based estimates. For higher resolution cases, the area-based estimates perform worse than the competing methods, whereas at the lowest resolution considered, the geometric mean of area provides a more accurate volume estimate than those derived from the merged silhouette sets. Arithmetic mean is the approach to volume estimation investigated by Taylor [126].

	1/4 resolution	1/8 resolution	1/16 resolution	1/32 resolution
quality	922.7	483.6	213.9	86.4
ET error	0.201 pixels	0.380 pixels	0.863 pixels	2.125 pixels
merged	*+3.70%*	*+3.38%*	*+1.47%*	*-8.32%*
pose est.	**1.28%**	**1.30%**	**1.61%**	**5.34%**
VH	(1.12%, 1.49%)	(1.15%, 1.50%)	(1.45%, 1.82%)	(4.89%, 5.86%)
merged	*+3.70%*	*+3.37%*	*+1.45%*	*-8.38%*
true pose	**1.28%**	**1.31%**	**1.62%**	**5.41%**
VH	(1.13%, 1.48%)	(1.15%, 1.50%)	(1.45%, 1.83%)	(4.94%, 6.02%)
merged	*+1.20%*	*+0.63%*	*-2.08%*	*-14.76%*
pose est.	**0.95%**	**1.00%**	**1.70%**	**7.13%**
VEMH	(0.83%, 1.11%)	(0.88%, 1.14%)	(1.55%, 1.88%)	(6.52%, 7.89%)
merged	*+1.19%*	*+0.62%*	*-2.09%*	*-14.71%*
true pose	**0.94%**	**0.99%**	**1.71%**	**7.33%**
VEMH	(0.83%, 1.10%)	(0.88%, 1.12%)	(1.56%, 1.88%)	(6.66%, 8.20%)
Run 1	*+8.37%*	*+8.15%*	*+6.74%*	*-0.66%*
6-view	**2.13%**	**2.15%**	**2.25%**	**4.06%**
VH	(1.88%, 2.55%)	(1.90%, 2.54%)	(2.01%, 2.59%)	(3.73%, 4.44%)
Run 2	*+8.21%*	*+8.00%*	*+6.54%*	*-0.90%*
6-view	**2.14%**	**2.17%**	**2.30%**	**4.34%**
VH	(1.96%, 2.36%)	(1.97%, 2.40%)	(2.09%, 2.53%)	(3.96%, 4.76%)
mean of	*+8.29%*	*+8.07%*	*+6.64%*	*-0.78%*
Run 1+2	**1.67%**	**1.70%**	**1.83%**	**3.85%**
VH	(1.50%, 1.93%)	(1.53%, 1.97%)	(1.66%, 2.07%)	(3.56%, 4.19%)
Run 1	*+1.72%*	*+1.39%*	*-0.44%*	*-9.83%*
6-view	**1.92%**	**1.94%**	**2.14%**	**5.36%**
VEMH	(1.71%, 2.24%)	(1.74%, 2.25%)	(1.95%, 2.41%)	(4.94%, 5.83%)
Run 2	*+1.56%*	*+1.25%*	*-0.56%*	*-10.13%*
6-view	**1.93%**	**1.99%**	**2.17%**	**5.80%**
VEMH	(1.76%, 2.12%)	(1.81%, 2.19%)	(1.97%, 2.40%)	(5.32%, 6.42%)
mean of	*+1.64%*	*+1.32%*	*-0.50%*	*-9.98%*
Run 1+2	**1.54%**	**1.58%**	**1.79%**	**5.25%**
VEMH	(1.40%, 1.75%)	(1.44%, 1.77%)	(1.63%, 1.97%)	(4.83%, 5.72%)
geometric	**4.41%**	**4.40%**	**4.44%**	**4.95%**
mean of area	(3.77%, 5.68%)	(3.78%, 5.73%)	(3.83%, 5.87%)	(4.42%, 6.00%)
arithmetic	**5.71%**	**5.71%**	**5.75%**	**6.19%**
mean of area	(4.84%, 7.48%)	(4.86%, 7.48%)	(4.91%, 7.50%)	(5.41%, 7.74%)

Table 6.2: Volume estimation using the six-view synthetic garnet data at various image resolution levels. Quality is the mean silhouette diameter divided by the mean ET error. ET error is the mean internal ET error over all silhouette sets. Mean percentage error is shown in italics. RMS percentage error is shown in boldface with 95% confidence intervals in brackets. Merged pose est. indicates that silhouette set pairs were merged using the proposed method. Merged true pose indicates that the ground truth pose value was used for merging. VH (visual hull) or VEMH indicates the method of 3D shape approximation used.

Table 6.3 shows the results of the volume estimation experiment applied to synthetic data formed using different numbers of cameras. Results for the two-camera setup clearly show that merging pairs of two-view silhouette sets provides poses that are insufficiently close to the true pose to provide improvements in volume estimation accuracy. Whereas the estimates based on merging using the true pose provide volume estimates that are more accurate than the competing methods, merging using the estimated pose provides volume estimates that are worse than the other corresponding hull-based methods. Increasing the number of cameras to three offers a substantial improvement: the volume estimates computed using the estimated pose are almost as accurate as those computed using the true pose. Increasing the number of cameras further provides a far greater improvement in the accuracy of methods based on the visual hull and the VEMH than the area-based methods.

6.4.2 Caliper Diameter Estimation with Synthetic Data

A further experiment to investigate the accuracy of caliper diameter estimation was carried out with the six-view silhouette sets at $1/4$ resolution level.

Ground truth values were determined for the shortest, intermediate, and longest diameters for the mesh models of stones.

Table 6.4 presents the results of estimating caliper diameters from the visual hulls and VEMHs of merged and original silhouette sets. Again, the estimates from the silhouette sets merged using the proposed method produce results that are very close to the results obtained using the ground truth poses for alignment. The proposed method also produces results that are more accurate than results that are computed from the original silhouette sets. The table also indicates that the longest diameter can be estimated more accurately than the shortest and intermediate diameters.

6.4.3 Mass Estimation with Data from the Two-Mirror Setup

The three runs of 5-view silhouette sets of the gravel data set were merged into 15-view silhouette sets using the proposed method. Figure 6.13 shows some examples of the 15-view visual hull models and photographs of the gravel from the same viewpoint (the photographs are cropped portions of the input images). Also shown are the three 5-view visual hulls from the original 5-view silhouette sets. The figure shows a version of the 15-view visual hull that is coloured according to which of the 5-view visual hulls share the surface region. This demonstrates that each of the three silhouette sets tends to contribute at least somewhat to the final 15-view visual hull.

Figure 6.14 shows some more examples of photographs of gravel and 15-view visual hulls rendered from the same viewpoint. These figures provide a qualitative illustration of the degree of accuracy that one can expect when using the two-mirror setup together with the proposed merging method. Since the visual hulls cannot

	2 cameras	3 cameras	4 cameras	10 cameras
merged	*+11.33%*	*+8.93%*	*+6.29%*	*+2.10%*
pose est.	**9.52%**	**2.82%**	**2.36%**	**0.65%**
VH	(8.46%, 11.21%)	(2.52%, 3.26%)	(2.06%, 2.75%)	(0.57%, 0.76%)
merged	*+17.22%*	*+8.96%*	*+6.29%*	*+2.09%*
true pose	**6.01%**	**2.79%**	**2.35%**	**0.65%**
VH	(5.15%, 7.88%)	(2.48%, 3.22%)	(2.06%, 2.75%)	(0.57%, 0.76%)
merged	*-5.70%*	*+1.60%*	*+1.62%*	*+0.77%*
pose est.	**13.98%**	**2.71%**	**1.78%**	**0.48%**
VEMH	(12.40%, 16.06%)	(2.37%, 3.27%)	(1.57%, 2.07%)	(0.43%, 0.56%)
merged	*+1.31%*	*+1.66%*	*+1.62%*	*+0.77%*
true pose	**5.95%**	**2.55%**	**1.78%**	**0.48%**
VEMH	(5.11%, 7.45%)	(2.30%, 2.94%)	(1.56%, 2.06%)	(0.43%, 0.56%)
Run 1	*+41.10%*	*+19.81%*	*+14.24%*	*+4.64%*
6-view	**7.57%**	**4.33%**	**3.99%**	**1.25%**
VH	(6.66%, 9.43%)	(3.71%, 5.41%)	(3.57%, 4.62%)	(1.11%, 1.44%)
Run 2	*+41.67%*	*+19.46%*	*+14.21%*	*+4.60%*
6-view	**9.36%**	**4.18%**	**3.69%**	**1.18%**
VH	(7.26%, 15.14%)	(3.70%, 5.40%)	(3.37%, 4.10%)	(1.04%, 1.46%)
mean of	*+41.38%*	*+19.63%*	*+14.23%*	*+4.62%*
Run 1+2	**6.85%**	**3.39%**	**2.96%**	**0.93%**
VH	(5.27%, 10.97%)	(2.87%, 4.49%)	(2.68%, 3.32%)	(0.84%, 1.06%)
Run 1	*-26.99%*	*-8.49%*	*-0.75%*	*+1.78%*
6-view	**7.65%**	**4.66%**	**3.99%**	**0.96%**
VEMH	(6.70%, 9.92%)	(3.94%, 5.85%)	(3.57%, 4.57%)	(0.86%, 1.12%)
Run 2	*-26.73%*	*-8.68%*	*-0.76%*	*+1.73%*
6-view	**9.39%**	**4.71%**	**3.60%**	**0.93%**
VEMH	(7.30%, 14.63%)	(4.11%, 6.38%)	(3.30%, 4.00%)	(0.82%, 1.10%)
mean of	*-26.86%*	*-8.58%*	*-0.76%*	*+1.76%*
Run 1+2	**7.15%**	**4.16%**	**3.02%**	**0.72%**
VEMH	(5.52%, 11.22%)	(3.55%, 5.54%)	(2.72%, 3.37%)	(0.65%, 0.81%)
geometric	**10.14%**	**5.80%**	**4.93%**	**4.24%**
mean of area	(8.97%, 12.78%)	(4.84%, 7.59%)	(4.32%, 6.08%)	(3.66%, 5.40%)
arithmetic	**10.29%**	**6.38%**	**6.04%**	**5.67%**
mean of area	(9.17%, 12.68%)	(5.33%, 8.31%)	(5.21%, 7.65%)	(4.80%, 7.32%)

Table 6.3: Volume estimation using the synthetic garnet data with different numbers of cameras at the 1/4 resolution level. Mean percentage error is shown in italics. RMS percentage error is shown in boldface with 95% confidence intervals in brackets. Merged pose est. indicates that silhouette set pairs were merged using the proposed method. Merged true pose indicates that the ground truth pose value was used for merging. VH (visual hull) or VEMH indicates the method of 3D shape approximation used.

Diameter	Shortest	Intermediate	Longest
merged	*+1.82%*	*-0.72%*	*-1.11%*
pose est.	**3.06%**	**2.76%**	**1.05%**
VH	(2.49%, 3.94%)	(2.47%, 3.14%)	(0.94%, 1.21%)
merged	*+1.81%*	*-0.78%*	*-1.13%*
true pose	**3.08%**	**2.80%**	**1.07%**
VH	(2.50%, 3.96%)	(2.51%, 3.17%)	(0.97%, 1.24%)
merged	*+0.58%*	*-1.63%*	*-1.83%*
pose est.	**2.34%**	**2.89%**	**1.02%**
VEMH	(1.99%, 2.80%)	(2.49%, 3.77%)	(0.91%, 1.19%)
merged	*+0.56%*	*-1.71%*	*-1.84%*
true pose	**2.35%**	**2.74%**	**1.06%**
VEMH	(2.03%, 2.77%)	(2.32%, 3.69%)	(0.95%, 1.22%)
Run 1	*+4.10%*	*+1.07%*	*+0.48%*
6-view	**4.93%**	**4.67%**	**1.42%**
VH	(4.22%, 6.03%)	(4.15%, 5.36%)	(1.29%, 1.56%)
Run 2	*+3.76%*	*+0.61%*	*+0.47%*
6-view	**4.38%**	**4.77%**	**1.46%**
VH	(3.84%, 5.15%)	(4.25%, 5.50%)	(1.33%, 1.65%)
mean of	*+3.93%*	*+0.84%*	*+0.48%*
Run 1+2	**3.85%**	**3.72%**	**1.27%**
VH	(3.31%, 4.60%)	(3.36%, 4.16%)	(1.16%, 1.42%)
Run 1	*+1.68%*	*-0.90%*	*-1.12%*
6-view	**3.58%**	**3.18%**	**1.10%**
VEMH	(3.13%, 4.36%)	(2.76%, 3.82%)	(0.99%, 1.23%)
Run 2	*+1.54%*	*-1.04%*	*-1.17%*
6-view	**3.67%**	**3.05%**	**1.06%**
VEMH	(3.25%, 4.26%)	(2.64%, 3.66%)	(0.96%, 1.18%)
mean of	*+1.61%*	*-0.97%*	*-1.15%*
Run 1+2	**3.03%**	**2.52%**	**0.91%**
VEMH	(2.65%, 3.62%)	(2.24%, 2.88%)	(0.82%, 1.04%)

Table 6.4: Estimating the three caliper diameters using pairs of synthetic 6-view silhouette sets at the 1/4 resolution level. Mean percentage error is shown in italics. RMS percentage error is shown in boldface with 95% confidence intervals in brackets. Merged pose est. indicates that silhouette set pairs were merged using the proposed method. Merged true pose indicates that the ground truth pose value was used for merging. VH (visual hull) or VEMH indicates the method of 3D shape approximation used.

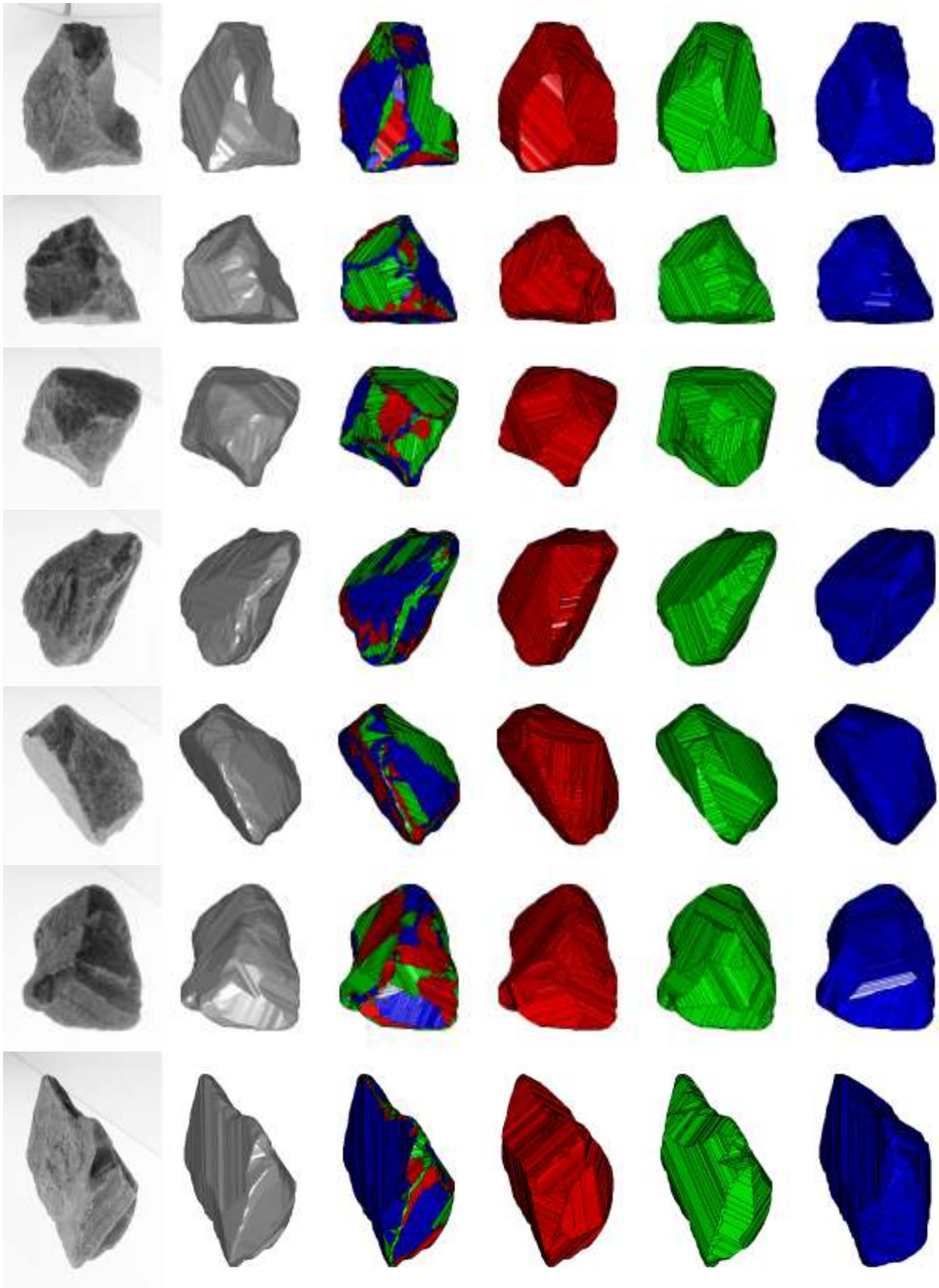


Figure 6.13: Some examples of visual hulls of pieces of gravel. The first column shows original images of the gravel. The second column shows the 15-view visual hull (formed from three 5-view silhouette sets) from the same viewpoint as the first column. The third column shows the 15-view visual hull surfaces coloured according to which of the three original 5-view visual hulls contributes to the surface region. The three original visual hull models are shown to the right in corresponding colours.

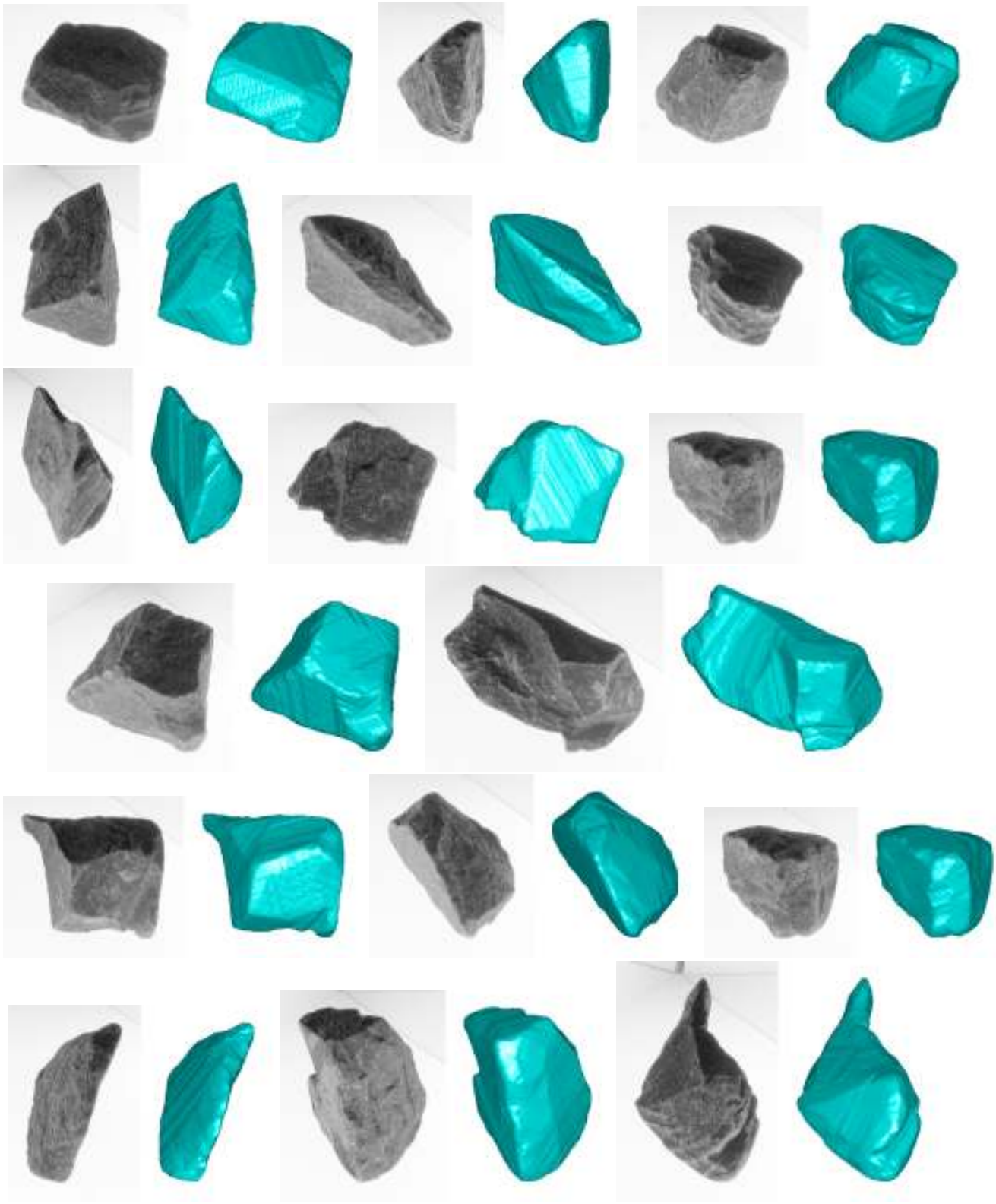


Figure 6.14: Images of pieces of gravel with visual hulls shown from the same viewpoint. The visual hulls were formed from three images of the stones, yielding $3 \times 5 = 15$ silhouettes for each visual hull.

model concavities, they exhibit regions of extra volume in certain places due to the lack of total coverage of the viewing sphere, and they show some striations due to image noise. Nonetheless, these 3D shapes appear to be likely to provide a better representation of particle shape than ellipsoidal models sometimes used in simulations.

Visual hulls of gravel were used to estimate the mass of gravel particles. VEMHs were not used as the gravel stones are nonconvex, whereas the VEMH approximates the convex hull of an object. The visual hull volume is used to form a mass estimate m_{est} as follows:

$$m_{\text{est}} = cV_{\text{VH}}, \tag{6.15}$$

where c is an empirically determined constant and V_{VH} is the visual hull volume. The constant c accounts for both the tendency of the visual hull to be an overestimate of stone volume and an implicit estimate of gravel density.

The mass estimates are limited by the extent to which gravel density varies from stone to stone. Attempts to measure ground truth volume (using the Archimedes Principle: weigh each stone in air and, using a cradle, underwater) rather than mass were abandoned, as the volume measurements were insufficiently repeatable.

Table 6.5 shows the results of gravel mass estimation. Note that unlike in the case of synthetic garnet data, the accuracy that can be achieved is limited by both the variation in density from stone to stone, and the variation in concavities from stone to stone. The table shows that the proposed merging method produces somewhat more accurate mass estimation results than averaging the volume estimation results from the three original silhouette sets. The table also indicates that the visual hull-based estimates are more accurate than the area-based estimates.

mass estimator	RMS%E	95% CI
merged 15-view visual hull volume	5.97%	(4.90%, 8.18%)
5-view visual hull volumes	7.63%	(6.80%, 9.41%)
mean of three 5-view visual hull volumes	6.54%	(5.62%, 8.60%)
5-view geometric mean of area	10.99%	(9.95%, 12.46%)
15-view geometric mean of area	10.11%	(9.04%, 11.60%)
5-view arithmetic mean of area	12.23%	(11.17%, 13.70%)
15-view arithmetic mean of area	11.00%	(9.87%, 12.52%)

Table 6.5: RMS percentage errors (RMS%E) and 95% confidence intervals for gravel mass estimates.

Figure 6.15 shows plots of mass versus visual hull volume for the 5-view and 15-view visual hulls. The plots show a linear relationship between mass and visual hull volume, with variability decreasing when fifteen views are used instead of five. Note that the data points associated with the largest error are gross overestimates of visual hull volume (due to unfavourable stone orientation), whereas gross underestimates

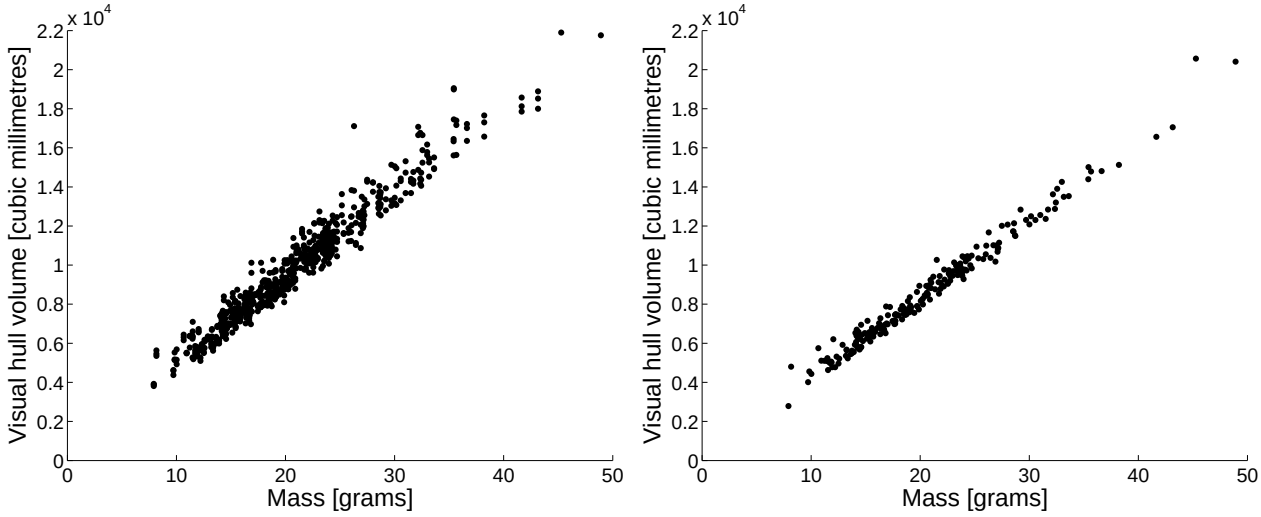


Figure 6.15: Plots of gravel mass versus visual hull volume for 5-view visual hulls *(left)* and 15-view visual hulls *(right)*.

of volume are not possible (in the absence of gross segmentation or calibration errors) since the visual hull is always larger than the stone.

6.4.4 Caliper Diameter Estimation with Data from the Two-Mirror Setup

Vernier calipers were used to manually measure the longest, intermediate, and shortest diameter of 100 of the stones from the gravel data set. Each stone was measured three times on three separate days, and the median value was used as a ground truth value.

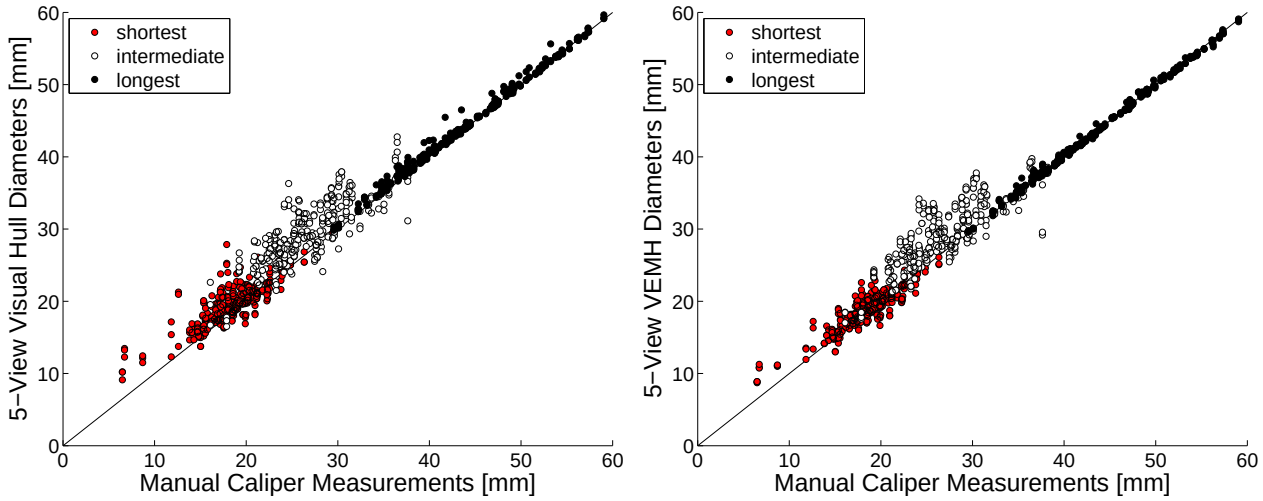
Figure 6.16 shows plots of the manually measured diameter values versus estimates based on 5-view silhouette sets using visual hulls and VEMHs. Silhouette-based estimates of the longest diameter agree more closely with manually estimated values for the longest diameter than for the intermediate and shortest diameter.

Table 6.6 shows error statistics for estimating caliper diameters using 5-view silhouette sets.

	Visual hull			VEMH		
	mean	RMS	RMS adjusted	mean	RMS	RMS adjusted
shortest	+8.04%	17.01%	14.05%	+3.10%	10.78%	10.08%
intermediate	+11.49%	15.59%	9.58%	+9.85%	14.08%	9.28%
longest	+1.01%	1.78%	1.45%	+0.15%	0.93%	0.92%

Table 6.6: Percentage errors for diameter estimates based on 5-view silhouette sets formed from the gravel data set. The 'RMS adjusted' value is computed after multiplying estimates by a constant to compensate for systematic error.

Coefficients of variation are shown for manual and silhouette-based caliper estimates in Table 6.7. The table



(a) (b)

Figure 6.16: Plot of manual caliper measurements versus estimates based on 5-view silhouette sets for the gravel data set using (a) visual hull-based caliper estimates, and (b) VEMH-based estimates.

	manual	visual hull	VEMH
shortest	6.53%	4.23%	2.83%
intermediate	5.68%	3.60%	2.34%
longest	1.34%	0.76%	0.42%

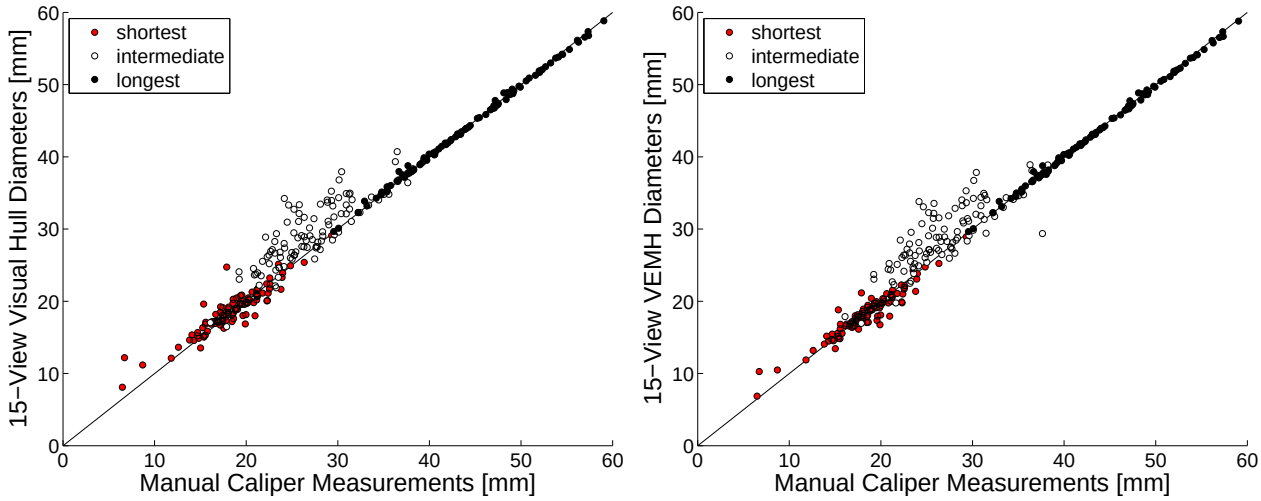
Table 6.7: Coefficients of variation of caliper diameters determined using different methods.

indicates that the manual measurements are the least repeatable. This means that inaccurate ground truth may account for the high errors observed in Table 6.6. The coefficients of variation indicate that the VEMH-based estimates are more repeatable than those based on visual hulls. For all three methods, estimates of the longest diameter are the most repeatable, whereas estimates of the shortest diameter are the least repeatable.

Figure 6.17 and Table 6.8 present the results of applying caliper diameter estimation to the 15-view merged silhouette sets formed from the original 5-view silhouette sets. The results indicate an improvement over the 5-view silhouette sets (see Figure 6.16 and Table 6.6).

	Visual hull			VEMH		
	mean	RMS	RMS adjusted	mean	RMS	RMS adjusted
shortest	+2.47%	11.37%	10.96%	-0.08%	7.91%	7.97%
intermediate	+9.86%	13.91%	9.02%	+9.11%	13.75%	9.56%
longest	+0.03%	0.75%	0.75%	-0.14%	0.77%	0.76%

Table 6.8: Percentage errors for diameter estimates based on 15-view silhouette sets formed from the gravel data set. The 'RMS adjusted' value is computed after multiplying estimates by a constant to compensate for systematic error.



(a) (b)

Figure 6.17: Plot of manual caliper measurements versus estimates based on merged 15-view silhouette sets for the gravel data set using (a) visual hull-based caliper estimates, and (b) VEMH-based estimates.

6.4.5 Mass Estimation with Data from the Six-Camera Setup

Mass measurements were made for the data set of 1423 uncut gemstones (illustrated on pages 222–224) using an electronic balance. Ten runs of 6-view silhouette sets were captured for each stone.

Mass estimates were carried out by multiplying the computed visual hull volume by a constant factor determined from the data (Equation 6.15).

Table 6.9 presents the results in terms of RMS percentage error for mass estimates computed using various silhouette-based methods. The table shows that greater volume estimation accuracy is achieved using merged visual hull volume, than by using the mean volume of the original 6-view visual hulls. However, both approaches increase in accuracy as the number of runs (and hence the number of available views) is increased. Results are also shown for visual hulls that are formed by aligning silhouette sets using the principal axes of visual hulls or VEMHs rather than adjusting pose to minimise ET error. These approaches produce inferior results to the ET minimised silhouette sets, and volume estimation error tends to *increase* as the number of runs is increased. Results are also shown for area-based mass estimates. These are substantially less accurate than visual hull-based estimates, and show only small improvements in accuracy as the number of available views is increased. Again, mass estimates based on the product of areas (geometric mean) outperform those based on the sum of areas (arithmetic mean).

Mass estimation was carried out on subsets of the 6-view silhouette sets to investigate performance using a small number of views. The n-view subsets are formed by discarding all but the first n views from the six available views. The results shown in Table 6.10 indicate that visual hull-based mass estimates outperform area-based methods even when as few as two views are used. However, the first column of the table shows

No. Runs	merged	mean VH vol.	VH aligned	VEMH aligned	geometric area mean	arithmetic area mean
1	4.88% (4.7, 5.3)				9.21% (8.8, 9.8)	10.97% (10.3, 11.8)
2	3.97% (3.8, 4.3)	4.67% (4.5, 5.0)	4.60% (4.4, 4.9)	4.61% (4.4, 4.9)	9.12% (8.7, 9.7)	10.92% (10.3, 11.7)
3	3.63% (3.5, 3.9)	4.66% (4.4, 5.0)	4.66% (4.5, 5.0)	4.65% (4.5, 5.0)	9.08% (8.6, 9.6)	10.92% (10.3, 11.7)
4	3.48% (3.3, 3.8)	4.60% (4.4, 5.0)	4.81% (4.6, 5.1)	4.77% (4.6, 5.1)	9.05% (8.6, 9.6)	10.92% (10.2, 11.7)
5	3.35% (3.2, 3.6)	4.55% (4.3, 4.9)	4.95% (4.7, 5.2)	4.90% (4.7, 5.2)	9.05% (8.6, 9.6)	10.92% (10.3, 11.8)
6	3.28% (3.1, 3.6)	4.52% (4.3, 4.9)	5.08% (4.9, 5.3)	5.03% (4.8, 5.3)	9.06% (8.6, 9.6)	10.92% (10.3, 11.7)
7	3.23% (3.1, 3.5)	4.49% (4.3, 4.9)	5.21% (5.0, 5.4)	5.14% (4.9, 5.4)	9.04% (8.6, 9.6)	10.92% (10.3, 11.7)
8	3.21% (3.0, 3.5)	4.49% (4.3, 4.8)	5.33% (5.1, 5.6)	5.26% (5.1, 5.5)	9.05% (8.6, 9.6)	10.92% (10.3, 11.7)
9	3.20% (3.0, 3.5)	4.50% (4.3, 4.8)	5.45% (5.3, 5.7)	5.37% (5.2, 5.6)	9.06% (8.6, 9.6)	10.93% (10.3, 11.7)
10	3.18% (3.0, 3.4)	4.50% (4.3, 4.9)	5.57% (5.4, 5.8)	5.48% (5.3, 5.7)	9.06% (8.6, 9.6)	10.93% (10.2, 11.7)

Table 6.9: RMS percentage errors for mass estimates based on 1–10 runs of 6-view silhouette sets of the data set of 1423 uncut gemstones: 'merged' is visual hulls formed from merging the available runs of silhouette sets with the proposed method; 'mean VH vol.' uses the mean value of the 6-view visual hull volumes for the available runs; 'VH aligned' uses merged visual hull volume, but without minimisation of ET error—visual hull principal axes and third order moments are used instead; 'VEMH aligned' uses merged visual hull volume with VEMH principal axes and third order moments used for merging; 'geometric' and 'arithmetic' use silhouette areas to estimate mass. Ninety-five percent confidence interval computed using a bootstrap approach are given in brackets.

cameras n	n-view VH	merged $2n$-view VH	geometric area mean	arithmetic area mean
2	13.80% (12.70%, 15.16%)	14.92% (13.93%, 16.21%)	15.99% (14.88%, 17.60%)	16.48% (15.32%, 18.12%)
3	9.67% (8.78%, 11.10%)	6.35% (5.98%, 6.84%)	13.21% (12.12%, 14.97%)	13.85% (12.74%, 15.68%)
4	6.41% (6.04%, 7.00%)	4.75% (4.53%, 5.06%)	10.69% (10.02%, 11.60%)	11.95% (11.15%, 13.03%)
5	5.25% (5.02%, 5.59%)	4.16% (3.98%, 4.42%)	9.32% (8.83%, 9.98%)	10.86% (10.24%, 11.68%)
6	4.88% (4.65%, 5.25%)	3.98% (3.79%, 4.27%)	9.21% (8.73%, 9.79%)	10.97% (10.30%, 11.73%)

Table 6.10: RMS percentage errors for mass estimation of 1423 uncut gemstones using subsets of the original 6-view silhouette sets: 'n-view VH' uses the n-view visual hull volumes to estimate mass; 'merged $2n$-view VH' uses visual hulls formed by merging two runs of n-view silhouette sets; 'geometric' and 'arithmetic' use n silhouette areas to estimate mass. Ninety-five percent confidence intervals are bracketed.

a substantial increase in accuracy as the number of views is increased from two to six. Merging pairs of 2-view silhouette sets produces less accurate results than considering the 2-view silhouette sets individually. This is because the 2-view silhouette sets do not provide sufficient constraints to produce accurate alignment. However, merging using three or more views leads to more accurate mass estimates than using the original silhouette sets before merging.

6.5 Summary

A method for merging more than one silhouette set of the same object into a single large silhouette set has been presented. The method adjusts relative pose to minimise the ET error across silhouette sets. Starting points for the minimisation are determined by using moments to align 3D approximations of the object computed from each of the original silhouette sets. When moment-based starting points do not lead to a sufficiently low ET error, starting points formed using a uniform random rotational component are considered.

Qualitative results computed using everyday objects such as toy animals demonstrate that better reconstructions can be obtained from a merged silhouette set than from any of the original silhouette sets used to form the merged set.

Experiments carried out using synthetic data demonstrate that volume estimates based on the merged silhouette sets are more accurate than those based on the original silhouette sets. Volume estimates computed using silhouette sets merged by minimising ET error are close to as accurate as those computed using silhouettes sets merged using the ground truth poses. Caliper diameter estimates are also more accurately estimated from merged silhouette sets than from the original silhouette sets.

The method is applied to data sets of stones captured using both the two-mirror setup and the six-camera setup. The accuracy with which mass and caliper diameters can be estimated is quantified. Mass estimates based on visual hull volume are demonstrated to be more accurate than those based on silhouette area. Results are compared with estimates based on merged silhouette sets. The merged silhouette sets show an improved accuracy for mass estimates and caliper diameter estimates. The accuracy associated with the caliper diameter estimates is likely underestimated, because of the difficultly in accurately manually measuring the ground truth values with a Vernier caliper. The silhouette-based methods are found to be more repeatable over multiple runs than the manual measurements.